Introducción a la Simulación y a la Teoría de Colas

Introducción a la Simulación y a la Teoría de Colas

Ricardo Cao Abad

Área de Estadística e Investigación Operativa

Departamento de Matemáticas, Facultad de Informática

Universidade da Coruña

netbiblo

www.netbiblo.com

Ficha de catalogación bibliográfica

Introducción a la simulación y a la teoría de colas • 1º Edición
Ricardo Cao Abad
NETBIBLO, S.L., A Coruña, 2002

ISBN: 84-9745-017-5

Materia: Matemáticas computacionales: 519.6

Formato: 17 x 24 • Páginas: 224

INTRODUCCIÓN A LA SIMULACIÓN Y A LA TEORÍA DE COLAS

http://www.netbiblo.com
editorial@netbiblo.com

ISBN: 84-9745-017-5
Deposito Legal: C-1867-2002

Editor: Carlos Iglesias
Diseño: Artia>DLF · Jose DelaFuente
Compuesto en: Centerprint, S.L. · Eduardo Bóveda
Impreso en: Josman Press

IMPRESO EN ESPAÑA - PRINTED IN SPAIN.

AUTOR

Catedrático de Universidad del área de Estadística e Investigación Operativa en el Departamento de Matemáticas de la Universidade da Coruña (UDC). En los últimos años ha impartido diversas asignaturas de Simulación y Teoría de Colas en las distintas titulaciones de la Facultad de Informática de la UDC.

Doctor en Matemáticas por la Universidad de Santiago de Compostela, Ricardo Cao es autor de otros seis libros y más de una treintena de artículos de investigación en revistas de prestigio internacional como *The Annals of Statistics*, *Journal of the American Statistical Association*, *Technometrics*, *Scandinavian Journal of Statistics*, *Journal of Econometrics*, *Test*, *Canadian Journal of Statistics*, *Computational Statistics and Data Analysis* y *Journal of Statistical Planning and Inference*, entre muchas otras.

Ha sido vocal del Consejo Académico de Estadística de la Sociedad de Estadística e Investigación Operativa (SEIO), es miembro fundador de la Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións (SGAPEIO) y miembro de la Sociedad Bernoulli y de la Sociedad Española de Biometría. Asimismo es Editor Asociado de la revista Test desde 1996.

Para más información: http://www.udc.es/dep/mate/ricardo/homepage.htm

Índice General

PARTE I

SIMULACIÓN ESTOCÁSTICA

1

Introducción a la simulación

1.1 Conceptos básicos

La simulación es la técnica que consiste en realizar experimentos de muestreo sobre el modelo de un sistema. Un modelo no es más que un conjunto de variables junto con ecuaciones matemáticas que las relacionan y restricciones sobre dichas variables. La modelización es una etapa presente en la mayor parte de los trabajos de investigación (especialmente en las ciencias experimentales). En muchas ocasiones, la realidad es bastante compleja como para ser estudiada directamente y es preferible la formulación de un modelo que contenga las variables más relevantes que aparececen en el fenómeno en estudio y las relaciones más importantes entre ellas.

Frecuentemente, la resolución de los problemas que se pretenden abordar puede realizarse por procedimientos analíticos sobre el modelo construido (normalmente mediante el uso de herramientas matemáticas como las de resolución de ecuaciones ordinarias o de ecuaciones diferenciales, el cálculo de probabilidades, etc.). En otras circunstancias dicha resolución analítica no es posible (o es tremendamente complicada o costosa) y es preferible una aproximación de la solución mediante simulación.

1.2 Experimentación real y simulación

La experimentación directa sobre la realidad puede tener muchos inconvenientes:

- un coste muy alto
- gran lentitud
- en ocasiones las pruebas son destructivas
- a veces no es ética (experimentación sobre seres humanos)
- puede resultar imposible (un acontecimiento futuro)

Razones como esas (y algunas otras) pueden indicar la ventaja de trabajar con un modelo del sistema real. La estadística es precisamente la ciencia que se preocupa de cómo estimar los parámetros y contrastar la validez de un modelo a partir de los datos observados del sistema real que se pretende modelizar.

1.3 Simulación necesaria e innecesaria

Una vez se ha construído un modelo, la primera tentativa debe ser siempre tratar de resolver analíticamente el problema que nos ocupa. En caso de ser esto posible, la solución es exacta (a menudo la resolución también es rápida). En caso contrario puede recurrirse a la simulación que involucrará mucha labor de procesado. Gracias a la gran potencia de cálculo de los computadores actuales los programas de simulación pueden ofrecer una solución aproximada rápida en la mayor parte de los problemas susceptibles de ser modelizados.

Ejemplo 1 *Supóngase que se quiere calcular la probabilidad de aparición de exactamente dos caras en tres lanzamientos de una moneda. La experimentación sobre la situación real consistiría en repetir numerosas veces los tres lanzamientos y observar con qué frecuencia se obtienen exactamente dos caras. El sistema real es el mecanismo por el cual se realizan los lanzamientos. Un modelo razonable para este sistema es el de utilizar una variable aleatoria* $X \stackrel{d}{=} B(3, 0.5)$ *(ya que se supone que la moneda tiene la misma probabilidad de cara que de cruz). Bajo este modelo, se trataría de calcular* $P(X = 2)$*. En este caso la resolución analítica es factible y muy sencilla:*

$$P(X = 2) = \binom{3}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1 = \frac{3}{8} = 0.375$$

La simulación consistiría en obtener números aleatorios (en ordenador) para replicar artificialmente los tres lanzamientos en gran cantidad de ocasiones, observando la frecuencia relativa con la que aparecen exactamente dos caras. Es obvio que, en una situación como esta, la experimentación real tiene la desventaja de la lentitud. Una vez construído, el modelo la resolución analítica resulta exacta e inmediata y, por tanto, preferible a la resolución aproximada mediante simulación. Cuestión distinta es si el modelo refleja adecuadamente la realidad. Así, por ejemplo, si la moneda no diese igual probabilidad a cara y cruz, el modelo no sería adecuado y la resolución mediante el mismo sería incorrecta. Obviamente este problema no lo tiene la experimentación sobre la situación real. En realidad, en un caso como ese, gracias a la inferencia estadística, podría formularse otro modelo más adecuado.

Ejemplo 2 *Supóngase el siguiente juego: un jugador lanza una moneda (abonando un euro por cada lanzamiento) hasta que el número de caras supere en tres al número de cruces obtenidas. En ese momento el jugador recibe 10 unidades monetarias. ¿Resulta rentable jugar? De nuevo aquí la experimentación real es muy lenta. La modelización puede realizarse de nuevo gracias a la teoría de la probabilidad. En esta ocasión, sin embargo, la resolución analítica sería complicada (salvo que se tengan conocimientos de cadenas de Markov). Parece, por tanto, conveniente una aproximación mediante simulación. Se trataría por tanto de ir replicando artificialmente el lanzamiento*

de la moneda simulando un gran número de posibles partidas y examinando la pérdida o ganancia media.

1.4 Ventajas e inconvenientes de la simulación

Ventajas:

1. En casos en los que la resolución analítica no puede llevarse a cabo.
2. Cuando existen medios de resolver analíticamente el problema pero dicha resolución es complicada y costosa.
3. Si se desea experimentar antes de que exista el sistema.
4. Cuando es imposible experimentar sobre el sistema real por ser dicha experimentación destructiva.
5. En ocasiones en las que la experimentación sobre el sistema es posible pero no ética.
6. Es de utilidad en sistemas que evolucionan muy lentamente en el tiempo.

Inconvenientes:

1. La construcción de un buen modelo puede ser una tarea muy laboriosa.
2. Frecuentemente el modelo omite variables o relaciones importantes entre ellas.
3. Resulta difícil conocer la precisión de la simulación, especialmente en lo relativo a la precisión del modelo formulado.

Ejercicio 1.4.1 *Una compañía petrolífera debe decidir en cuál de dos puntos (A ó B) ha de realizar una perforación para extraer petróleo. La profundidad estimada de la bolsa de petróleo en ambos puntos es de 57 m. en A y 40 m. en B. Desde este momento hasta el instante en que se comenzará a perforar es posible que llegue una nueva maquinaria (lo cual ocurre con una probabilidad de* 0.71*). Durante la perforación puede producirse una avería en la maquinaria. En base a estudios retrospectivos se estima que dicha avería ocurre con una probabilidad de* 0.14 *ó* 0.16 *si se perfora en A ó B, respectivamente, usando la maquinaria nueva. En caso de utilizar la vieja maquinaria, las probabilidades de avería son* 0.28 *y* 0.19 *para ambos puntos (A y B). Independientemente del tipo de maquinaria utilizada y del punto donde se ha decidido perforar, cuando se produce avería, ésta puede ser grande (con duración de tres días) con probabilidad de* 0.35*. En otro caso la avería es pequeña e implica un sólo día sin*

perforar. Sabiendo que la perforación avanza a una velocidad de 3 m. por día en el punto A y 2 m. por día en el punto B, se trata de dar respuesta, mediante simulación, a la pregunta ¿en qué punto es más rentable perforar?, atendiendo a que el tiempo hasta que se llega al petróleo sea mínimo. Hacer un diagrama de flujo del algoritmo de simulación en este contexto.

1.5 Contenidos de la parte de simulación

A lo largo del presente texto, los contenidos de simulación que se verán son los siguientes:

1. Estudio de los métodos más importantes de generación de números pseudoaleatorios en el intervalo $(0, 1)$, centrando nuestra atención en los generadores congruenciales. Se revisarán los métodos congruenciales implementados en algunos lenguajes de programación y bibliotecas de rutinas estadísticas. También se dará una visión de las diferentes medidas estadísticas para contrastar la calidad de una generador de números aleatorios.

2. Métodos universales (o generales) de simulación de distribuciones continuas: abordando los métodos de inversión y de aceptación rechazo.

3. Métodos universales de simulación de distribuciones discretas. Aquí se estudiarán el método de la transformación cuantil (con búsqueda secuencial y elección de etiquetado óptimo), algoritmos basados en árboles de Huffman, el método de la tabla guía y el método de truncamiento.

4. Un conjunto de método específicos de simulación para las distribuciones discretas y continuas más importantes.

5. Una breve incursión a la simulación de distribuciones multidimensionales, comentando algunos métodos generales y particularizando también al caso importante de la normal multivariante.

6. Algunas cuestiones prácticas sobre el diseño de experimentos de simulación, dedicando especial atención a las técnicas de reducción de la varianza.

2

Generación de números pseudoaleatorios uniformes en (0,1)

2.1 Introducción

Casi todos los métodos de simulación se basan en la posibilidad de generar números aleatorios con distribución $U(0,1)$. Hasta el gran desarrollo de los ordenadores los números aleatorios se obtenían por procedimientos experimentales (loterías, ruletas) y se almacenaban en tablas. En la actualidad estos números son generados por ordenador y se denominan pseudoaleatorios ya que, en realidad, todos los números de la sucesión que se genera son predecibles a partir del primero, llamado semilla. En cualquier caso, todo generador de números pseudoaleatorios mínimamente aceptable debe comportarse como si se tratase de una muestra genuina de datos independientes de una $U(0,1)$.

2.1.1 Propiedades deseables de un generador de números pseudoaleatorios

Para poder utilizar sin reservas un generador de números pseudoaleatorios éste debe satisfacer los contrastes estadísticos más habituales en este contexto: los de aleatoriedad (los contrastes de rachas o los de saltos), los de independencia (como los basados en autocorrelaciones, el test de Ljung-Box, el contraste de pares seriados, etc) y los de bondad de ajuste a una $U(0,1)$ (entre ellos el test chi-cuadrado y el de Kolmogoroff-Smirnoff). También existen otros contrastes específicos que tratan de indagar a la vez sobre varios de los aspectos anteriores. Entre ellos destacamos el contraste del poker y el del coleccionista.

Además de estas propiedades de tipo estadístico existen otros requisitos computacionales. Unos y otros pueden resumirse en la siguiente lista.

Requisitos deseables para un generador

1. Producir muestras según una distribución $U(0,1)$.
2. Pasar los contrastes de aleatoriedad e independencia más habituales.
3. Que la sucesión generada sea reproducible a partir de la semilla.
4. Tener una longitud de ciclo tan grande como se desee.
5. Generar valores a alta velocidad.
6. Ocupar poca memoria.

2.2 Método de los cuadrados medios

Es debido a von Neumann y tiene fundamentalmente sólo interés histórico.

1. Se toma un número entero inicial, x_0, llamado semilla, de $2n$ cifras.
2. Se eleva al cuadrado, obteniendo un número de $4n$ cifras (completando, quizá, con ceros a la izquierda).
3. Se considera x_1 el número entero formado por las $2n$ cifras centrales.
4. Se eleva al cuadrado x_1 y se repite el proceso anterior tantas veces como sea preciso.
5. Finalmente se consideran los números $u_i = \dfrac{x_i}{10^{2n}}$, ya en el intervalo $(0,1)$.

Ejemplo 3 *Tómese* $n = 2$ *y* $x_0 = 4122$. *Resulta:*

$$\begin{array}{llll} x_0 = 4122 & x_0^2 = 16|9908|84 & x_1 = 9908 & x_1^2 = 98|1684|64 \\ x_2 = 1684 & x_2^2 = 02|8358|56 & x_3 = 8358 & x_3^2 = 69|8561|64 \\ x_4 = 8561 & x_4^2 = 73|2907|21 & x_5 = 2907 & x_5^2 = 08|4506|49 \end{array}$$

De esta forma, los números pseudoaleatorios en $(0,1)$ *son*

$$\begin{array}{llll} u_0 = 0.4122 & u_1 = 0.9908 & u_2 = 0.1684 & u_3 = 0.8385 \\ u_4 = 0.8561 & u_5 = 0.2907 & & \end{array}$$

Siguiendo, de nuevo, con $n = 2$*, pero tomando como semilla* $x_0 = 3708$*, se obtiene*

$$\begin{array}{llll} x_0 = 3708 & x_0^2 = 13|7492|64 & x_1 = 7292 & x_1^2 = 56|1300|64 \\ x_2 = 1300 & x_2^2 = 01|6900|00 & x_3 = 6900 & x_3^2 = 47|6100|00 \\ x_4 = 6100 & x_4^2 = 37|2100|00 & x_5 = 2100 & x_5^2 = 04|4100|00 \\ x_6 = 4100 & x_6^2 = 16|8100|00 & x_7 = 8100 & x_7^2 = 65|6100|00 \\ x_8 = 6100 & & & \end{array}$$

Así pues, como $x_8 = x_4$*, los valores* u_4, u_5, u_6, u_7 *se repetirán ciclicamente de forma indefinida. Este tipo de fenómenos de ciclo corto son su mayor inconveniente.*

2.3 Método de Lehmer

El método consiste en los siguientes pasos:

1. Se toma como semilla un número entero, x_0, de n cifras.

2. Se elige otro entero, c, de k cifras. Suele tomarse $k < n$.

3. Se calcula $x_0 \cdot c$, número de, a lo sumo, $n + k$ cifras.

4. Se separan las k cifras de la izquierda de $x_0 \cdot c$ y al número formado por las n cifras restantes se le resta el que forman esas k cifras de la izquierda, dando lugar a x_1.

5. Se repite este proceso tantas veces como sea necesario.

6. Se devuelven los valores $u_i = \dfrac{x_i}{10^{2n}}$.

Ejemplo 4 *Tomando $n = 4$, $k = 2$, $x_0 = 4122$ y $c = 76$, se obtiene*

$$\begin{array}{lll} x_0 = 4122 & x_0 \cdot c = 31|3272 & 3272 - 31 = 3241 \\ x_1 = 3241 & x_1 \cdot c = 24|6316 & 6316 - 24 = 6292 \\ x_2 = 6292 & x_2 \cdot c = 47|8192 & 8192 - 47 = 8145 \\ x_3 = 8145 & x_3 \cdot c = 61|9020 & 9020 - 61 = 8959 \\ x_4 = 8959 & x_4 \cdot c = 68|0884 & 0884 - 68 = 0816 \\ x_5 = 0816 & x_5 \cdot c = 06|2016 & 2016 - 06 = 2010 \end{array}$$

De esta forma

$$\begin{array}{llll} u_0 = 0.4122 & u_1 = 0.3241 & u_2 = 0.6292 & u_3 = 0.8145 \\ u_4 = 0.8959 & u_5 = 0.0816 & & \end{array}$$

Todavía en el caso de que $n = 4$ y $k = 2$, pero con $x_0 = 2000$ y $c = 50$, se tiene $x_0 \cdot c = 10|0000$ y así $x_1 = 0000 - 10 = -10 < 0$. Este es precisamente uno de los peores inconvenientes de este método: la aparición de iterantes negativos. También aparecen, con frecuencia, ciclos cortos (en particular, el cero es un valor absorbente de este generador).

2.4 Métodos congruenciales

Se basan en la idea de considerar una combinación lineal de los últimos k enteros generados y calcular su resto al dividir por un entero fijo m. Consideraremos tan sólo el método congruencial simple que procede como sigue:

1. Elegir un número entero positivo m (normalmente en relación con el tipo de enteros que se va a usar) y otros dos números enteros, a y c, tales que $0 < a < m$ y $0 \leq c < m$.

2. Fijar la semilla x_0, un valor entero inicial que cumpla $0 \leq x_0 < m$.

3. Obtener de forma recurrente

$$x_n = (ax_{n-1} + c) \bmod m$$

para $n = 1, 2, \ldots$

4. Devolver los valores $u_n = \dfrac{x_n}{m}$, $n = 0, 1, \ldots$

Dadas las propiedades algebraicas de la suma y el producto en el conjunto de clases de resto módulo m (que es un anillo) se tiene que cualquier otra elección de a ó c mayores o iguales que m tiene un equivalente verificando la restricción impuesta. Cuando tomamos $c = 0$ el generador se dice congruencial multiplicativo.

Ejemplo 5 *Considérese un generador congruencial con* $m = 8$, $a = 5$, $c = 4$*:*

$$x_n = (5x_{n-1} + 4) \bmod 8$$

Tomando como semilla los valores 5 *ó* 2 *se obtiene:*

$$\begin{array}{llll} x_0 = 5 & x_1 = 5 & x_2 = 5 & \cdots \\ x_0 = 2 & x_1 = 6 & x_2 = 2 & x_3 = 6 \quad \cdots \end{array}$$

que presentan ciclos de longitud 1 y 2 respectivamente.

Cambiando el valor de c *a* 2 *se tiene* $x_n = (5x_{n-1} + 2) \bmod 8$ *y así,*

$$\begin{array}{lllll} x_0 = 5 & x_1 = 3 & x_2 = 1 & x_3 = 7 & x_4 = 5 \quad \cdots \\ x_0 = 2 & x_1 = 4 & x_2 = 6 & x_3 = 0 & x_4 = 2 \quad \cdots \end{array}$$

donde ambos ciclos son de longitud cuatro.

Finalmente dejando el mismo valor de m *pero eligiendo* $a = 5$ *y* $c = 5$*, se tiene* $x_n = (5x_{n-1} + 5) \bmod 8$*, que conduce a*

$$\begin{array}{lllll} x_0 = 5 & x_1 = 6 & x_2 = 3 & x_3 = 4 & x_4 = 1 \\ x_5 = 2 & x_6 = 7 & x_7 = 0 & x_8 = 5 & \cdots \\ x_0 = 2 & x_1 = 7 & x_2 = 0 & x_3 = 5 & x_4 = 6 \\ x_5 = 3 & x_6 = 4 & x_7 = 1 & x_8 = 2 & \cdots \end{array}$$

con ciclo de longitud 8, que es el máximo valor posible.

2.4.1 Generadores congruenciales de ciclo máximo

Además de las propiedades estadísticas deseables para cualquier generador, una cuestión importante para los generadores congruenciales (como se ha visto en los ejemplos previos) es la de garantizar que el ciclo del generador sea máximo

(o, cuando menos, muy elevado). En la práctica tratará de tomarse el ciclo igual o muy próximo al número de enteros de tipo largo del lenguaje en cuestión.

En general, se define la longitud del ciclo (o período) de un generador de números pseudoaleatorios uniformes, como el menor número entero positivo, p, que cumple que existe un n_0 natural tal que $x_{i+p} = x_i$ para todo $i = n_0, n_0 + 1, \ldots$ En el caso de un generador congruencial mixto el máximo valor para el período es m. Esto es así porque en este tipo de generadores x_n es función unívoca del anterior iterante, x_{n-1}. Así, basta, por tanto, que se repita algun valor en la sucesión de los x_n para que todo se repita cíclicamente a partir de él. Cómo los valores posibles para los x_n son los enteros $0, 1, \ldots, m-1$, entonces es obvio que, a lo sumo después de m generaciones debe haber una repetición en la sucesión. También es muy fácil demostrar que si un generador congruencial tiene ciclo máximo para cierta elección de la semilla, entonces lo tiene para cualquier otra.

Un resultado que sirve para caracterizar qué propiedades deben cumplir los parámetros de un generador congruencial para que tenga período máximo es el Teorema de Knuth (1969).

Teorema 6 (Knuth) *Las siguientes condiciones son necesarias y suficientes para que un generador congruencial con parámetros m, a y c, tenga período máximo (i.e. $p = m$).*

1. c y m son primos entre sí (i.e. m.c.d. $(c, m) = 1$).
2. $a - 1$ es múltiplo de todos los factores primos de m (i.e. $a \equiv 1 \bmod g$, para todo g factor primo de m).
3. Si m es múltiplo de 4, entonces $a - 1$ también lo ha de ser (i.e. $m \equiv 0 \bmod 4 \Rightarrow a \equiv 1 \bmod 4$).

A la luz del Teorema de Knuth, es fácil darse cuenta porqué sólo el tercero de los generadores del ejemplo anterior tenía período óptimo.

2.4.2 Generadores congruenciales de algunos lenguajes y bibliotecas de rutinas

En ordenadores binarios es muy común elegir $m = 2^\beta$ ó $m = 2^\beta - 1$, donde β depende del tamaño de palabra (típicamente m será el mayor entero representable en el ordenador o una unidad mayor que él). En lenguajes, como el C, en el que existen tipos de enteros positivos (entre 0 y $2^\beta - 1$) la elección $m = 2^\beta$ permite "ahorrar" la operación de calcular el resto de la división entre m que aparece en el generador congruencial. La razón es que, al calcular $ax_{n-1} + c$ y sólo poder almacenarlo con β bits, los bits sobrantes (correspondientes a 2^β, $2^{\beta+1}, \ldots$) se desechan, obteniendo precisamente un número entre 0 y $2^\beta - 1$ que difiere del valor exacto de $ax_{n-1} + c$ en un múltiplo de $m = 2^\beta$. Esto es tanto

como decir que el valor que almacena el ordenador al calcular $ax_{n-1}+c$ (perdiendo los bits más significativos) es precisamente el que se deseaba calcular $(ax_{n-1}+c) \bmod m$.

En los generadores con $m = 2^\beta$ resulta especialmente fácil expresar las condiciones del Teorema de Knuth de forma mucho más sencilla. Así, al ser m una potencia de 2, su único factor primo es el 2 y, por tanto la primera condición equivale a que c sea impar. Para $\beta \geq 2$ se tiene que m es múltiplo de 4 y, por tanto la tercera condición impone que $a-1$ también lo sea. Por último, de nuevo por ser el 2 el único factor primo de m, la segunda condición pediría que $a-1$ fuese par, lo cual ya es consecuencia de que sea múltiplo de 4. En resumen, si $m = 2^\beta$, con $\beta \geq 2$, el generador congruencial tiene período máximo si y sólo sí c es impar y $a = 4k+1$, siendo k un número natural.

Algunos generadores habituales en lenguajes con enteros (con signo) de 36 bits corresponden a las elecciones

$$m = 2^{35} \quad a = 2^7 + 1 \quad c = 1$$
$$m = 2^{35} \quad a = 5^{15} \quad c = 1$$

Todos ellos tienen tienen período máximo (e igual a $2^{35} \simeq 3.44 \times 10^{10}$).

Otros generadores congruenciales para enteros (con o sin signo) de 32 bits y algunos lenguaje o bibliotecas que los usan o los han usado en el pasado son

$m = 2^{31}$	$a = 314159269$	$c = 453805245$	
$m = 2^{31}-1$	$a = 16807$	$c = 0$	APL, IMSL y SIMPL/I
$m = 2^{31}-1$	$a = 630360016$	$c = 0$	Algunos FORTRAN
$m = 2^{32}$	$a = 663608941$	$c = 0$	Ahrens y Dieter (1974)

Aunque sólo el primero tiene período máximo, los demás lo tienen muy elevado.

El lenguaje C (bajo UNIX) posee un generador congruencial de números pseudoaleatorios de 48 bits: el `drand48`. Sus parámetros son $m = 2^{48}$, $a = 25214903917$ y $c = 11$. La semilla se inicializa mediante la sentencia `srand48()` introduciendo como argumento un entero de 32 bits que corresponde a los 32 bits más significativos de x_0 (entero de 48 bits). Los 16 bits de menor orden se toman siempre coincidentes con el número (decimal) 13070. Los parámetros a y c se pueden modificar por el usuario desde otras rutinas del C. Los valores por defecto para estas cantidades ofrecen un generador de período máximo ya que $m = 2^\beta$, con $\beta = 48$, c es impar y $a = 6303725979 \cdot 4 + 1$.

2.5 Medidas estadísticas de la calidad de un generador de números aleatorios

La mayoría de los contrastes estadísticos para estudiar la calidad de un generador de números aleatorios se basan en medir posibles discrepancias (en algún

sentido) de las muestras generadas por el método en cuestión con respecto a las hipótesis de aleatoriedad, independencia y ajuste a una distribución $U(0,1)$.

En casi todos los casos se acaba recurriendo a un contraste de tipo chi-cuadrado en el que se comparan las frecuencias esperadas, e_i, de ciertas modalidades, con las observadas, o_i, mediante el estadístico

$$D = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

que seguirá, aproximadamente, una distribución χ^2_{k-1}, bajo la hipótesis nula que se contrasta. A continuación detallamos algunos de los contrastes más habituales.

2.5.1 El contraste chi-cuadrado

Para llevarlo a acabo, es necesario dividir el intervalo $(0,1)$ en k subintervalos (normalmente de la misma longitud). De esa forma la modalidad i-ésima será simplemente el subintervalo $[\frac{i-1}{k}, \frac{i}{k})$, para $i = 1, 2, \ldots, k$. De esta forma, después de simular un gran número, n, de valores, $X_1, X_2, \ldots, X_n$, mediante el generador calcularemos las frecuencias observadas en cada intervalo y las confrontaremos con las esperadas ($e_i = \frac{n}{k}$) mediante el estadístico D. Para que la distribución χ^2_{k-1} sea una aproximación razonable a la del estadístico y así el contraste pueda aplicarse suele pedirse la condición mínima de que $e_i \geq 5$ para todo i.

2.5.2 El contraste de Kolmogoroff-Smirnoff

Se basa en el cálculo de la máxima discrepancia entre la función de distribución teórica ($F(x) = x$, si $x \in [0,1]$, en este caso) y la distribución empírica de la muestra generada, dada por

$$F_n(x) = \frac{\#\{X_i \leq x\}}{n}, \forall x \in [0,1].$$

Esta discrepancia se calcula mediante el estadístico

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

cuya distribución (la K de Kolmogoroff) está tabulada.

2.5.3 El contraste del coleccionista

Consiste en fijar un entero positivo, M, y discretizar los valores generados, $X_1, X_2, \ldots, X_n$, de la forma $\lceil M \cdot X_i \rceil + 1$, donde $\lceil x \rceil$ denota la parte entera de x. De esta forma se consigue una sucesión de enteros aleatorios cuyos valores están comprendidos entre 1 y M. Ahora se procede (como un coleccionista) a contabilizar cuál es el número, Q, (aleatorio) de valores a generar hasta que se completa la colección de todos los enteros entre 1 y M. Obviamente, bajo las hipótesis de aleatoriedad y distribución $U(0,1)$, cada posible entero entre 1 y M tiene la misma probabilidad de aparecer en cada generación y, por tanto, resulta posible calcular la distribución de probabilidad de Q. De esta forma podemos utilizar los valores calculados de las probabilidades

$$P(Q = M),\ P(Q = M+1),\ \ldots$$

para calcular las frecuencias esperadas de cada clase y confrontarlas con las observadas vía el estadístico chi-cuadrado.

Existen varias elecciones comunes de M, así como pequeñas variantes del contraste:

Tomando $M = 5$ con clases $Q = 5$, $Q = 6$, $\ldots$, $Q = 19$, $Q \geq 20$.

Aplicando el test a partir de la tercera cifra decimal

Esto equivale a considerar $100 \cdot X_i - \lceil 100 \cdot X_i \rceil$ en lugar de X_i en el planteamiento anterior.

En ambos casos, las probabilidades de las clases vienen dadas por

$$\begin{array}{ll} P(Q=5) = 0.03840000 & P(Q=6) = 0.07680000 \\ P(Q=7) = 0.09984000 & P(Q=8) = 0.10752000 \\ P(Q=9) = 0.10450944 & P(Q=10) = 0.09547776 \\ P(Q=11) = 0.08381645 & P(Q=12) = 0.07163904 \\ P(Q=13) = 0.06011299 & P(Q=14) = 0.04979157 \\ P(Q=15) = 0.04086200 & P(Q=16) = 0.03331007 \\ P(Q=17) = 0.02702163 & P(Q=18) = 0.02184196 \\ P(Q=19) = 0.01760857 & P(Q \geq 20) = 0.07144851 \end{array}$$

Eligiendo $M = 10$

Tomando las categorías (con sus correspondientes probabilidades) como sigue:

$$\begin{array}{ll} P(10 \leq Q \leq 19) = 0.17321155 & P(20 \leq Q = 23) = 0.17492380 \\ P(24 \leq Q \leq 27) = 0.17150818 & P(28 \leq Q \leq 32) = 0.17134210 \\ P(33 \leq Q \leq 39) = 0.15216056 & P(Q \geq 40) = 0.15685380 \end{array}$$

2.5.4 Contrastes de salto

Dados dos números α y β tales que $0 \leq \alpha < \beta \leq 1$, los contrastes de saltos tratan de examinar, para cada valor generado, X_i, si se cumple $\alpha \leq X_i \leq \beta$, anotando, en ese caso, un 1 (0 en caso contrario). En estas condiciones, la probabilidad de que aparezca un 1 es $p = \beta - \alpha$ y la de que aparezcan j ceros desde la aparición de un uno hasta la del siguiente uno es $p_j = p\,(1-p)^j$, $j = 0, 1, 2, \ldots$ (que corresponde a una distribución geométrica). De nuevo puede aplicarse el test chi-cuadrado a las clases resultantes.

Las elecciones más habituales de α y β dan lugar a los siguientes contrastes:

El test de rachas bajo la mediana

Consiste en tomar $\alpha = 0$ y $\beta = 1/2$.

El test de rachas sobre la mediana

Corresponde al caso $\alpha = 1/2$ y $\beta = 1$.

El test del tercio medio

Que no es más que la elección $\alpha = 1/3$ y $\beta = 2/3$.

2.5.5 El contraste de permutaciones

Dada la sucesión de números pseudoaleatorios generada se consideran bloques de T valores consecutivos. Cada uno de los bloques puede presentar una cualquiera de las $T!$ posibles ordenaciones de esos T valores. Además, de ser el generador adecuado, cada posible ordenación ocurrirá con igual probabilidad: $\frac{1}{T!}$. El test consiste en observar una gran número de bloques y comparar las frecuencias observadas de cada posible ordenación con las esperadas mediante el test chi-cuadrado. Las elecciones más comunes son $T = 3$, 4, ó 5.

2.5.6 El contraste del poker

En un primer momento se procede como en el contraste del coleccionista con $M = 10$. A partir de aquí hay varias formas de actuar:

Poker 1

Se toman conjuntos sucesivos de cinco enteros y, para cada uno, se determina cuál de las siguientes posibilidades se da:

1. Un mismo entero se repite cinco veces (abreviadamente, $AAAAA$).
2. Un mismo entero se repite cuatro veces y otro distinto aparece una vez ($AAAAB$).
3. Un entero se repite tres veces y otro distinto se repite dos ($AAABB$).
4. Un entero se repite tres veces y otros dos distintos aparecen una vez cada uno ($AAABC$).
5. Un entero se repite dos veces, otro distinto se repite también dos veces y un tercer entero diferente aparece una sóla vez ($AABBC$).
6. Un entero se repite dos veces y otros tres distintos aparecen una vez cada uno ($AABCD$).
7. Los cinco enteros que aparecen son todos distintos ($ABCDE$).

Bajo la hipótesis de aleatoriedad y ajuste a una $U(0,1)$, pueden calcularse las probabilidades de estas modalidades:

$$\begin{aligned} P(AAAAA) &= 0.0001,\ P(AAAAB) = 0.0045,\ P(AAABB) = 0.0090, \\ P(AAABC) &= 0.0720,\ P(AABBC) = 0.1080,\ P(AABCD) = 0.5040, \\ P(ABCDE) &= 0.3024. \end{aligned}$$

Es frecuente que las clases $AAAAA$ y $AAAAB$ se agrupen a la hora de aplicar el test chi-cuadrado, ya que, en caso contrario, la restricción habitual $e_i \geq 5$ llevaría a que $0.0001 \cdot \frac{n}{5} \geq 5$, es decir, $n \geq 250000$.

Poker 2

Algo también bastante habitual es usar conjuntos de cinco enteros (como en el caso anterior) pero definiendo las categorías según el número de enteros distintos de entre los cinco observados. Así

$$\begin{aligned} P(\text{1 entero diferente}) &= 0.0001,\ P(\text{2 enteros diferentes}) = 0.0135, \\ P(\text{3 enteros diferentes}) &= 0.1800,\ P(\text{4 enteros diferentes}) = 0.5040, \\ P(\text{5 enteros diferentes}) &= 0.3024, \end{aligned}$$

procediendo frecuentemente a agrupar las dos primeras modalidades.

Poker 3

A menudo se consideran conjuntos de cuatro enteros. En tal caso,

$$\begin{aligned} P(AAAA) &= 0.001,\ P(AAAB) = 0.036,\ P(AABB) = 0.027, \\ P(AABC) &= 0.432,\ P(ABCD) = 0.504, \end{aligned}$$

siendo también bastante habitual el agrupar las dos primeras categorías.

2.5.7 El contraste de las rachas ascendentes

Es bastante más complicado de justificar que los anteriores. En términos intuitivos, se basa en las longitudes de las rachas ascendentes que aparecen en la sucesión de números aleatorios.

2.5.8 El contraste de pares seriados

La idea consiste en fijar un entero $M \geq 2$ y considerar los enteros $\lceil M \cdot X_i \rceil + 1$, tomar ahora estos valores apareados y utilizar el contrate chi-cuadrado considerando como categorías los posibles pares (i, j) tales que $i, j \in \{1, 2, \ldots, M\}$. Así se medirá la discrepancia entre la frecuencias observadas en estas categorías y las esperadas, iguales todas a $\frac{n}{2}\frac{1}{M^2}$. La elecciones más frecuentes son $M = 3$, 10 ó 20.

2.5.9 El contraste de Ljung-Box

Consiste en fijar un entero positivo, m, calcular las autocorrelaciones muestrales:

$$r(k) = \frac{\sum_{i=k+1}^{n} (x_i - \overline{x})(x_{i-k} - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

y, apartir de ellas, obtener el estadístico

$$Q = n(n+2) \sum_{k=1}^{m} \frac{r(k)^2}{n-k},$$

que, bajo la hipótesis de independencia, se distribuye aproximadamente según una χ^2_{m-1}.

2.5.10 Chi-cuadrado sobre chi-cuadrado

Todos los contrastes anteriores se han planteado desde la perspectiva de la realización de una única prueba. Es decir, se toma un número, n (normalmente grande), de valores obtenidos por el generador y se realiza el contraste evaluando el estadístico y comparándolo con el punto crítico de una chi-cuadrado para decidir si se acepta o rechaza la hipótesis (independencia, ajuste, aleatoriedad). En realidad tiene mucho más sentido la realización de un gran número de pruebas, evaluando en cada una el valor del estadístico y, o bien observar

que la proporción de rechazos del test se aproxima al valor nominal fijado (normalmente $\alpha = 0.01$ ó $\alpha = 0.05$), o más precisamente aplicando, de nuevo, el contraste chi cuadrado para comprobar el ajuste de la distribución del estadístico a la chi-cuadrado especificada bajo la hipótesis nula.

2.6 Ejercicios propuestos

1. Estudiar el período del generador congruencial: $x_{n+1} = (1002\ x_n\ +\ 40) \bmod 1002001$, $u_n = x_n/1002001$.

2. ¿Tienen periodo máximo los siguientes generadores?

   (a) $x_{n+1} = (2x_n + 3) \bmod 39$
   (b) $x_{n+1} = (2x_n) \bmod 39$
   (c) $x_{n+1} = (2x_n + 3) \bmod 40$
   (d) $x_{n+1} = (x_n + 2) \bmod 43$

3. Dar un generador congruencial mixto, de período máximo, usando enteros de 8 bits. Justifíquese su optimalidad. Utilizando como semilla el entero cero, obtener los cuatro primeros valores generados según el algoritmo obtenido anteriormente

4. Dado el generador congruencial definido por

$$\begin{aligned} x_{n+1} &= (x_n + 1) \bmod 2^{16} \\ u_{n+1} &= \frac{x_{n+1}}{2^{16}},\ n = 0, 1, \ldots \end{aligned}$$

   para enteros positivos de 16 bits, ¿es de ciclo máximo? ¿Goza de buenas propiedades estadísticas? Justificar las respuestas. Si no lo es, construir otro generador, para ese mismo tamaño de palabra, que sea de ciclo máximo.

5. ¿Es siempre preferible un generador congruencial con período máximo a otro cuyo período sea la mitad? Razonar la respuesta.

6. ¿Es de ciclo óptimo el siguiente generador congruencial de números pseudoaleatorios?

$$\begin{aligned} x_n &= (25x_{n-1} + 81) \bmod 1024, \\ u_n &= \frac{x_n}{m},\ n = 1, 2, \ldots \end{aligned}$$

   Justificar la respuesta. Tomando como semilla $x_0 = 40$, ¿cúanto vale el valor u_{51202}?

3

Métodos universales para la simulación de variables contínuas

En lo que sigue se expondrán dos de los métodos generales para simular distribuciones continuas: el método de inversión y el de aceptación/rechazo. Ambos son aplicables a gran número de contextos, sin más que la distribución que se desea simular tenga ciertas características. En ambos casos la herramienta indispensable es algún método de generación de números pseudoaleatorios uniformes en (0,1).

3.1 El método de inversión

Es el método universal por antonomasia para simular distribuciones continuas. También a veces se conoce como método de Montecarlo. Está basado en el siguiente resultado teórico.

Teorema 7 (de inversión) *Sea X una variable aleatoria con función de distribución F, continua e invertible. Entonces, la variable aleatoria $U = F(X)$, transformada de la original mediante su propia función de distribución, tiene distribución $U(0,1)$. Como consecuencia, si $U \stackrel{d}{=} U(0,1)$ entonces la variable $F^{-1}(U)$ tiene función de distribución F (la misma distribución que la de X).*

Demostración: Denotando por G la función de distribución de U y dado un valor $u \in (0,1)$, se tiene

$$G(u) = P(Y \leq u) = P(F(X) \leq u) = P\left(X \leq F^{-1}(u)\right) = F\left(F^{-1}(u)\right) = u.$$

Por otra parte es obvio que $G(u) = 0$ si $u \leq 0$ y $G(u) = 1$ si $u \geq 1$, con lo cual G es la función de distribución de una $U(0,1)$. La segunda parte es trivial, ya que, al ser F invertible

$$X = F^{-1}(F(X)) = F^{-1}(U),$$

teniendo U distribución $U(0,1)$.

El resultado anterior da pie al siguiente algoritmo genérico para simular cualquier variable continua con función de distribución F invertible:

Algoritmo (método de inversión)

1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = F^{-1}(U)$.

Ejemplo 8 *Dar un algoritmo, basado en el método de inversión, para simular la distribución exponencial de parámetro* $\lambda > 0$.

La función de densidad de una $\exp(\lambda)$ *es*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases}$$

y su función de distribución:

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

que es continua e invertible en el intervalo $[0, \infty)$. *Obsérvese que*

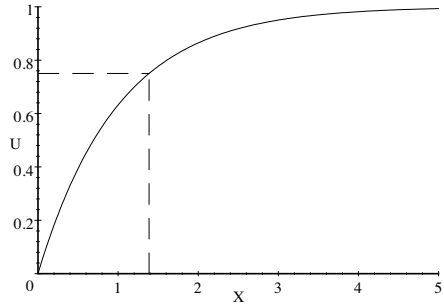
$$x = F^{-1}(u) \Leftrightarrow F(x) = u \Leftrightarrow 1 - e^{-\lambda x} = u$$
$$\Leftrightarrow 1 - u = e^{-\lambda x} \Leftrightarrow x = -\frac{\ln(1-u)}{\lambda}.$$

Como consecuencia, el algoritmo sería

1. *Generar* $U \sim U(0,1)$.
2. *Devolver* $X = -\dfrac{\ln(1-U)}{\lambda}$.

El algoritmo anterior puede abreviarse en tiempo de cálculo si en lugar de usar $1-U$ *se utiliza el número aleatorio generado* U. *Esto no afecta a la validez del método pues si* $U \overset{d}{=} U(0,1)$ *entonces* $1-U \overset{d}{=} U(0,1)$ *y, por tanto,* $g(U) \overset{d}{=} g(1-U)$ *para cualquier transformación* g. *Otro hecho que permite ahorra gran cantidad de cálculos cuando (como suele ser habitual) se va a llamar un gran número de veces al generador, consiste en definir inicialmente una variable* $L = -\frac{1}{\lambda}$, *que permitirá evitar la operación de cambio de signo en llamadas sucesivas y substituir una división por una multiplicación (más rápida de procesar). En resumen esta versión simplificada del algoritmo resulta:*

0. *Hacer* $L = -1/\lambda$.
1. *Generar* $U \sim U(0,1)$.
2. *Devolver* $X = L \cdot \ln U$
3. *Repetir los pasos 1-2 tantas veces como se precise.*



Interpretación gráfica del algoritmo de inversión.

3.1.1 Ventajas e inconvenientes del método de inversión

La ventaja más importante del método de inversión es que, en general, es aplicable a cualquier distribución continua. No obstante el método presenta algunos inconvenientes.

Inconvenientes del método de inversión

1. En ocasiones la función de distribución no tiene una expresión explícita (por ejemplo para la distribución normal).

2. A veces, aún teniendo una expresión explícita para $F(x)$, es imposible despejar x en la ecuación $F(x) = u$ (es decir, encontrar una expresión explícita para $F^{-1}(u)$).

3. Aún siendo posible encontrar $x = F^{-1}(u)$, puede ocurrir que esta expresión sea complicada y conlleve una gran lentitud de cálculo.

4. El primero de los inconvenientes expuesto puede, a veces, subsanarse mediante el uso de aproximaciones de la distribución en cuestión o mediante tabulaciones de la misma. El segundo suele abordarse mediante la utilización de métodos numéricos para la resolución aproximada de la ecuación $F(x) = u$. El mayor problema práctico que esto conlleva es la necesidad de resolver numéricamente una ecuación cada vez que se desee generar un nuevo número aleatorio que siga esa distribución (sin que los cálculos hechos para el anterior valor simulado sean de ayuda).

3.1.2 Algunas distribuciones que pueden simularse por el método de inversión

A continuación se incluyen algunas distribuciones que se pueden simular fácilmente mediante el método de inversión. Se adjunta una forma simplificada del método que tiene por objeto evitar cálculos innecesarios (tal y como se hizo en el ejemplo de la exponencial).

Nombre y densidad	$F(x)$	$F^{-1}(U)$	Forma simplificada
$\exp(\lambda)$ $(\lambda > 0)$ $\lambda e^{-\lambda x}$, si $x \geq 0$	$1 - e^{-\lambda x}$	$-\dfrac{\ln(1-U)}{\lambda}$	$-\dfrac{\ln U}{\lambda}$
Cauchy $\dfrac{1}{\pi(1+x^2)}$	$\dfrac{1}{2} + \dfrac{\arctan x}{\pi}$	$\tan\left(\pi\left(U - \dfrac{1}{2}\right)\right)$	$\tan \pi U$
Triangular en $(0,a)$ $\dfrac{2}{a}\left(1 - \dfrac{x}{a}\right)$, si $0 \leq x \leq a$	$\dfrac{2}{a}\left(x - \dfrac{x^2}{2a}\right)$	$a\left(1 - \sqrt{1-U}\right)$	$a\left(1 - \sqrt{U}\right)$
Pareto $(a, b > 0)$ $\dfrac{ab^a}{x^{a+1}}$, si $x \geq b$	$1 - \left(\dfrac{b}{x}\right)^a$	$\dfrac{b}{(1-U)^{1/a}}$	$\dfrac{b}{U^{1/a}}$
Weibull $(\lambda, \alpha > 0)$ $\alpha\lambda^\alpha x^{\alpha-1} e^{-(\lambda x)^\alpha}$, si $x \geq 0$	$1 - e^{-(\lambda x)^\alpha}$	$\dfrac{(-\ln(1-U))^{1/\alpha}}{\lambda}$	$\dfrac{(-\ln U)^{1/\alpha}}{\lambda}$

3.1.3 Inversión aproximada

Como se comentó anteriormente, en casos en los que no es posible determinar una expresión explícita para $F(x)$ o en los que no se puede hallar la de su inversa, puede optarse por encontrar expresiones sencillas que aproximen razonablemente bien la función $F^{-1}(u)$. A continuación se detalla la aproximación encontrada por Odeh y Evans para la normal estándar.

Estos autores consideran la función auxiliar

$$g(v) = \sqrt{-2\ln v}\frac{A\left(\sqrt{-2\ln v}\right)}{B\left(\sqrt{-2\ln v}\right)},$$

siendo $A(x) = \sum_{i=0}^{4} a_i x^i$ y $B(x) = \sum_{i=0}^{4} b_i x^i$, con

$a_0 = -0.322232431088$ $\quad a_1 = -1$
$a_2 = -0.342242088547$ $\quad a_3 = -0.0204231210245$
$a_4 = -0.0000453642210148$ $\quad b_0 = 0.0993484626060$
$b_1 = 0.588581570495$ $\quad b_2 = 0.531103462366$
$b_3 = 0.103537752850$ $\quad b_4 = 0.0038560700634$

La aproximación consiste en utilizar $g(1-u)$ en lugar de $F^{-1}(u)$ para los valores de $u \in [10^{-20}, \frac{1}{2}]$ y $-g(u)$ si $u \in [\frac{1}{2}, 1-10^{-20}]$. Para $u \notin [10^{-20}, 1-10^{-20}]$ (que sólo ocurre con una probabilidad de $2 \cdot 10^{-20}$) la aproximación no es recomendable.

Algoritmo de Odeh y Evans
1. `Generar` $U \sim U(0,1)$.
2. `Si` $U < 10^{-20}$ `ó` $U > 1 - 10^{-20}$ `entonces volver a 1.`
3. `Si` $U < 0.5$ `entonces hacer` $X = g(1-U)$ `sino hacer` $X = -g(U)$.
4. `Devolver` X.

3.2 El método de aceptación rechazo

Es un método universal alternativo al de inversión que está adaptado al caso en que, aunque se desconozca una fórmula explícita para $F(x)$ o sea difícil de resolver $F(x) = u$, sí se disponga de una expresión (preferiblemente sencilla) para la función de densidad $f(x)$. El método está basado en el siguiente resultado teórico.

Teorema 9 (de aceptacion/rechazo) *Sea X una variable aleatoria con función de densidad f y sea U otra variable aleatoria, independiente de la anterior, con distribución $U(0,1)$. Entonces, para cada $c > 0$, la variable aleatoria bidimensional $(X, c \cdot U \cdot f(X))$ tiene distribución uniforme en el recinto $A = \{(x,y) \in \mathbb{R}^2 / 0 \leq y \leq cf(x)\}$. Recíprocamente, si dada una función de densidad f, un vector aleatorio (X,Y) tiene distribución uniforme sobre el conjunto A, entonces, su primera componente, X, es una variable aleatoria unidimensional con función de densidad f.*

El teorema anterior establece la equivalencia entre la simulación de densidades unidimensionales y la simulación de variables bidimensionales con distribución uniforme sobre el hipografo de $c \cdot f(x)$ (el conjunto de puntos del plano que quedan por debajo de la gráfica de $c \cdot f$ pero por encima del eje OX). La idea del algoritmo consistirá en utilizar el recíproco en el teorema para simular valores de ese tipo de distribuciones bidimensionales y luego tomar la primera componente. Para simular valores de esa distribución bidimensional se usa también el teorema en sentido directo aplicándolo a otra densidad auxiliar g, fácil de simular.

Supóngase que se desea simular una distribución con densidad f y que no es factible hacerlo por el método de inversión. Considérese otra distribución, con densidad g, fácil de simular, de forma que exista cierta constante $c > 0$ tal que

$$f(x) \leq c \cdot g(x), \text{ para todo } x \in \mathbb{R}.$$

Definamos los hipografos de f y de $c \cdot g$:

$$\begin{aligned} A_f &= \{(x,y)/0 \leq y \leq f(x)\}, \\ A_{cg} &= \{(x,y)/0 \leq y \leq c \cdot g(x)\}. \end{aligned}$$

Gracias a la condición anterior, se tiene que $A_f \subset A_{cg}$.

Dado que la densidad g es fácil de simular, puede aplicarse la primera parte del teorema de aceptación/rechazo para encontrar una variable aleatoria bidimensional, (T, Y), con distribución uniforme sobre A_{cg}. Aceptando tan sólo los valores de (T, Y) que pertenezcan a A_f se tendrá una variable bidimensional con distribución uniforme sobre A_f. Técnicamente hablando estamos afirmando que la distribución condicionada $(T, Y)\,|_{(T,Y)\in A_f}$ es uniforme sobre A_f. Finalmente la segunda parte del teorema permite obtener una variable con densidad f sin más que tomar la primera componente del par obtenido.

De forma más detallada, el método constaría de los siguientes pasos:

1. Generar un valor T con densidad g.
2. Utilizar el teorema para encontrar un par (T, Y) con distribución uniforme en A_{cg}.
3. Comprobar si $(T, Y) \in A_f$ y en caso afirmativo hacer $X = T$. En caso contrario volver al paso 1.

El par de valores (T, Y) se obtiene simplemente simulando $U \sim U(0, 1)$ y definiendo $Y = c \cdot U \cdot g(T)$. Además, la condición $(T, Y) \in A_f$ que hay que comprobar en el último paso equivale a $Y \leq f(T)$. Teniendo todo esto en cuenta el algoritmo procedería como sigue.

Algortimo de aceptación/rechazo

1. `Repetir`
1.1. `Generar` $U \sim U(0, 1)$ `y` T `con densidad` g.
2. `Hasta que` $c \cdot U \cdot g(T) \leq f(T)$.
3. `Devolver` $X = T$.

Ejemplo 10 (densidades acotadas en un intervalo cerrado) *Sea f una función de densidad cualquiera con soporte en un intervalo cerrado $[a, b]$ (es decir, $\{x/f(x) \neq 0\} = [a, b]$) de tal forma que $\exists M > 0$ tal que $f(x) \leq M$ $\forall x$ (es decir, f es acotada superiormente). En este caso puede tomarse como densidad auxiliar g, la de una $U[a, b]$. En efecto, tomando $c = M(b - a)$ y teniendo en cuenta que*

$$g(x) = \begin{cases} \frac{1}{b-a} & si\ x \in [a, b] \\ 0 & si\ x \notin [a, b] \end{cases}$$

se tiene que $f(x) \leq M = \frac{c}{b-a} = c \cdot g(x)$, $\forall x \in [a, b]$. Así pues, el algoritmo quedaría como sigue:

1. Repetir
1.1. Generar $U, V \sim U(0, 1)$.
1.2. Hacer $T = a + (b - a)V$.
2. Hasta que $M \cdot U \leq f(T)$.
3. Devolver $X = T$.

3.2.1 Eficiencia del algoritmo

Dado que el algoritmo de aceptación/rechazo repite los pasos 1-2 un número aleatorio de veces, será importante medir, de alguna forma, la eficiencia del mismo. En primer lugar, existen restricciones obvias para la constante c que ha de elegirse en el algortimo. Así, debido al hecho de que $f(x) \leq c \cdot g(x)$, se tiene

$$1 = \int f(x)\,dx \leq c \int g(x)\,dx = c,$$

luego $c \geq 1$. Puede demostrarse además que si $c = 1$ entonces f y g serían densidades correspondientes a la misma distribución (iguales salvo en un conjunto de probabilidad cero) y, por tanto, si g es fácil de simular igualmente fácil lo sería f. Así pues, se tiene $c > 1$.

La comprobación que aparece en el paso 2 del algoritmo es $c \cdot U \cdot g(T) \leq f(T)$. La probabilidad de aceptación de esta condición es

$$p = \frac{\text{area}(A_f)}{\text{area}(A_{cg})} = \frac{\int f(x)\,dx}{\int c \cdot g(x)\,dx} = \frac{1}{c}.$$

De ésta se obtiene la probabilidad de rechazo: $q = \frac{c-1}{c}$. El flujo del algoritmo es aleatorio y el número de repeticiones de los pasos 1-2 hasta poder generar un valor de f (paso 3) es una variable aleatoria, N, con distribución geométrica (entendida ésta como el número de pruebas necesarias hasta obtener el primer éxito). En tales circunstancias el número medio de repeticiones de los pasos 1-2 es

$$E(N) = \frac{1}{p} = c,$$

luego c puede interpretarse como el número medio de comparaciones necesarias (o de repeticiones de los pasos 1-2, o de pares de variables (T, U) que se necesitan generar) hasta obtener un valor simulado de la variable X. Es obvio, por tanto, que cuanto más cercano a 1 sea el valor de c más eficiente será el algoritmo.

Ejercicio 3.2.1 *Dar un algoritmo para simular la función de densidad dada por* $f(x) = \frac{1}{16}(3x^2 + 2x + 2)$ *si* $x \in [0, 2]$, *cero en otro caso. Estudiar su eficiencia.*

3.2.2 Elección de c

Una vez fijada la densidad g es obvio que el mejor valor de c (que denotaremos por c_{opt}) se obtiene al encontrar el más pequeño número real c que verifica

$f(x) \leq c \cdot g(x)$, es decir: $c \geq \frac{f(x)}{g(x)}$, para todo x del soporte de g (que ha de contener al de f).

De esta forma, ha de cumplirse que $f(x) \neq 0 \Rightarrow g(x) \neq 0$ y además

$$c \geq \max_{x/g(x)>0} \frac{f(x)}{g(x)}.$$

Así pues, el menor valor posible que cumple esta condición es

$$c_{\text{opt}} = \max_{x/g(x)>0} \frac{f(x)}{g(x)}.$$

Ejemplo 11 (Simulación de la normal mediante la doble exponencial) *Se trata de simular la distribución normal estándar, cuya función de densidad viene dada por*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \text{ para todo } x \in \mathbb{R},$$

mediante aceptación/rechazo, utilizando como densidad auxiliar la doble exponencial de parámetro 1 (o distribución de Laplace). Esta última distribución es la asociada al experimento consistente en simular una distribución exponencial de parámetro 1 y luego otorgarle un signo positivo o negativo con equiprobabilidad. Su función de densidad viene dada por

$$g(x) = \frac{1}{2} e^{-|x|}, \text{ para todo } x \in \mathbb{R}.$$

Aunque la distribución doble exponencial puede simularse mediante el experimento compuesto antes comentado, resulta muy sencillo calcular su función de distribución:

$$G(x) = \int_{-\infty}^{x} g(t)\, dt = \begin{cases} \int_{-\infty}^{x} \frac{1}{2} e^{t} dt = \frac{1}{2} e^{x} & si\ x < 0 \\ \int_{-\infty}^{0} \frac{1}{2} e^{t} dt + \int_{0}^{x} \frac{1}{2} e^{-t} dt = 1 - \frac{1}{2} e^{-x} & si\ x \geq 0 \end{cases}$$

y aplicar directamente el método de inversión para obtener el algoritmo

1. *Generar* $V \sim U(0,1)$.
2. *Si* $V < 0.5$ *entonces* $T = \ln 2V$, *sino hacer* $T = -\ln 2(1-V)$.
3. *Devolver* T.

Por otra parte el soporte de la densidad g contiene al de f (de hecho ambos son toda la recta real) y, por tanto, el valor óptimo para c será

$$c_{opt} = \max_{x \in \mathbb{R}} \frac{f(x)}{g(x)} = \max_{x \in \mathbb{R}} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\frac{1}{2} e^{-|x|}} = \sqrt{\frac{2}{\pi}} \max_{x \in \mathbb{R}} e^{\varphi(x)} = \sqrt{\frac{2}{\pi}} e^{\max_{x \in \mathbb{R}} \varphi(x)},$$

donde $\varphi(x) = -\frac{x^2}{2} + |x|$. *Dado que esta función es simétrica, continua en toda la recta real y diferenciable tantas veces como se desee salvo en* $x = 0$, *bastará encontrar su máximo absoluto en el intervalo* $[0, \infty)$*:*

$$\begin{aligned} x &> 0 \Rightarrow \varphi'(x) = -x + 1, \varphi''(x) = -1; \\ \{x &> 0, \varphi'(x) = 0\} \Leftrightarrow x = 1 \\ \varphi''(1) &< 0. \end{aligned}$$

De esta forma, φ *alcanza un máximo relativo en* $x = 1$ *y otro de idéntico valor en* $x = -1$. *Resulta fácil demostrar que ambos son máximos absolutos (por los intervalos de crecimiento y decrecimiento de la función). Consiguientemente,*

$$c_{opt} = \sqrt{\frac{2}{\pi}} e^{\varphi(1)} = \sqrt{\frac{2}{\pi}} e^{1/2} = \sqrt{\frac{2e}{\pi}} \simeq 1.3155.$$

Como consecuencia el algoritmo procedería del siguiente modo:

1. Repetir

1.1. Generar $U, V \sim U(0,1)$.

1.2. Si $V < 0.5$ *hacer* $T = \ln 2V$, *sino hacer* $T = -\ln 2(1-V)$.

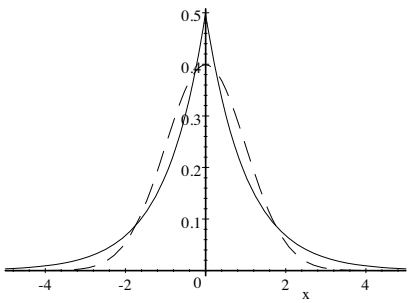
2. Hasta que $U \cdot \exp\left(\frac{T^2}{2} - |T| + \frac{1}{2}\right) \leq 1$.

3. Devolver $X = T$.

La condición que hay que comprobar para decidir si hay aceptación o rechazo surge de que

$$c \cdot U \cdot \frac{g(T)}{f(T)} = \sqrt{\frac{2e}{\pi}} U \sqrt{\frac{\pi}{2}} \exp\left(\frac{T^2}{2} - |T|\right) = U \cdot \exp\left(\frac{T^2}{2} - |T| + \frac{1}{2}\right).$$

Dado que el número medio de repeticiones de los pasos 1-2 hasta que se obtiene un valor simulado para X *es* $c \simeq 1.3155$ *y la probabilidad de aceptación en el paso 2 es* $p = 1/c = \sqrt{\frac{\pi}{2e}} = 0.76017$, *puede decirse que el algoritmo es bastante eficiente.*



Densidades normal estándar (trazo discontinuo)
y doble exponencial auxiliar (trazo continuo)

3.2.3 Elección de la densidad auxiliar g

Como se ha comentado anteriormente, un aspecto importante que influye en la eficiencia del método de aceptación/rechazo es el valor de la constante c. Conocida la densidad auxiliar g sabemos cómo elegir c de forma que el algoritmo sea lo más eficiente posible, sin embargo es obvio que algunas densidades auxiliares serían mejores candidatas que otras para conseguir un método eficiente.

En general, cuanto más parecida sea la forma de g a la de f, más pequeño es el mínimo c necesario para conseguir que la gráfica de $c \cdot g$ quede por encima de la de f. De todas formas, el problema de encontrar la densidad auxiliar g que ofrezca un c (óptimo) lo menor posible, no tiene solución. Mejor dicho, tiene la solución trivial $g = f$, que es absolutamente inútil para la implementación del algoritmo, pues si f era difícil de simular, no podemos tomar como g la propia f (ya que sería igual de difícil de simular).

Una solución intermedia al problema de elegir una función de densidad auxiliar, g, adecuada consiste en tomar cierta familia paramétrica de densidades que presenten un abanico de formas entre las que haya alguna que se parece bastante a la de f: $\{g_\theta / \theta \in \Theta\}$, encontrar el valor de c óptimo para cada densidad de esa familia:

$$c_\theta = \max_x \frac{f(x)}{g_\theta(x)}$$

y, finalmente, elegir el mejor valor del parámetro, θ_0, en el sentido de ofrecer el menor posible c_θ:

$$c_{\theta_0} = \min_{\theta \in \Theta} \max_x \frac{f(x)}{g_\theta(x)}.$$

Veamos un ejemplo.

Ejemplo 12 *Supóngase que se desea utilizar la doble exponencial de parámetro $\lambda > 0$ como densidad auxiliar en el método de aceptación rechazo para simular la normal estándar. La densidad doble exponencial ya fue introducida en el ejemplo anterior en el caso en el que el parámetro valga $\lambda = 1$. En un caso general su fórmula es*

$$g_\lambda(x) = \frac{\lambda}{2} e^{-\lambda|x|}, \text{ para todo } x \in \mathbb{R}.$$

Siguiendo los mismos pasos que para el caso $\lambda = 1$, anteriormente estudiado, un algoritmo para simular esta distribución basado el método de inversión es el siguiente:

1. **Generar** $V \sim U(0,1)$.
2. **Si** $V < 0.5$ **hacer** $T = \frac{1}{\lambda} \ln 2V$, **sino hacer** $T = -\frac{1}{\lambda} \ln 2(1-V)$.
3. **Devolver** T.

Si pretendemos encontrar el mejor valor de λ, en términos de eficiencia del algoritmo, debemos calcular

$$c_{\lambda_0} = \min_{\lambda>0} \max_{x\in\mathbb{R}} \frac{f(x)}{g_\lambda(x)} = \min_{\lambda>0} \max_{x\in\mathbb{R}} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\frac{\lambda}{2} e^{-\lambda|x|}}.$$

análogamente a lo visto para el caso $\lambda = 1$, se tiene

$$c_\lambda = \max_{x\in\mathbb{R}} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\frac{\lambda}{2} e^{-\lambda|x|}} = \frac{1}{\lambda}\sqrt{\frac{2}{\pi}} \max_{x\in\mathbb{R}} e^{\varphi_\lambda(x)} = \frac{1}{\lambda}\sqrt{\frac{2}{\pi}} e^{\max_{x\in\mathbb{R}} \varphi_\lambda(x)},$$

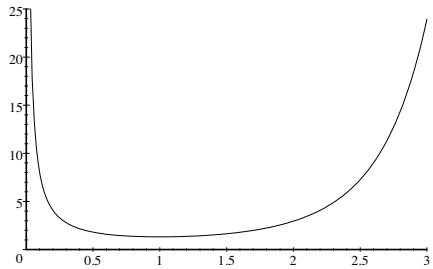
donde $\varphi_\lambda(x) = -\frac{x^2}{2} + \lambda|x|$. De forma totalmente similar a aquel caso puede probarse que φ_λ alcanza su máximo absoluto en los puntos $x = \pm\lambda$, siendo dicho valor máximo $\varphi_\lambda(\pm\lambda) = \frac{\lambda^2}{2}$. Como consecuencia,

$$c_\lambda = \frac{1}{\lambda}\sqrt{\frac{2}{\pi}} e^{\varphi_\lambda(\pm\lambda)} = \frac{e^{\frac{\lambda^2}{2}}}{\lambda}\sqrt{\frac{2}{\pi}}$$

Ahora debemos encontrar λ_0 tal que $c_{\lambda_0} = \min_{\lambda>0} c_\lambda$:

$$\begin{aligned}
\frac{dc_\lambda}{d\lambda} &= \sqrt{\frac{2}{\pi}} \frac{\lambda e^{\frac{\lambda^2}{2}}\lambda - e^{\frac{\lambda^2}{2}}}{\lambda^2} = \sqrt{\frac{2}{\pi}} \frac{e^{\frac{\lambda^2}{2}}\left(\lambda^2 - 1\right)}{\lambda^2}, \\
\frac{d^2c_\lambda}{d\lambda^2} &= \sqrt{\frac{2}{\pi}} \frac{[\lambda e^{\frac{\lambda^2}{2}}\left(\lambda^2 - 1\right) + e^{\frac{\lambda^2}{2}} 2\lambda]\lambda^2 - e^{\frac{\lambda^2}{2}}\left(\lambda^2 - 1\right) 2\lambda}{\lambda^4} \\
&= \sqrt{\frac{2}{\pi}} \frac{e^{\frac{\lambda^2}{2}}\left(\lambda^5 + \lambda^3 - 2\lambda^3 + 2\lambda\right)}{\lambda^4} = \sqrt{\frac{2}{\pi}} \frac{e^{\frac{\lambda^2}{2}}\left(\lambda^5 - \lambda^3 + 2\lambda\right)}{\lambda^4}, \\
\frac{dc_\lambda}{d\lambda} &= 0 \Leftrightarrow \lambda = 1, \text{ ya que } \lambda > 0 \\
\left.\frac{d^2c_\lambda}{d\lambda^2}\right|_{\lambda=1} &= 2\sqrt{\frac{2e}{\pi}} > 0, \text{ luego } \lambda = 1 \text{ es un punto de mínimo.}
\end{aligned}$$

De esto se deduce que la mejor doble exponencial, como densidad auxiliar en el algoritmo, es la correspondiente a $\lambda = 1$. Esta fue la usada en el ejercicio anterior, así pues, el algortimo más eficiente (con densidad auxiliar doble exponencial) es el expuesto en dicho ejercicio.



Gráfica de c_λ frente a λ

3.3 Ejercicios propuestos

1. Dar un algoritmo para simular la variable aleatoria continua con densidad

$$f(x) = \left(\frac{2x}{e^x}\right)^2, \text{ si } x \geq 0,$$

utilizando, de forma auxiliar, el generador de una exponencial de parámetro uno. Comentar su eficiencia.

2. Dada una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} 6x(1-x) & \text{si } x \in [0,1] \\ 0 & \text{en otro caso} \end{cases}$$

dar un algoritmo para simular valores de la misma. Analizar la eficiencia de dicho algoritmo.

3. Al bombardear una lámina circular de radio 1cm, hecha de plata, con partículas α, la distancia de cada impacto al centro del círculo resulta ser una variable aleatoria continua con función de densidad dada por $f(x) = 3x^2$, si $0 \leq x \leq 1$. Detallar un algoritmo, lo más sencillo posible, para simular la distancia al centro de la lámina en sucesivos impactos.

4. El tiempo de respuesta (en centésimas de segundo) de un servidor informático es una variable con función de distribución dada por $F(x) = 1 - (x+1)e^{-x}$, si $x \geq 0$ y cero en otro caso. Dar un algoritmo, lo más detallado posible, para simular valores de dicho tiempo de respuesta.

5. La densidad de probabilidad de una variable aleatoria viene dada por

$$f(x) = \begin{cases} (2x^2+1)\, e^{-2x} & \text{si } x \geq 0 \\ 0 & \text{en otro caso} \end{cases}$$

Encontrar un algoritmo para simular valores procedentes de dicha distribución y comentar su eficiencia.

6. Dada la función de densidad $f(x) = \frac{x^3-12x+20}{48}$, si $x \in [0,4]$ y cero en el resto, detallar un algoritmo que permita simular valores de la misma.

7. El tiempo (en años) entre dos inspecciones fiscales a una empresa es una variable aleatoria con densidad dada por

$$f(x) = \begin{cases} 0.11 - 0.0018x - 0.00003x^2 & \text{si} \quad x \in [0,10] \\ 0 & \text{si} \quad x \notin [0,10] \end{cases}$$

Dar un algoritmo, lo más sencillo posible, que permita simular valores de esa variable. Comentar el grado de eficiencia de dicho algoritmo frente

a otras alternativas posibles para conseguir el mismo fin. Utilizando un generador congruencial con parámetros $a = 5$, $c = 33$ y $m = 1024$ y semilla $x_0 = 27$, simular tres tiempos entre inspecciones, por medio del algoritmo encontrado. Describir un método para simular el número de inspecciones realizadas desde una dada hasta dentro de veinte años.

8. El tiempo, en milisegundos, de anticipación (o retraso) de una señal respecto de otra en un sistema de comunicaciones puede considerarse una variable aleatoria con función de distribución dada por $F(x) = \dfrac{e^x}{1+e^x}$. Dar un algoritmo, lo más detallado y simple posible, para simular valores de dicho tiempo de respuesta. Comparar la eficiencia computacional del método con la del algoritmo clásico para simular la distribución exponencial.

9. El número de horas diarias durante las cuales un terminal (de tipo A) de un laboratorio está siendo utilizado es una variable con función de distribución:

$$F(x) = \begin{cases} \dfrac{e^{x-2}}{1+e^{x-2}} & \text{si} \quad x \geq 0 \\ 0 & \text{en otro caso} \end{cases}$$

Se pide:

(a) Dar un algoritmo para simular dicha variable.

(b) Si para otro tipo de terminales (el tipo B) el tiempo de utilización tiene distribución $\exp\left(\frac{1}{2}\right)$, detallar un algoritmo para aproximar, por simulación, la probabilidad de que un terminal de tipo B sea más utilizado que uno de tipo A.

4

Métodos universales para la simulación de variables discretas

En lo que sigue se expondrán algunos métodos generales para simular distribuciones discretas. En concreto, se estudiará el método de la transformación cuantil en su versión clásica y con etiquetados óptimos, el método de las tablas guía y los métodos de truncamiento.

El problema que nos ocupa ahora es simular una variable aleatoria discreta, X, que toma los valores $x_1,\ x_2,\ \ldots,\ x_n\ (\ldots)$, con probabilidades $p_j = P(X = x_j),\ j = 1, 2, \ldots, n\ (\ldots)$. Un planteamiento estándar, equivalente al anterior, consiste en resolver la cuestión de simular la variable aleatoria I que toma los valores $1, 2, \ldots, n\ (\ldots)$ con las mismas probabilidades $p_j,\ j = 1, 2, \ldots, n\ (\ldots)$. Esto es consecuencia del uso de una aplicación inyectiva llamada de codificación o etiquetado, dada por,

$$\begin{aligned} l &: B \to \mathbb{N}, \\ \text{con } l(x_i) &= i, \end{aligned}$$

siendo B el conjunto de valores que toma la variable aleatoria X.

4.1 El método de la transformación cuantil

Este método es una adaptación del método de inversión (válido para el caso continuo) a distribuciones discretas. En primer lugar veamos porqué el método de inversión no es aplicable directamente en este caso.

Dada una variable aletoria discreta, su función de distribución viene dada por

$$F(x) = \sum_{x_j \le x} p_j,\ \forall x \in \mathbb{R}.$$

Se supondrá (por comodidad) que los valores que toma la variable ya están ordenados y nos ceñiremos al caso finito. En caso contrario se ordenarían denotando x_1 el menor, x_2 el segundo y así sucesivamente. De esta forma tendríamos: $x_1 < x_2 < \cdots < x_n$. En este caso es obvio que el resultado dado por el teorema de inversión no es cierto ya que la variable aleatoria $F(X)$ toma sólo los valores $F(p_1), F(p_1 + p_2), \ldots, F(p_1 + p_2 + \cdots + p_n)$ siendo, por tanto, discreta y no pudiendo tener distribución $U(0, 1)$.

Ejercicio 4.1.1 *Determinar la distribución de* $F(X)$.

De la misma forma, dada una variable $U \sim U(0,1)$, tampoco puede ser cierto que $F^{-1}(U)$ tenga la misma distribución que X. De hecho F^{-1} no está definida de forma única pues las funciones de distribución discretas no tienen inversa, pues para casi todo $u \in [0,1]$ no hay ningún x tal que $F(x) = u$ y para un número finito (o infinito numerable) de $u \in [0,1]$ se tiene que existe todo un intervalo de valores para x cumpliendo $F(x) = u$. A pesar de ello puede definirse la llamada función cuantil (o inversa generalidada) de una distribución cualquiera F a partir de

$$Q(u) = \inf\{x \in \mathbb{R}/F(x) \geq u\}, \ \forall u \in (0,1).$$

Es obvio que esta función siempre está definida y que cuando F sea invertible, $Q = F^{-1}$.

El siguiente teorema da un resultado que generaliza al teorema de inversión a situaciones en las que F no es invertible.

Teorema 13 (de inversión generalizada) *Sea X una variable aleatoria con función de distribución F y con función cuantil Q. Considérese una variable aleatoria, U, con distribución $U(0,1)$, entonces, la variable $Q(U)$ tiene la misma distribución que X.*

Demostración: Sea G la función de distribución de $Q(U)$. Dado $x \in \mathbb{R}$, se tiene

$$\begin{aligned} G(x) &= P(Q(U) \leq x) = P(\inf\{y \in \mathbb{R}/F(y) \geq U\} \leq x) \\ &= P(F(x) \geq U) = \int_0^{F(x)} du = F(x). \end{aligned}$$

A partir del teorema de inversión generalizada puede obtenerse un algoritmo general para simular cualquier distribución de probabilidad discreta. Es el llamado algoritmo de transformación cuantil o de inversión generalizada.

Algortimo de transformación cuantil

1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = Q(U)$.

La mayor dificultad en la implementación del algoritmo radica en el cálculo de

$$\begin{aligned} Q(U) &= \inf\{x \in \mathbb{R}/F(x) \geq U\} = \inf\left\{x_j \Big/ \sum_{i=1}^{j} p_i \geq U\right\} \\ &= x_k, \text{ tal que } \sum_{i=1}^{k} p_i \geq U > \sum_{i=1}^{k-1} p_i. \end{aligned}$$

Todo el problema radica, por tanto, en encontrar el valor, k, de la variable, I, que guarda las etiquetas, para el cual la función de distribución supera o

iguala por primera vez al valor de U. Este valor puede hallarse mediante una búsqueda secuencial, utilizando el siguiente algoritmo:

Algortimo de transformación cuantil con búsqueda secuencial

1. `Generar` $U \sim U(0,1)$.
2. `Hacer` $I = 1$ `y` $S = p_1$.
3. `Mientras` $U > S$ `hacer`

3.1. $I = I + 1$ `y` $S = S + p_I$.

4. `Devolver` $X = x_I$.

Si se desea generar un gran número de valores de la variable X (que es lo más habitual) puede resultar más eficiente calcular previamente las cantidades $S_j = \sum_{i=1}^{j} p_j$ de forma recursiva: $S_1 = p_1$, $S_j = S_{j-1} + p_j$ para $j = 2, 3, \ldots, n$ y hacer la comparación $U > S_I$ en el paso 3 del algoritmo anterior. De esta forma se evita lo que podrían ser cálculos repetitivos de las mismas sumas de probabilidades al simular distintos valores de X.

Ejemplo 14 (Simulación de la distribución de Poisson) *Tómese una variable, X, con distribución de Poisson de parámetro λ, que toma los valores $x_1 = 0$, $x_2 = 1$, ... con probabilidades*

$$p_j = P(X = x_j) = P(X = j-1) = \frac{e^{-\lambda}\lambda^{j-1}}{(j-1)!},\ j = 1, 2, \ldots$$

En este caso el algoritmo de inversión con búsqueda secuencial viene dado por

1. `Generar` $U \sim U(0,1)$.
2. `Hacer` $I = 1$ `y` $S = e^{-\lambda}$.
3. `Mientras` $U > S$ `hacer`

3.1. $I = I + 1$ `y` $S = S + \frac{e^{-\lambda}\lambda^{I-1}}{(I-1)!}$.

4. `Devolver` $X = I - 1$.

Debido a que la forma de etiquetar los valores de la variable conlleva el desfase de una unidad en los índices, es recomendable ajustar el algoritmo para evitar este efecto:

1. `Generar` $U \sim U(0,1)$.
2. `Hacer` $I = 0$ `y` $S = e^{-\lambda}$.
3. `Mientras` $U > S$ `hacer`

3.1. $I = I + 1$ `y` $S = S + \frac{e^{-\lambda}\lambda^{I}}{I!}$.

4. `Devolver` $X = I$.

Por último, teniendo en cuenta que las probabilidades pueden calcularse de forma recursiva

$$P(X = j) = \frac{e^{-\lambda}\lambda^{j}}{j!} = \frac{\lambda}{j}\frac{e^{-\lambda}\lambda^{j-1}}{(j-1)!} = \frac{\lambda}{j}P(X = j-1),$$

se pueden simplificar los cálculos que aparecen en el paso 3.1 del algoritmo dando lugar a

1. `Generar` $U \sim U(0,1)$.
2. `Hacer` $I = 0$, $p = e^{-\lambda}$ `y` $S = p$.

3. Mientras $U > S$ *hacer*
3.1. $I = I + 1$, $p = \frac{\lambda}{I} p$ *y* $S = S + p$.
4. Devolver $X = I$.

4.1.1 Eficiencia del algoritmo

Dada la forma del algoritmo general para simular una distribución discreta mediante el método de la transformación cuantil utilizando búsqueda secuencial, es fácil probar que el número de comprobaciones de la forma $U > S$ es precisamente igual a I, el valor de la variable que contiene las etiquetas. Como el valor de I es aleatorio y variará con cada ejecución del algortimo, una medida de la eficiencia del mismo será el número medio de comparaciones del paso 3, es decir,

$$E(I) = \begin{cases} \sum_{j=1}^{n} j p_j & \text{si } X \text{ toma un número finito } (n) \text{ de valores} \\ \sum_{j=1}^{\infty} j p_j & \text{si } X \text{ toma un infinitos valores} \end{cases}$$

Resulta pues evidente que, como no existe una única forma de etiquetar los valores que toma la variable en cuestión, habrá quizá algún etiquetado que ofrezca un menor número medio de comparaciones en el paso 3 del algoritmo que el etiquetado original (que obedece a la idea de ordenar de forma creciente los valores que toma la variable).

Es obvio que los demás etiquetados no serán equivalentes al método de transformación cuantil aplicado directamente sobre la variable X, sinó más bien consistirán en una fase previa en la que se etiquetarán (de la forma más conveniente) los posibles valores de la variable X, dando lugar a una variable de etiquetas, I, y luego se simularán valores de dicha variable para posteriormente recuperar los valores de X mediante l^{-1}, la inversa del etiquetado o función de decodificación.

Ejemplo 15 *Considérese la variable aleatoria discreta* X *con distribución dada por*

$$P(X = 3) = 0.1,\ P(X = 5) = 0.3,\ P(X = 7) = 0.6$$

Definiendo

$$\begin{aligned} l &: \{3, 5, 7\} \to \{1, 2, 3\} \\ l(3) &= 1,\ l(5) = 2,\ l(7) = 3 \\ (\textit{i.e. } x_1 &= 3,\ x_2 = 5,\ x_3 = 7) \end{aligned}$$

se tiene un etiquetado I *con distribución*

$$P(I = 1) = 0.1,\ P(I = 2) = 0.3,\ P(I = 3) = 0.6$$

y, por tanto, con media $E(I) = 1 \cdot 0.1 + 2 \cdot 0.3 + 3 \cdot 0.6 = 2.5$.

Si, por el contrario, definimos el etiquetado

$$\begin{aligned} h &: \{3,5,7\} \to \{1,2,3\} \\ h(3) &= 3,\ l(5) = 2,\ l(7) = 1 \\ (i.e.\ x'_1 &= 7,\ x'_2 = 5,\ x'_3 = 3) \end{aligned}$$

se tiene que

$$P(I' = 1) = 0.6,\ P(I' = 2) = 0.3,\ P(I' = 3) = 0.1$$

y así $E(I') = 1 \cdot 0.6 + 2 \cdot 0.3 + 3 \cdot 0.1 = 1.5$. *Se observa, por tanto, como* $E(I')$ *es sensiblemente inferior a* $E(I)$ *y, por lo tanto, como el segundo etiquetado, dado por* h*, proporciona un algoritmo más eficiente que el dado por el etiquetado original,* l.

Como parece deducirse del ejemplo anterior, un etiquetado será tanto mejor cuanto menores sean las etiquetas que se dan a los valores que tienen mayor probabilidad. Así, dada la variable correspondiente al etiquetado original, I, con masa de probabilidad $p_1, p_2, \ldots, p_n$, el mejor etiquetado es el dado por $l(i) = n + 1 - \text{rango}(p_i)$, siendo el rango de p_i el entero que indica el lugar que ocupa p_i en una ordenación conjunta de los n valores $p_1, p_2, \ldots, p_n$. Dicho de otra forma, el etiquetado l otorga el índice 1 al entero i con mayor probabilidad, p_i; el índice 2 al de segunda mayor probabilidad y así sucesivamente.

La propiedad anterior resulta de fácil demostración. Supóngase una variable, I, asociada a un etiquetado distinto del l, recién definido. Es decir, se trata de un etiquetado h verificando que existen un par de índices j y k tales que $h(k) < h(j)$ y además $p_k < p_j$. A partir de este etiquetado se define otro nuevo, g, de la forma

$$g(i) = \begin{cases} h(i) & \text{si } i \neq j \text{ e } i \neq k \\ h(k) & \text{si } i = j \\ h(j) & \text{si } i = k \end{cases}$$

Probaremos muy fácilmente que el número medio de comparaciones asociado a este etiquetado es menor que el del de partida:

$$\begin{aligned} E(I_g) &= \sum_{i=1}^{n} g(i)\, p_i = \sum_{i=1, i\neq j, i\neq k}^{n} g(i)\, p_i + g(j)\, p_j + g(k)\, p_k \\ &= \sum_{i=1, i\neq j, i\neq k}^{n} h(i)\, p_i + h(k)\, p_j + h(j)\, p_k \\ &= \sum_{i=1}^{n} h(i)\, p_i + h(k)\, p_j + h(j)\, p_k - h(j)\, p_j - h(k)\, p_k \\ &= E(I_h) + (h(k) - h(j))\,(p_j - p_k) < E(I_h)\,. \end{aligned}$$

Con esto se demuestra que un etiquetado que no proporcionase un orden decreciente en las probabilidades podría ser mejorado (en términos de hacer mínimo el número medio de comparaciones del algoritmo) por otro nuevo. Como consecuencia, los etiquetados óptimos serán aquellos que provoquen ordenaciones decrecientes sobre las probabilidades, es decir,

$$p_{l^{-1}(1)} \geq p_{l^{-1}(2)} \geq \cdots \geq p_{l^{-1}(n)}.$$

Obviamente pueden existir varios etiquetados óptimos si hay empates entre algunas de estas probabilidades.

Cuando la variable a simular tiene un número finito de valores: x_1, x_2, $\ldots$, x_n, al implementar el método de la transformación cuantil con búsqueda secuencia directa, una vez comprobado que $U > \sum_{j=1}^{n-1} p_j$, no es necesario comprobar $U > \sum_{j=1}^{n} p_j = 1$ (que siempre es falso), sinó que generamos x_n sin necesidad de efectuar esa comparación. Por ese motivo el número medio de comparaciones sería realmente:

$$\sum_{j=1}^{n-1} j p_j + (n-1)\, p_n.$$

Todo lo dicho anteriormente acerca de la elección óptima del etiquetado sigue siendo válido salvo permutaciones entre los valores con las dos últimas etiquetas.

Ejemplo 16 *Consideremos la variable aleatoria discreta con distibución*

$$\begin{aligned} P(X=1) &= 0.11,\ P(X=3)=0.3,\ P(X=5)=0.25, \\ P(X=7) &= 0.21,\ P(X=9)=0.13. \end{aligned}$$

El conjunto de valores que toma la variable es $B = \{1,3,5,7,9\}$ *y la función de etiquetado asociada al algoritmo de inversión generalizada en su versión clásica es*

$$h : B \longrightarrow \mathbb{N}$$

dada por

$$\begin{aligned} h(1) &= 1,\ h(3)=2,\ h(5)=3, \\ h(7) &= 4,\ h(9)=5. \end{aligned}$$

El número de comparaciones en el algoritmo asociado a este etiquetado es una variable discreta, I_h*, con distribución*

$$\begin{aligned} P(I_h=1) &= 0.11,\ P(I_h=2)=0.3,\ P(I_h=3)=0.25, \\ P(I_h=4) &= 0.21,\ P(I_h=5)=0.13 \end{aligned}$$

Por tanto, el número medio de comparaciones del algoritmo es

$$E(I_h) = 0.11 \cdot 1 + 0.3 \cdot 2 + 0.25 \cdot 3 + (0.21 + 0.13) \cdot 4 = 2.82$$

Un etiquetado óptimo sería

$$l : B \longrightarrow \mathbb{N}$$

dado por

$$\begin{aligned} l(1) &= 5,\ l(3) = 1,\ l(5) = 2, \\ l(7) &= 3,\ l(9) = 4. \end{aligned}$$

cuyo número medio de comparaciones viene dado por

$$E(I_l) = 0.3 \cdot 1 + 0.25 \cdot 2 + 0.21 \cdot 3 + (0.13 + 0.11) \cdot 4 = 2.39$$

que mejora sensiblemente el valor del método de transformación cuantil estándar. Este etiquetado corresponde a la notación $x_1 = 3$, $x_2 = 5$, $x_3 = 7$, $x_4 = 9$ *y* $x_5 = 1$.

4.1.2 Cálculo directo de la función cuantil

En ocasiones el método de la transformación cuantil puede acelerarse computacionalmente porque, mediante cálculos directos, es posible encontrar el valor de la función cuantil en cualquier U, en un tiempo de computación mínimo (evitando el bucle de búsqueda en el que se van acumulando las probabilidades). Veamos algunos ejemplos.

Ejemplo 17 (la distribución uniforme discreta en $\{1, 2, \ldots, n\}$**)** *En este caso la masa de probabilidad viene dada por*

$$p_j = \frac{1}{n},\ \textit{para}\ j = 1, 2, \ldots n.$$

De esta forma se tiene

$$\sum_{i=1}^{k} p_i \geq U > \sum_{i=1}^{k-1} p_i \Leftrightarrow \frac{k}{n} \geq U > \frac{k-1}{n} \Leftrightarrow k \geq nU > k - 1.$$

Salvo para un conjunto finito de posibles valores (para $U \in \{0, \frac{1}{n}, \cdots, \frac{n-1}{n}\}$*), esta última condición equivale a* $k > nU \geq k - 1$ *que, a su vez, es lo mismo que* $k = \lceil nU \rceil + 1$*, siendo* $\lceil x \rceil$ *la parte entera de* x.

De esta forma, el algoritmo resulta:

1. **Generar** $U \sim U(0, 1)$.
2. **Devolver** $X = \lceil nU \rceil + 1$.

que, obviamente, es un algoritmo enórmemente eficiente.

Ejemplo 18 (la distribución geométrica) *La distribución geométrica representa el número de fracasos antes del primer éxito y tiene la siguiente masa de probabilidad*

$$P(X=j) = p(1-p)^j,\ j = 0, 1, \ldots$$

Para un valor j entero no negativo su función de distribución viene dada por

$$F(j) = \sum_{i=0}^{j} p(1-p)^i = \frac{p(1-p)^{j+1} - p}{1-p-1} = 1 - (1-p)^{j+1}.$$

Como consecuencia se tiene

$$\begin{aligned} F(k) \geq U > F(k-1) &\Leftrightarrow 1-(1-p)^{k+1} \geq U > 1-(1-p)^k \\ &\Leftrightarrow (1-p)^k > 1-U \geq (1-p)^{k+1} \\ &\Leftrightarrow k\ln(1-p) > \ln(1-U) \geq (k+1)\ln(1-p) \\ &\Leftrightarrow k < \frac{\ln(1-U)}{\ln(1-p)} \leq k+1 \end{aligned}$$

condición esta última que, salvo para un número infiniro pero numerable de valores de U (de probabilidad cero), equivale a

$$k = \left\lceil \frac{\ln(1-U)}{\ln(1-p)} \right\rceil.$$

El algoritmo procedería de la siguiente forma:

1. *Generar* $U \sim U(0,1)$.
2. *Devolver* $X = \left\lceil \frac{\ln(1-U)}{\ln(1-p)} \right\rceil$.

o, ahorrando operaciones,

0. *Hacer* $a = 1/\ln(1-p)$.
1. *Generar* $U \sim U(0,1)$.
2. *Devolver* $X = \lceil a \ln U \rceil$.
3. *Repetir los pasos 1-2 tantas veces como sea necesario.*

4.2 Algoritmos basados en árboles binarios. Árboles de Huffman.

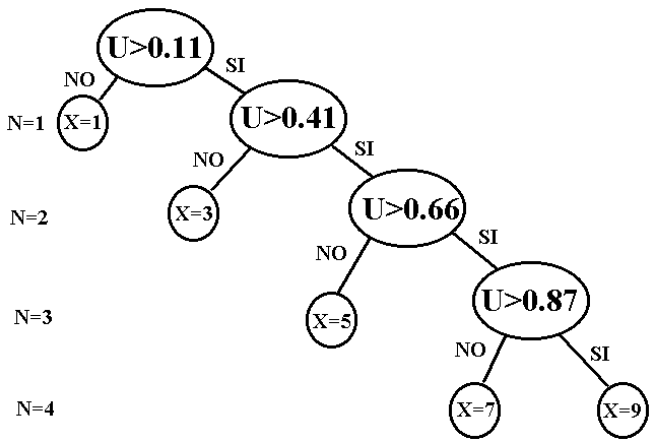
En los algoritmos de búsqueda secuencial existe un paso en el que, en función del resultado de una comparación ($U > \sum_{j=1}^{i} p_j$) se decide si el valor a generar para X va a ser uno concreto ($X = x_i$) o seguiremos buscando entre los restantes ($X \in \{x_j / j > i\}$). Esta forma de proceder lleva a la elección de un único valor si la condición $U > \sum_{j=1}^{i} p_j$ es falsa.

Una forma más general de proceder consiste en comparar (en cada paso) el valor de U con cierta suma de probabilidades y elegir entre dos conjuntos de valores posibles (no necesariamente de un único elemento ninguno de ellos) según el resultado de la comparación.

Ejemplo 19 *Considérese la variable aleatoria X, dada por*

$$\begin{aligned} P(X=1) &= 0.11,\ P(X=3)=0.3,\ P(X=5)=0.25, \\ P(X=7) &= 0.21,\ P(X=9)=0.13. \end{aligned}$$

Es decir, denotando $x_1=1$, $x_2=3$, $x_3=5$, $x_4=7$, $x_5=9$, se tiene que $p_1=0.11$, $p_2=0.3$, $p_3=0.25$, $p_4=0.21$, $p_5=0.13$. El algoritmo de búsqueda secuencial tiene una estructura dada por la figura adjunta.
Donde N indica el número de comparaciones para obtener uno de los valores simulados. En este caso $E(N)=1\cdot 0.11+2\cdot 0.3+3\cdot 0.25+4\cdot(0.21+0.13)=2.82$. Además, resulta fácil ver que una versión del algoritmo en la que se etiquetasen los valores de forma que las probabilidades resultasen ordenadas de forma decreciente ofrecería un número medio de comparaciones igual a $E(N')=1\cdot 0.3+2\cdot 0.25+3\cdot 0.21+4\cdot(0.13+0.11)=2.39$.



Representación gráfica del algoritmo en forma de árbol de Huffman.

La pregunta que se plantea en el ejemplo anterior (y en general) es ¿no existiría algún otro tipo de búsqueda en forma de árbol binario (aunque con nodos hijos por la izquierda que no tengan que ser necesariamente nodos terminales u hojas del árbol) que produzca un menor número medio de comparaciones? El problema puede plantearse de forma equivalente en términos de árboles binarios.

Un árbol no es más que un grafo orientado en el que existe un nodo singular llamado raíz (o nodo origen) del cual salen arcos pero al cual no llegan arcos. Los nodos a los que llegan arcos procedentes de otro nodo son llamados hijos de dicho nodo. En nuestro caso consideraremos que cada nodo, excepto el

raíz, es hijo de exactamente otro nodo. Los nodos de los que no salen arcos (o que no tienen hijos) son llamados nodos terminales u hojas del árbol. En estas condiciones si el grafo es finito (i.e. si tiene un número finito de nodos) todo nodo desciende del raíz en un número finito de generaciones (es decir, es hijo de un hijo de un hijo ... de un hijo del raíz). Dicho número es llamado profundidad del nodo. Aquellos árboles que cumplen que todos los nodos, excepto los terminales, tienen exactamente dos hijos se llaman árboles binarios.

El problema que nos preocupa, pensado en términos de árboles binarios, consiste en dados n nodos $\{1, 2, \ldots, n\}$ con probabilidades respectivas p_1, p_2, $\ldots$, p_n, encontrar un árbol binario que tiene precisamente nodos terminales $\{1, 2, \ldots, n\}$ y profundidades respectivas d_1, d_2, $\ldots$, d_n, de tal forma que se minimice la función

$$\sum_{j=1}^{n} d_j p_j.$$

Con esta formulación, las probabilidades otorgadas a los nodos terminales son precisamente $p_j = P(X = x_j)$.

Al objeto de construir dicho árbol (o uno de ellos si hay varios que minimizan la expresión), resulta fácil percatarse de que ha de cumplir las siguientes propiedades:

1. Existen al menos dos nodos terminales de profundida máxima. Es obvio que el nodo de máxima profundidad es terminal. Por ser un árbol binario, ese nodo ha de tener un hermano que es de la misma profundidad y, por tanto, también terminal.

2. Los dos nodos de probabilidad más pequeña del árbol óptimo son terminales y de profundidad máxima. De no ser así habría un nodo de máxima profundidad con probabilidad mayor que la de otro nodo terminal con menor profundidad. Permutando ambos nodos encontraríamos otro nuevo árbol en el que la función del objetivo sería menor.

3. Siempre existe algún árbol óptimo en el que los dos nodos de más pequeña probabilidad son hermanos (y ya sabemos que de profundidad máxima). Basta considerar un árbol óptimo (en el que ambos nodos han de tener máxima profundidad -en particular igual-) y en caso de que los nodos en cuestión no fuesen hermanos realizar una permutación entre uno de ellos y el hermano del otro. Este proceso no altera la función del objetivo y, por tanto, el nuevo árbol construído, en el que esos dos nodos ya sí son hermanos, también es óptimo.

4. Fusionando los dos nodos a los que se refiere el apartado anterior (por comodidad notacional supongamos que son los nodos con etiquetas $n-1$ y n) con su nodo padre y otorgándole a éste la suma de sus probabilidades $(p_{n-1} + p_n)$, el árbol resultante es óptimo entre los árboles binarios

factibles para el problema con un nodo terminal menos y con probabilidades $p_1, p_2, \ldots, p_{n-2}, p_{n-1}+p_n$. Dado que en el proceso de "fusión" de estos nodos, lo que se hace es reducir la función del objetivo en el valor $p_{n-1}+p_n$, el suponer que este árbol reducido no fuese óptimo dentro de los de $n-1$ nodos terminales, implicaría la existencia de otro mejor que él, el cual podría ser "expandido" -mediante un proceso justamente contrario al anterior- a un árbol con n nodos terminales (donde el $n-1$ y el n ya han sido separados) a costa de aumentar la función del objetivo en exactamente $p_{n-1}+p_n$. Este árbol presentaría un valor de la función del objetivo menor que el de partida, lo cual contradice la optimalidad de aquél.

Siguiendo las pautas anteriores podemos construír un árbol óptimo de la siguiente forma:

1. Agrupamos los dos nodos con probabilidades más pequeñas en un sólo nodo con probabilidad igual a la suma de las de ambos. En caso de empates los deshacemos arbitrariamente.

2. En el árbol resultante (con un nodo menos) procedemos como en el paso 1, repitiendo esto hasta finalizar con un problema para árboles de sólo dos nodos terminales, cuya solución es trivial.

El árbol binario óptimo así construído se denomina árbol de Huffman.

Ejemplo 20 *Retomemos el ejemplo anterior. En principio tenemos tantos nodos como posibles valores de la variable (cinco) con sus respectivas probabilidades*

nodos	1	3	5	7	9
probabilidades	**0.11**	0.3	0.25	0.21	**0.13**

Así, se crea el nodo $\{1,9\}$ *con probabilidad asociada* $p_{\{1,9\}} = p_1 + p_9 = 0.24$. *De esta forma se va reduciendo un nodo en cada etapa:*

nodos	$\{1,9\}$	3	5	7
probabilidades	**0.24**	0.3	0.25	**0.21**

nodos	$\{1,7,9\}$	3	5
probabilidades	0.45	**0.3**	**0.25**

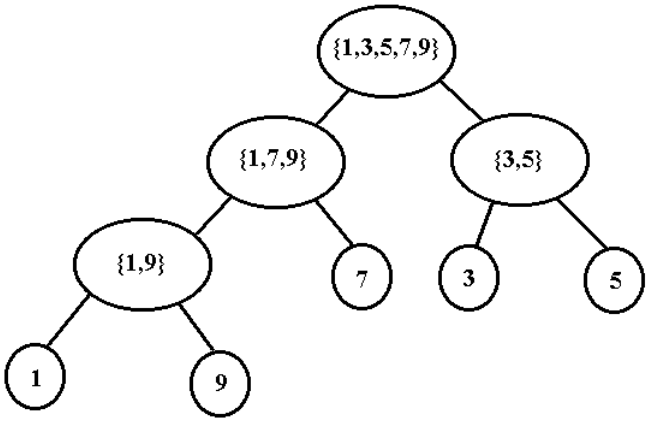
nodos	$\{1,7,9\}$	$\{3,5\}$
probabilidades	0.45	0.55

De esta forma, el árbol de Huffman viene dado por la primera figura adjunta.

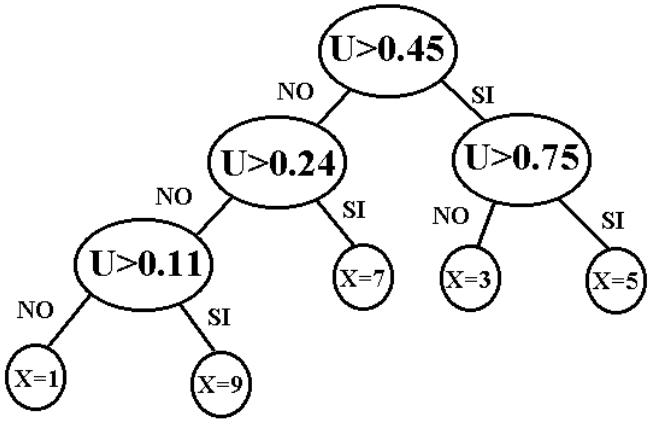
Con lo cual el algoritmo podría esquematizarse de la forma indicada en la segunda figura.

En forma más detallada puede escribirse así:

1. Generar $U \sim U(0,1)$.
2. Si $U > 0.45$ *entonces*
2.1. Empezar
2.2. Si $U > 0.75$ *entonces hacer* $X = 5$, *en otro caso hacer* $X = 3$.
2.3. Terminar.
En otro caso.
2.4. Empezar
2.5. Si $U > 0.24$ *entonces hacer* $X = 7$, *en otro caso*
2.5.1. Empezar
2.5.2. Si $U > 0.11$ *entonces hacer* $X = 9$, *sino hacer* $X = 1$.
2.5.3. Terminar.
2.6. Terminar.



Árbol de Huffman.



Esquema del algoritmo.

Un algoritmo general para construir el árbol de Huffman puede basarse en usar una estructura de punteros.

Supónganse las probabilidades ordenadas de forma decreciente: $p_1 \geq p_2 \geq \cdots \geq p_n$, lo cual puede hacerse eficientemente usando algoritmos como el quicksort. Sean x_1, x_2, ..., x_n los valores que toma la variable ya arrastrando la reordenación de las probabilidades. El algortimo procedería del siguiente modo:

1. Crear una estructura de punteros de la forma

$$\begin{aligned} n &\rightarrow n+1 \rightarrow \cdots \rightarrow 2n-2 \rightarrow 2n-1, \\ p[n] &= p_1,\, p[n+1]=p_2,\, \ldots,\, p[2n-1]=p_n. \end{aligned}$$

2. Desde $i=1$ hasta $n-1$ hacer

2.1. Eliminar los dos últimos elementos del puntero: j y k.

2.2. Hacer Left[i]=j y Right[i]=k.

2.3. Insertar i en el puntero en orden decreciente de las probabilidades y hacer $p[i]=p+q$.

Finalizado el algoritmo tenemos los valores $(i, p[i])$, para $i = 1, 2, \ldots, n-1, n, n+1, \ldots, 2n-1$, donde los índices $\{1, 2, \ldots, n-1\}$ representan los nodos internos y $\{n, n+1, \ldots, 2n-1\}$ son los nodos terminales, u hojas, almacenando en los vectores Letf[i] y Right[i] la estructura del árbol de Huffman.

Para simular valores mediante el método basado en el árbol de Huffman podemos proceder como sigue:

1. Generar $U \sim U(0,1)$.
2. Hacer $i = n-1$.
3. Repetir

3.1. l =Left[i], r =Right[i], p =p[l].

3.2. Si $U > p$ entonces hacer

3.2.1. $i = r$, $U = U - p$.

En otro caso hacer

3.2.2. $i = l$.

4. Hasta que $i \geq n$.
5. Hacer $X = x_{i-n+1}$.

4.3 El método de la tabla guía

El mayor problema computacional del método de la transformación cuantil consiste en encontrar el índice k que cumple $\sum_{i=1}^{k} p_i \geq U > \sum_{i=1}^{k-1} p_i$. Como ya se ha visto en los dos últimos ejemplos existen distribuciones para las cuales este valor k se puede calcular directamente. El método de la tabla guía consiste en hacer uso de la rapidez de cálculo de la función cuantil para alguna de esas distribuciones (fácilmente simulable mediante el método de inversión generalizada) para abreviar al máximo el número de comparaciones necesarias a la hora de comprobar la condición $\sum_{i=1}^{k} p_i \geq U > \sum_{i=1}^{k-1} p_i$.

Considérese una variable aleatoria con distribución discreta y masa de probabilidad dada por p_j, $j = 1, 2, \ldots, n$ y defínanse las sumas acumulativas de estas probabilidades (que no son otra cosa que los valores que toma la función de distribución): $q_j = \sum_{i=1}^{j} p_i$. Como ya se comentó anteriormente, estos valores deben calcularse de forma recursiva (para evitar cálculos innecesarios): $q_0 = 0$, $q_j = q_{j-1} + p_j$, $j = 1, 2, \ldots, n$. Dada la variable aleatoria I, asociada al etiquetado original (o a otro) la idea del método consiste en construir n subintervalos equiespaciados contenidos en $[0, 1]$ de la forma $J_i = [\frac{i-1}{n}, \frac{i}{n})$ para $i = 1, 2, \ldots, n$ y luego definir

$$\begin{aligned} g_i &= \max\left\{ j \,\middle/\, q_j < \frac{i}{n} \right\}, \text{ para } i = 1, 2, \ldots, n, \\ \text{tomándose } g_i &= 0 \text{ si el conjunto anterior es vacío} \end{aligned}$$

es decir, para cada intervalo se considera el valor más alto del índice entero tal que la suma acumulada de probabilidades hasta él es menor que el extremo superior de dicho intervalo.

Ejemplo 21 *Tomemos como ejemplo la distribución discreta dada por* $p_1 = 0.13$, $p_2 = 0.25$, $p_3 = 0.17$, $p_4 = 0.1$, $p_5 = 0.24$ *y* $p_6 = 0.11$. *Se tiene que* $q_1 = 0.13$, $q_2 = 0.38$, $q_3 = 0.55$, $q_4 = 0.65$, $q_5 = 0.89$ *y* $q_6 = 1$. *Los valores de la tabla guía son*

$$g_1 = 1,\ g_2 = 1,\ g_3 = 2,\ g_4 = 4,\ g_5 = 4,\ g_6 = 5.$$

A la hora de aplicar el método de la transformación cuantil, dado el valor de U, es inmediato detectar en cuál de los intervalos J_i ha caído, basta con hacer $i = \lceil nU \rceil + 1$. Lo único que resta por hacer, una vez encontrado este índice, es obtener el valor del índice I a simular. Dicho valor será $g_i + 1$ si ya ocurre que $U > q_{g_i}$. En caso contrario deberemos encontrar el primer índice $j = g_i - 1, g_i - 2, \ldots, 0$, para el cual se cumple $U > q_j$ y luego hacer $I = j + 1$.

Algoritmo de simulación mediante una tabla guía

1. `Generar` $U \sim U(0, 1)$.
2. `Hacer` $i = \lceil nU \rceil + 1$.
3. `Hacer` $j = g_i$.
4. `Mientras` $U \leq q_j$ `hacer` $j = j - 1$.
5. `Devolver` $I = j + 1$.

Por su parte, los valores de la tabla guía pueden calcularse fácilmente de forma rápida según el siguiente algoritmo:

Algoritmo de cálculo de la tabla guía

1. `Desde` $i = 1$ `hasta` $n - 1$ `hacer` $g_i = 0$.
2. `Hacer` $S = 0$.
3. `Desde` $i = 1$ `hasta` $n - 1$ `hacer`

3.1. $S = S + p_i$

3.2. $j = \lceil nS \rceil + 1$

3.3. $g_j = i$

4. `Desde` $i = 2$ `hasta` $n-1$ `hacer` $g_i = \max(g_{i-1}, g_i)$.

Aunque los bucles de los pasos 1 y 4 del algoritmo pueden extenderse hasta n (en lugar de hasta $n-1$) en la práctica resulta innecesario inicializando $g_n = n-1$, conocido de antemano.

Cuando el valor U cae en el intervalo J_i, es obvio que el número medio de comparaciones en el paso 4 del algoritmo es menor o igual que 1 más el número de valores q_j pertenecientes al intervalo J_i. Utilizando este hecho, la esperanza del número de comparaciones (N) puede acotarse mediante

$$\begin{aligned} E(N) &\leq \frac{1}{n}\sum_{i=1}^{n}\left(1+\#\left\{j\,/q_j \in J_i\right\}\right) = 1+\frac{1}{n}\sum_{i=1}^{n}\#\left\{j\,/q_j \in J_i\right\} \\ &= 1+\frac{1}{n}\#\left\{j\,/q_j \in [0,1)\right\} = 1+\frac{n-1}{n} < 2. \end{aligned}$$

En general, el método es aplicable para tablas guía de m elementos (donde m no tiene porqué ser necesariamente igual a n). En tal caso el intervalo $[0,1)$ se divide en m subintervalos, pudiendo acotar el número medio de comparaciones mediante $E(N) \leq 1+\frac{n}{m}$. Gracias a este argumento, para variables con un número exorbitante de posibles valores, pueden utilizarse tablas guía de un número más moderado de elementos de forma que la tabla no ocupe demasiada memoria y que, a la vez, el número medio de comparaciones esté acotado por un valor moderado. Así, por ejemplo, para una variable discreta con 1000000 de posibles valores podríamos utilizar una tabla guía de sólo 10000 elementos (para que no ocupe demasiado en memoria) obteniendo que el número medio de comparaciones estaría acotado por 101.

4.4 Métodos de truncamiento

La idea general de este tipo de métodos consiste en hacer uso de una distribución continua auxiliar cuya función de distribución se parezca (en cierto sentido que se precisará más adelante) a la función de distribución de la variable discreta que se desea simular.

Supóngase, sin pérdida de generalidad, que se desea simular la variable I, que toma los valores 1, 2, ..., n, con probabilidades p_1, p_2, ..., p_n. En este caso, la función de distribución de I viene dada por

$$F(x) = \sum_{i \leq x} p_i.$$

Supóngase, además, que tenemos otra variable aleatoria continua, con función de distribución $G(x)$ y ciertos valores $a_0 = -\infty < a_1 < a_2 < \cdots < a_{n-1} < a_n = \infty$, tales que $F(i) - F(i^-) = p_i = G(a_i) - G(a_{i-1})$, $i = 1, 2, \ldots, n$. Esta última condición viene a garantizar que la probabilidad de que la variable

continua caiga en el intervalo $[a_{i-1}, a_i)$ coincide con la probabilidad con la que la variable discreta original tome el valor i.

Si la distribución continua es fácil de simular, simplemente deberemos generar valores de la misma y luego transformarlos en valores de la variable I.

Algoritmo de simulación por truncamiento

1. Generar T con distribución G.
2. Encontrar el valor i tal que $a_{i-1} \leq T < a_i$.
3. Devolver $I = i$.

El método se hace especialmente rápido cuando el valor de i puede obtenerse de forma inmediata a partir del valor de T. Uno de los casos en los que esto es así se da cuando $G(0) = 0$ y los valores $a_i = i$, $i = 0, 1, \ldots, n$ (o, incluso, infinitos valores a_i de esta forma). En este caso el algoritmo resulta:

Algoritmo de simulación por truncamiento a la parte entera

1. Generar T con distribución G.
2. Hacer $i = \lceil T \rceil + 1$.
3. Devolver $I = i$.

Ejemplo 22 (simulación de la geométrica por truncamiento) *La masa de probabilidad de la distribución geométrica es*

$$P(X = j) = P(I = j + 1) = p(1-p)^j,\ j = 0, 1, \ldots$$

Considérese como variable aleatoria continua auxiliar la exponencial, que tiene función de distribución dada por

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

Ahora bien,

$$\begin{aligned} G(i) - G(i-1) &= 1 - e^{-\lambda i} - \left(1 - e^{-\lambda(i-1)}\right) = e^{-\lambda(i-1)} - e^{-\lambda i} \\ &= e^{-\lambda(i-1)}\left(1 - e^{-\lambda}\right) = \left(1 - e^{-\lambda}\right)\left(e^{-\lambda}\right)^{i-1} = p(1-p)^{i-1} \end{aligned}$$

siempre que tomemos $p = 1 - e^{-\lambda}$. De esta forma se tiene

$$G(i) - G(i-1) = P(X = i-1) = P(I = i) = p_i$$

y el algoritmo resultaría:

0. Hacer $L = \ln(1-p)$.
1. Generar $U \sim U(0,1)$.
2. Hacer $T = L \cdot \ln U$.
3. Devolver $X = \lceil T \rceil$.

Este es el mismo algoritmo que se obtenía anteriormente cuando razonábamos cómo calcular directamente el valor de la función cuantil para la distribución geométrica.

4.5 Ejercicios propuestos

1. Cuando se produce una sobrecarga eléctrica en una cadena productiva el número de componentes averiadas está descrito mediante una variable aleatoria, X, que puede tomar los valores 1, 2, 3, 4 y 5, con probabilidades

$$p_1 = 0.11,\ p_2 = 0.22,\ p_3 = 0.45,\ p_4 = 0.13,\ p_5 = 0.09.$$

Para simular dicha variable se usa el algoritmo:

```
1.  Generar U, con distribucion U(0,1).
2.1 Si U<=0.11 entonces hacer X=1, sino
 2.2 Si U<=0.33 entonces hacer X=2, sino
  2.3 Si U<=0.78 entonces hacer X=3, sino
   2.4 Si U<=0.91 entonces hacer X=4, sino hacer X=5.
3.  Devolver X.
```

¿Cuál es el número medio de comparaciones para simular un valor de X? Construir un algoritmo basado en un árbol binario óptimo. ¿Cuál es ahora el número medio de comparaciones? Dar un algoritmo basado en el método de la tabla guía. Calcular, también en este caso, el numero medio de comparaciones.

2. El número de procesos de usuarios activos en una estación de trabajo es una variable aleatoria, X, con distribución dada por:

$$\begin{aligned} P(X &= 0) = 0.33,\ P(X = 1) = 0.22,\ P(X = 2) = 0.15 \\ P(X &= 3) = 0.11,\ P(X = 4) = 0.09,\ P(X = 5) = 0.06, \\ P(X &= 6) = 0.04. \end{aligned}$$

Dar un algoritmo para simular esta variable mediante búsqueda en forma de árbol de Huffman. Calcular el número medio de comparaciones para simular cada valor de la variable X. ¿Cuál es una cota para dicho número medio de comparaciones en caso de usar el método de la tabla guía? Para esta último método, calcular el valor exacto del número medio de comparaciones.

3. El número de veces que se produce una caida de un sistema informático durante un mes es una variable aleatoria, X, con distribución dada por $P(X = 0) = 0.09$, $P(X = 1) = 0.21$, $P(X = 2) = 0.39$, $P(X = 3) = 0.19$, $P(X = 4) = 0.08$ y $P(X = 5) = 0.04$. Dar un algoritmo para simular esta variable mediante búsqueda en forma de árbol de Huffman. Calcular el número medio de comparaciones para simular cada valor de la variable X. ¿Qué ventajas se obtienen para este método de simulación en comparación con la implementación mediante el método secuencial (tanto en su versión directa como en la del etiquetado más eficiente)?

4. El código de error que puede suministrar una rutina implementada en un lenguaje de alto nivel puede describirse mediante una variable aleatoria, C, que puede tomar los valores 0, 1, 3, 4 y 7, con probabilidades:

$$p_0 = 0.28,\ p_1 = 0.35,\ p_3 = 0.16,\ p_4 = 0.12,\ p_7 = 0.09.$$

   Dentro de un complejo modelo de simulación se quiere generar dicha variable usando, para cada valor generado, un único número pseudoaleatorio $U(0,1)$. Encontrar algoritmos, para tal fin, basados en la construcción de un árbol de Huffman y en el método de las tablas guía. Estudiar la eficiencia de ambos algoritmos y compararla.

5. Considérese la variable aleatoria discreta, X, número de fallos mensuales en el suministro eléctrico de un equipo informático, que tiene la siguiente masa de probabilidad:

x	0	1	2	3	4	5	6	7
$P(X=x)$	0.35	0.20	0.16	0.12	0.08	0.05	0.03	0.01

   Construir una tabla guía y, basándose en ella, dar un algoritmo para simular esta variable. ¿Cuál es el número medio exacto de comparaciones para generar un valor mediante ese algoritmo?

6. Dando por supuesto un algoritmo que permita simular la variable aleatoria del ejercicio anterior, deducir un método para aproximar por simulación la probabilidad de que en un año haya más de 30 fallos en el suministro eléctrico. Plantear el cálculo analítico de la cantidad anterior. ¿Sería sencillo obtener teóricamente la masa de probabilidad de la variable aleatoria número de fallos anuales?

5

Métodos específicos para la simulación de distribuciones notables

En este capítulo se estudiarán algoritmos específicos para simular algunas de las distribuciones de probabilidad más importantes. La mayoría de ellos son aplicaciones de los métodos generales ya expuestos, quizá con alguna particularidad. Distribución por distribución, se irán comentando los métodos más habituales de simulación de las mismas, comenzando por las continuas y pasando posteriormente a las variables discretas.

5.1 Distribuciones continuas

5.1.1 La distribución normal

Es obvio que nuestro objetivo será saber simular una normal estándar, X, ya que la distribución general $N(\mu, \sigma)$, con parámetros arbitrarios, puede simularse mediante $\mu + \sigma X$.

Método de Box-Müller

Se basa en la siguiente propiedad. Dadas dos variables aleatorias independientes $E \stackrel{\mathrm{d}}{=} \exp(1)$ y $U \stackrel{\mathrm{d}}{=} U(0,1)$, las variables $\sqrt{2E}\cos 2\pi U$ y $\sqrt{2E}$ sen $2\pi U$ son $N(0,1)$ independientes.

Algoritmo de Box-Müller

1. `Generar` $U, V \sim U(0,1)$.
2. `Hacer` $W_1 = \sqrt{-2\ln U}$ `y` $W_2 = 2\pi V$.
3. `Devolver` $X_1 = W_1 \cos W_2$, $X_2 = W_1$ sen W_2.

Método polar

De nuevo es un método cuyo fundamento teórico descansa en una propiedad que nos da la distribución condicionada a cierto suceso de un par de variables transformadas de otras uniformes. Su nombre procede de que el uso de coordenadas polares es muy útil en la demostración de dicho resultado.

Dadas dos variables independientes V_1 y V_2, con distribución $U(-1,1)$, entonces la distribución condicionada

$$\left(V_1 \sqrt{\frac{-2\ln(V_1^2+V_2^2)}{V_1^2+V_2^2}}, V_2 \sqrt{\frac{-2\ln(V_1^2+V_2^2)}{V_1^2+V_2^2}} \right) \Bigg|_{V_1^2+V_2^2 \leq 1}$$

es una $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$.

Algoritmo polar

1. `Generar` $U_1, U_2 \sim U(0,1)$.
2. `Hacer` $V_1 = 2U_1 - 1$, $V_2 = 2U_2 - 1$ `y` $W = V_1^2 + V_2^2$.
3. `Si` $W > 1$ `entonces volver a 1.`
4. `Hacer` $Y = \sqrt{\frac{-2\ln W}{W}}$.
5. `Devolver` $X_1 = V_1 Y$, $X_2 = V_2 Y$.

Método del Teorema Central del Límite

Su fundamento teórico, como su propio nombre indica, es el Teorema Central del Límite. Dadas variables aleatorias T_1, T_2, ..., T_n, independientes y con distribución cualquiera, se tiene que

$$\frac{\overline{T} - \mu_T}{\frac{\sigma_T}{\sqrt{n}}} = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} T_i - E(T_1)\right)}{\sqrt{Var(T_1)}} \overset{\text{d}}{\simeq} N(0,1),$$

si n es suficientemente grande.

Este teorema puede aplicarse para simular una $N(0,1)$ tomando variables con otra distribución más fácil de simular. El caso más habitual es elegir $T_i = U_i \overset{\text{d}}{=} U(0,1)$ y $n = 12$ (por simplicidad de cálculo). De esta forma, la siguiente variable resulta tener distrribución aproximada $N(0,1)$,

$$\frac{\overline{U} - \mu_T}{\frac{\sigma_U}{\sqrt{n}}} = \frac{\sqrt{12}\left(\frac{1}{12}\sum_{i=1}^{12} U_i - \frac{1}{2}\right)}{\sqrt{\frac{1}{12}}} = \sum_{i=1}^{12} U_i - 6.$$

Algoritmo basado en el TCL

1. `Generar` $U_1, U_2, \ldots, U_{12} \sim U(0,1)$.
2. `Devolver` $X = U_1 + U_2 + \cdots + U_{12} - 6$.

5.1.2 La distribución de Cauchy

Esta distribución puede definirse, de forma general, dependiendo de dos parámetros: μ el de localización y σ el de escala. Su función de densidad viene dada por

$$f(x) = \frac{\sigma}{\pi\left(\sigma^2 + (x-\mu)^2\right)}, \text{ para todo } x \in \mathbb{R}.$$

Un sencillo cálculo permite hallar su función de distribución:

$$F(x) = \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{2},$$

la cual permite implementar el método de inversión. Con razonamientos semejantes a los ya realizados para el algoritmo de la exponencial (en este caso basados en que si $U \sim U(0,1)$ entonces $\tan \pi \left(U - \frac{1}{2}\right) \stackrel{\mathrm{d}}{=} \tan \pi U$), puede encontrarse un algoritmo ligeramente más eficiente desde el punto de vista computacional:

1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = \sigma \tan(\pi U) + \mu$.

5.1.3 La distribución exponencial

Se simula utilizando el método de inversión con la simplificación que ya se indicó anteriormente:

0. `Hacer` $L = -\frac{1}{\lambda}$.
1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = L \cdot \ln U$.

5.1.4 Las distribuciones gamma y de Erlang

La distribución gamma, $\Gamma(a,p)$, depende de dos parámetros: $a > 0$, parámetro de escala, y $p > 0$, parámetro de forma. La distribución de Erlang no es más que la particularización de la gamma al caso en que $p \in \mathbb{N}$. La función de densidad de una $\Gamma(a,p)$ viene dada por

$$f(x) = \begin{cases} k(a,p)\, x^{p-1} e^{-ax} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

donde $k(a,p)$ es la única constante posible para que la función anterior sea una función de densidad. Es obvio que, con tal de que $k(a,p) \geq 0$, se tiene que $f(x) \geq 0$. Para que $\int f(x)\, dx = 1$, ha de verificarse:

$$\begin{aligned} 1 &= \int f(x)\, dx = \int_0^\infty k(a,p)\, x^{p-1} e^{-ax} dx = \frac{k(a,p)}{a} \int_0^\infty \left(\frac{y}{a}\right)^{p-1} e^{-y} dy \\ &= \frac{k(a,p)}{a^p} \int_0^\infty y^{p-1} e^{-y} dy = k(a,p)\, \frac{\Gamma(p)}{a^p}, \end{aligned}$$

donde se define $\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx$, que es la llamada función gamma de Euler. A partir de todo lo anterior se tiene $k(a,p) = \frac{a^p}{\Gamma(p)}$, y, por tanto, la densidad de la gamma resulta ser

$$f(x) = \begin{cases} \frac{a^p}{\Gamma(p)} x^{p-1} e^{-ax} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Puede demostrarse una relación recursiva para $\Gamma(p)$ sin más que hacer una integración por partes, tomando $u = x^{p-1}$ y $dv = e^{-x}dx$. Así, se tiene,

$$\begin{aligned}\Gamma(p) &= \int_0^\infty x^{p-1}e^{-x}dx = \left[x^{p-1}\left(-e^{-x}\right)\right]_0^\infty - \int_0^\infty (p-1)\,x^{p-2}\left(-e^{-x}\right)dx \\ &= (p-1)\int_0^\infty x^{p-2}e^{-x}dx = (p-1)\,\Gamma(p-1).\end{aligned}$$

Esto permite reducir el cálculo de $\Gamma(p)$ al caso en que $p \in (0,1]$, ya que

$$\Gamma(p) = (p-1)(p-2)\cdots(p-\lceil p\rceil)\,\Gamma(p-\lceil p\rceil).$$

Dado que $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$, la fórmula anterior se simplifica cuando $p \in \mathbb{N}$, dando lugar a $\Gamma(p) = (p-1)!$ Por este motivo se dice que la función gamma de Euler generaliza la noción de factorial de un número a cualquier valor real positivo. Además, cuando $p = 1$ la densidad de la gamma es

$$f(x) = \begin{cases} ae^{-ax} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

es decir $\Gamma(a,1) \overset{\mathrm{d}}{=} \exp(a)$.

Una propiedad muy importante de la distribución gamma es la llamada propiedad de reproductividad, que afirma que si se dispone de dos variables aleatorias independientes, $X \overset{\mathrm{d}}{=} \Gamma(a,p_1)$ e $X \overset{\mathrm{d}}{=} \Gamma(a,p_2)$, entonces la suma de ambas también es una gamma: $X+Y \overset{\mathrm{d}}{=} \Gamma(a,p_1+p_2)$. Este resultado se puede generalizar, por inducción a la suma de cualquier número finito de variables gamma independientes con primer parámetro, a, coincidente. En virtud de ello, si p es entero, dadas $X_1, X_2, \cdots, X_p$ variables independientes con distrinución $\exp(a)$ (o, lo que es lo mismo $\Gamma(a,1)$) se tiene que su suma, $\sum_{i=1}^p X_i$, tiene distribución $\Gamma(a,p)$. Como consecuencia, la distribución de Erlang se puede simular fácilmente como suma de exponenciales:

Algoritmo reproductivo de simulación de la Erlang

`1. Desde` $i = 1$ `hasta` p `hacer`
`1.1. Generar` $U_i \sim U(0,1)$.
`1.2. Hacer` $X_i = -\frac{\ln U_i}{a}$.
`2. Devolver` $X = \sum_{i=1}^p X_i$.

Este algoritmo puede agilizarse computacionalmente definiendo previamente el valor $L = -\frac{1}{a}$ y calculando un único logaritmo (en lugar de p) gracias a que $\sum_{i=1}^p X_i = -\sum_{i=1}^p \frac{\ln U_i}{a} = -\frac{1}{a}\ln\left(\prod_{i=1}^p U_i\right)$. Así se tiene:

Algoritmo reproductivo de simulación de la Erlang optimizado

`0. Hacer` $L = -\frac{1}{a}$.
`1. Hacer` $S = 1$.
`2. Desde` $i = 1$ `hasta` p `hacer`
`2.1. Generar` $U \sim U(0,1)$.
`2.2. Hacer` $S = S \cdot U$.
`3. Devolver` $X = L \cdot \ln S$.

Los algoritmos anteriores sólo son válidos para p entero, siendo además muy lentos si p es grande. Por contra son muy simples y de fácil implementación. Como alternativa existen otros algoritmos más complicados que cubren también el caso en que p no sea entero. Veremos el algoritmo de Tadikamalla (1978) para simular una $\Gamma(1,p)$. Se trata de un algortimo de aceptación/rechazo, que sólo es válido si $p > \frac{3}{4}$ y que usa como densidad auxiliar una doble exponencial centrada en $p-1$ y con parámetro de escala dado por

$$\lambda = \frac{1}{\theta} = \frac{2}{1+\sqrt{4p-3}}.$$

Para la implementación del algoritmo debe definirse la función

$$T(x) = \left|\frac{(\theta-1)\,x}{\theta\,(p-1)}\right|^{p-1} \exp\left(-x + \frac{|x-(p-1)| + (p-1)\,(\theta+1)}{\theta}\right).$$

Algoritmo de Tadikamalla

1. Generar X, doble exponencial con media $p-1$ y escala λ.
2. Si $X < 0$ entonces volver a 1.
3. Generar $U \sim U(0,1)$.
4. Si $U \leq T(X)$ entonces devolver X, sinó volver a 1.

Para simular una $\Gamma(a,p)$ con un a cualquiera, usaremos el hecho de que si $X \stackrel{\mathrm{d}}{=} \Gamma(a,p)$ entonces $Y = aX \stackrel{\mathrm{d}}{=} \Gamma(1,p)$. Asi pues bastará simular Y según una $\Gamma(1,p)$ y luego hacer $X = \frac{Y}{a}$.

5.1.5 La distribución beta

Dadas dos variables aleatorias $Y \stackrel{\mathrm{d}}{=} \Gamma(1,p)$ y $Z \stackrel{\mathrm{d}}{=} \Gamma(1,q)$, independientes, se dice que la variable

$$X = \frac{Y}{Y+Z}$$

tiene distribución $\beta(p,q)$, beta de parámetros p y q. Puede demostrarse que en la definición podrían haberse tomado variables con distribuciones $Y \stackrel{\mathrm{d}}{=} \Gamma(a,p)$ y $Z \stackrel{\mathrm{d}}{=} \Gamma(a,q)$, ya que la distribución de la X resultante no depende del a elegido. La función de densidad de una $\beta(p,q)$ viene dada por

$$f(x) = \begin{cases} \dfrac{x^{p-1}\,(1-x)^{q-1}}{\beta(p,q)} & \text{si } x \in [0,1] \\ 0 & \text{en otro caso} \end{cases}$$

siendo $\beta(p,q) = \int_0^1 x^{p-1}\,(1-x)^{q-1}\,dx$.

Existen multitud de algoritmos para simular la distribución $\beta(p,q)$. Probablemente, el más sencillo de todos es el que se obtiene, a partir de la distribución

gamma, como consecuencia de la propia definición. El algoritmo de Fox (1963) es adecuado para simular la distribución beta cuando $p, q \in \mathbb{N}$ y son valores pequeños.

Algortimo de Fox

1. `Generar` $U_1, U_2, \ldots, U_{p+q-1} \sim U(0,1)$.
2. `Ordenarlos:` $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(p+q-1)}$.
3. `Devolver` $X = U_{(p)}$.

Es obvio que este algoritmo puede resultar muy lento si alguno de los dos parámetros es elevado (pues habrá que simular muchas uniformes para conseguir un valor simulado de la beta). Además, en función de cuál de los dos parámetros, p ó q, sea mayor, resultará más eficiente, en el paso 2, comenzar a ordenar por el mayor, luego el segundo mayor, y así sucesivamente, o hacerlo empezando por el menor. En cualquier caso, es obvio que no es necesario odenar todos los valores U_i generados, sino tan sólo encontrar el que ocupa el lugar p-ésimo.

Un método válido aunque p ó q no sean enteros es el dado por el algoritmo de Jöhnk (1964).

Algoritmo de Jöhnk

1. `Repetir.`
1.1. `Generar` $U, V \sim U(0,1)$.
1.2. `Hacer` $Y = U^{\frac{1}{p}}$, $Z = V^{\frac{1}{q}}$, $S = Y + Z$.
2. `Hasta que` $S \leq 1$.
3. `Hacer` $X = \frac{Y}{S}$.

El método resulta extremadamente ineficiente para p ó q mayores que 1. Esto es debido a que la condición $S \leq 1$ del paso 2 puede tardar muchísimo en verificarse. Por este motivo, el algoritmo de Jöhnk sólo es recomendable para $p < 1$ y $q < 1$. Como remedio a esto puede usarse el algoritmo de Cheng (1978) que es bastante más largo de implementar pero también mucho más eficiente.

Algoritmo de Cheng

0.1. `Hacer` $\alpha = p + q$.
0.2. `Si` $\min(p,q) \leq 1$ `entonces hacer` $\beta = \frac{1}{\min(p,q)}$, `en otro caso hacer` $\beta = \sqrt{\frac{\alpha - 2}{2pq - \alpha}}$.
0.3. `Hacer` $\gamma = p + \frac{1}{\beta}$.
1. `Generar` $U_1, U_2 \sim U(0,1)$.
2. `Hacer` $V = \beta \cdot \ln\left(\frac{U_1}{1-U_1}\right)$ `y` $W = p \cdot e^V$.
3. `Si` $\alpha \cdot \ln\left(\frac{\alpha}{q+W}\right) + \gamma V - \ln 4 < \ln(U_1^2 U_2)$ `entonces volver a 1.`
4. `Devolver` $X = \frac{W}{q+W}$.

5.1.6 La distribución chi-cuadrado de Pearson

Dadas variables aleatorias $Z_1, Z_2, \ldots, Z_n$ independientes y con distribución común $N(0,1)$, la variable $X = \sum_{i=1}^{n} Z_i^2$ se dice de distribución chi-cuadrado con n grados de libertad, i.e. χ_n^2. Su función de densidad viene dada por

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, \text{ para todo } x \geq 0.$$

Como consecuencia, $\chi_n^2 \stackrel{\mathrm{d}}{=} \Gamma\left(\frac{1}{2}, \frac{n}{2}\right)$ y los algoritmos vistos para la distribución gamma son aplicables a este caso. Además, debido a la reproductividad ya mencionada se tiene que $\Gamma\left(\frac{1}{2}, \frac{n}{2}\right) \stackrel{\mathrm{d}}{=} \Gamma\left(\frac{1}{2}, \left\lceil \frac{n}{2} \right\rceil\right) + \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$, cuando n no sea par, siendo esta última distribución, $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$, la del cuadrado de una normal estándar. De esta forma, se tiene el siguiente algoritmo para la simulación de la chi-cuadrado:

Algoritmo reproductivo para simular la chi-cuadrado

`0.` Hacer $m = \left\lceil \frac{n}{2} \right\rceil$.
`1.` Hacer $S = 1$.
`2.` Desde $i = 1$ hasta m hacer
`2.1.` Generar $U \sim U(0,1)$.
`2.2.` Hacer $S = S \cdot U$.
`3.` Hacer $X = -2 \ln S$.
`4.` Si n es impar hacer
`4.1.` Generar $Z \sim N(0,1)$.
`4.2.` Hacer $X = X + Z^2$.
`5.` Devolver X.

De todas formas, si n es grande es recomendable usar el algoritmo de Tadikamalla.

5.1.7 La distribución F de Fisher-Snedecor

Dadas dos variables aleatorias $Y_1 \stackrel{\mathrm{d}}{=} \chi_m^2$ e $Y_2 \stackrel{\mathrm{d}}{=} \chi_n^2$ independientes, la variable aleatoria definida por

$$X = \frac{\frac{Y_1}{m}}{\frac{Y_2}{n}}$$

se dice de distribución F de Snedecor (o de Fisher) con m y n grados de libertad ($F_{m,n}$). Su densidad es

$$f(x) = \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{\beta\left(\frac{m}{2}, \frac{n}{2}\right)} x^{\frac{m}{2}-1} \left(x + \frac{n}{m}\right)^{-\frac{m+n}{2}}, \text{ para todo } x \geq 0,$$

que recuerda a la de una beta. Además de poder simularse a través de algoritmos de generación de la chi-cuadrado (como consecuencia de su definición), también puede simularse mediante el uso de una distribución beta.

Algoritmo de simulación de la F a través de la beta

1. `Generar` $Z \sim \beta\left(\frac{m}{2}, \frac{n}{2}\right)$.
2. `Hacer` $X = \frac{nZ}{m(1-Z)}$.

5.1.8 La distribución t de Student

Dadas dos variables independientes $Y_1 \overset{\mathrm{d}}{=} N(0,1)$ e $Y_2 \overset{\mathrm{d}}{=} \chi^2_n$, la variable aleatoria

$$X = \frac{Y_1}{\sqrt{\frac{Y_2}{n}}}$$

se dice que tiene distribución t de Student con n grados de libertad (t_n). La función de densidad de una t_n es

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \text{ para todo } x \in \mathbb{R}.$$

Teniendo en cuenta la simetría de esta distribución y su relación con la F de Snedecor: $(t_n)^2 \overset{\mathrm{d}}{=} F_{1,n}$, puede simularse fácilmente la t de Student.

Algoritmo de simulación de la t de Student a partir de la F

1. `Generar` $U \sim U(0,1)$ `y` $Z \sim F_{1,n}$.
2. `Si` $U < 0.5$ `entonces devolver` $X = \sqrt{Z}$`, sino devolver` $X = -\sqrt{Z}$.

5.1.9 La distribución de Weibull

La distribución de Weibull, $W(\lambda, \alpha)$, es una generalización de la distribución $\exp(\alpha)$. Su función de densidad de probabilidad es

$$f(x) = \alpha\lambda^\alpha x^{\alpha-1}e^{-(\lambda x)^\alpha}, \text{ para todo } x \geq 0.$$

Como se ve, $W(\lambda, 1) \overset{\mathrm{d}}{=} \exp(\lambda)$. Esta distribución se utiliza mucho en fiabilidad y análisis de supervivencia para modelizar el tiempo de vida de una componente o de un ser vivo. Puede simularse fácilmente mediante el método de inversión ligeramente optimizado.

Algoritmo de inversión para simular la distribución de Weibull

1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = \frac{(-\ln U)^{\frac{1}{\alpha}}}{\lambda}$.

5.1.10 La distribución logística

Es la que tiene por función de distribución:

$$F(x) = \frac{1}{1 + e^{-\frac{x-a}{b}}}, \forall x \in \mathbb{R},$$

siendo $a \in \mathbb{R}$ y $b > 0$. Puede simularse fácilmente mediante el método de inversión.

Algoritmo para simular la distribución logística mediante inversión

1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = a - b \ln\left(\frac{1}{U} - 1\right)$.

5.1.11 La distribución de Pareto

Tiene utilidad en ciencias como la Economía, donde en ocasiones sirve para modelizar distribuciones de rentas. Su densidad viene dada por

$$f(x) = \begin{cases} \frac{ab^a}{x^{a+1}} & \text{si } x \geq b \\ 0 & \text{si } x < b \end{cases}$$

Como consecuencia, su función de distribución resulta

$$F(x) = \begin{cases} 0 & \text{si } x < b \\ 1 - \left(\frac{b}{x}\right)^a & \text{si } x \geq b \end{cases}$$

y, por tanto, es simulable mediante inversión. Una versión óptimizada del algoritmo es:

Algoritmo de inversión para simular la distribución de Pareto

1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = \frac{b}{U^{\frac{1}{a}}}$.

5.2 Distribuciones discretas

5.2.1 La distribución uniforme discreta

Dado un conjunto finito de N elementos (sin pérdida de generalidad supondremos el conjunto $\{1, 2, \ldots, N\}$) la distribución uniforme discreta en dicho conjunto (o equiprobable sobre dicho conjunto) es la definida por $P(X = i) = \frac{1}{N}$, para $i = 1, 2, \ldots, N$. Tanto el método de inversión (calculando explícitamente la función cuantil) como el de truncamiento dan lugar al siguiente algoritmo.

Algoritmo para simular la distribución uniforme discreta en $\{1, 2, \ldots, N\}$

1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = \lceil N \cdot U \rceil + 1$.

5.2.2 La distribución binomial

La distribución binomial de extensión n y probabilidad de éxito p, $B(n,p)$, se define como el número de éxitos en n pruebas idénticas, independientes, en las que la probabilidad de éxito es p. Su masa de probabilidad es

$$P(X=i) = \binom{n}{i} p^i (1-p)^{n-i}, \text{ para } i = 0, 1, \ldots, n.$$

Puede simularse a partir de su definición:

1. `Hacer` $S = 0$.
2. `Repetir` n `veces`
2.1. `Generar` $U \sim U(0,1)$.
2.2. `Si` $U \leq p$ `entonces hacer` $S = S + 1$.
3. `Devolver` $X = S$.

Este método es extremadamente lento cuando n es grande, por eso, en ese caso, resulta más ventajoso utilizar el método de la tabla guía.

5.2.3 La distribución de Poisson

Una variable aleatoria discreta, X, tiene distribución de Poisson de parámetro $\lambda > 0$ si su masa de probabilidad viene dada por

$$P(X=i) = \frac{e^{-\lambda}\lambda^i}{i!}, \text{ para } i = 0, 1, \ldots$$

La distribución de Poisson puede simularse mediante el método de la transformación cuantil con búsqueda secuencial. También puede simularse haciendo uso de la relación que guarda con la distribución exponencial. Así, dadas variables aleatorias $T_1, T_2, \ldots, T_n, \ldots$ independientes y con distribución $\exp(\lambda)$, la variable aleatoria entera, X, que verifica

$$\sum_{i=1}^{X} T_i \leq 1 < \sum_{i=1}^{X+1} T_i,$$

(definiendo $X = 0$ si $T_1 > 1$) tiene distribución Pois(λ).

Las variables aleatorias T_i pueden simularse, utilizando valores U_i de una uniforme, mediante $T_i = -\frac{\ln U_i}{\lambda}$. En virtud de ello, se tiene

$$\begin{aligned}
\sum_{i=1}^{X} T_i &\leq 1 < \sum_{i=1}^{X+1} T_i \Leftrightarrow -\sum_{i=1}^{X} \frac{\ln U_i}{\lambda} \leq 1 < -\sum_{i=1}^{X+1} \frac{\ln U_i}{\lambda} \Leftrightarrow \\
-\frac{\ln\left(\prod_{i=1}^{X} U_i\right)}{\lambda} &\leq 1 < -\frac{\ln\left(\prod_{i=1}^{X+1} U_i\right)}{\lambda} \Leftrightarrow \ln\left(\prod_{i=1}^{X} U_i\right) \geq -\lambda > \ln\left(\prod_{i=1}^{X+1} U_i\right) \Leftrightarrow \\
\prod_{i=1}^{X} U_i &\geq e^{-\lambda} > \prod_{i=1}^{X+1} U_i.
\end{aligned}$$

Así, puede utilizarse el siguiente algoritmo:

Algoritmo de simulación de la Poisson a traves de la exponencial

1. `Hacer` $p = 1$ `y` $S = -1$.
2. `Repetir`
2.1. `Generar` $U \sim U(0,1)$.
2.2. `Hacer` $p = p \cdot U$ `y` $S = S + 1$.
3. `Hasta que` $p < e^{-\lambda}$.
4. `Hacer` $X = S$.

Tanto este algoritmo como el de la transformación cuantil tienen el inconveniente de ser muy ineficientes cuando λ es grande. En ese caso, aunque la distribución de Poisson tiene un número infinito de posibles resultados, es perfectamente aplicable el método de la tabla guía, agrupando todos los infinitos valores a partir de uno en adelante. Esto desemboca en una búsqueda secuencial cuando el intervalo elegido sea el último de la tabla. De esta forma se mejora muy considerablemente la eficiencia del método.

5.2.4 La distribución geométrica

Su masa de probabilidad es

$$P(X = i) = p \cdot (1-p)^i, \text{ para } i = 0, 1, \ldots$$

Aparte de poder simularse a partir de su definición (número de fracasos antes del primer éxito) también puede hacerse por truncamiento. Como ya se vió en su momento, el algoritmo que resulta por este método es equivalente al basado en la expresión explícita de la función cuantil.

Algoritmo de truncamiento para la distribución geométrica

0. `Hacer` $L = -\frac{1}{\ln(1-p)}$.
1. `Generar` $U \sim U(0,1)$.
2. `Hacer` $T = L \cdot \ln U$.
3. `Devolver` $X = \lceil T \rceil$.

5.2.5 La distribución binomial negativa

Como es sabido, la distribución binomial negativa, $BN(r,p)$, generaliza la geométrica. Puede interpretarse como el número de fracasos antes del r-ésimo éxito. Su función de masa de probabilidad es

$$P(X=i)=\binom{i+r-1}{i}p^r(1-p)^i\text{, para } i=0,1,\ldots$$

Debido a su reproductividad en el parámetro r, puede simularse como suma de r variables geométricas, aunque este algoritmo puede ser muy costoso en tiempo de computación si r es elevado. Existe también un método específico basado en la propiedad

$$X|_Y \overset{\mathrm{d}}{=} \text{Pois}(Y)\,,\ Y \overset{\mathrm{d}}{=} \Gamma\left(\frac{p}{1-p},r\right) \Rightarrow X \overset{\mathrm{d}}{=} BN(r,p)\,.$$

Algoritmo condicional para simular la binomial negativa

1. `Simular` $L\sim\Gamma\left(\frac{p}{1-p},r\right)$.
2. `Simular` $X\sim$`Pois`(L).
3. `Devolver` X.

5.3 Ejercicios propuestos

1. ¿Podría modificarse el algoritmo dado como solución al ejercicio 1, propuesto en el tema 3, para mejorar su eficiencia, de forma que se utilicen ahora tres variables exponenciales de parámetro 2? ¿Cómo?

2. Un programa antivirus comprueba, tan solo, algunas posiciones de memoria (elegidas aleatoriamente) de un ordenador. Debido a esto, en ordenadores con virus, la probabilidad de que el programa lo detecte en una única ejecución es del 80%. Suponiendo que las distintas ejecuciones del programa actúan de forma independiente, dado un ordenador con virus, ¿cómo podría simularse la variable *número de ejecuciones hasta que se detecta el virus*? ¿Podría hacerse lo anterior mediante el uso de un solo número pseudoaleatorio uniforme por cada valor a simular de la variable de interés?

3. La variable X, tiempo que transcurre entre las respuestas a dos comandos consecutivos ejecutados por un usuario de un laboratorio de informática, puede modelizarse como suma del tiempo que el usuario tarda en "pensar" el siguiente comando a ejecutar (*thinking time*) y del tiempo de respuesta del sistema ante la ejecución del comando en cuestión (*tiempo*

de respuesta). Suponiendo que el *thinking time* se modeliza según una distribución $\Gamma(2,3)$ y que para el *tiempo de respuesta* se usa una normal de media 0.5 y desviación típica 0.1, se pide:

(a) Detallar un algoritmo para simular la variable aleatoria X.

(b) Comentar la eficiencia del algoritmo encontrado.

4. En el laboratorio de informática del ejercicio anterior hay tres impresoras. Para cada impresora, la probabilidad de encontrarla atascada es de un 10%.

   (a) Hallar un algoritmo para simular la variable "número de impresoras operativas del laboratorio".

   (b) Encontrar un algoritmo, para el mismo fin, utilizando un sólo número aleatorio uniforme en $(0,1)$, por cada valor de la variable a simular.

5. La fracción de tiempo (o tanto por uno) que cada impresora del ejercicio anterior está ocupada es una variable continua con función de densidad:

$$f(x) = \begin{cases} 110x(1-x)^9 & \text{si} \quad x \in [0,1] \\ 0 & \text{en otro caso} \end{cases}$$

   (a) Obtener un algoritmo para simular la fracción de tiempo de ocupación de una impresora.

   (b) Comentar la eficiencia del algoritmo del apartado anterior.

   (c) ¿Es posible encontrar un algoritmo para el que sepamos de antemano el número de números aleatorios uniformes necesarios para simular un valor de la fracción de tiempo de ocupación? En caso afirmativo, detallar cómo.

6

Simulación de distribuciones multidimensionales

La simulación de vectores aleatorios $\overrightarrow{X} = (X_1, X_2, \ldots, X_d)$ que sigan cierta distribución dada no es tarea siempre sencilla. En general, no resulta una extensión inmediata del caso unidimensional, aúnque, si las variables que componen el vector son independientes, entonces bastará simular cada X_i con la distribución marginal deseada (F_i) y luego agrupar los valores simulados para cada componente en un vector.

En la mayor parte de los casos de interés, las componentes del vector aleatorio son dependientes y el método anterior no es válido. A continuación se verán algunos métodos generales para la simulación de distribuciones multidimensionales.

6.1 Método de las distribuciones condicionadas

Supóngase un vector aleatorio d-dimensional, con distribución continua. Denótese por $f(x_1, x_2, \ldots, x_n)$ su función de densidad conjunta y considérese la primera densidad marginal, $f_1(x_1)$, y las sucesivas densidades condicionales $f_2(x_2|x_1)$, $f_3(x_3|x_1, x_2)$, $\ldots$, $f_d(x_d|x_1, x_2, \ldots, x_{d-1})$. Gracias a la regla del producto, generalizada a funciones de densidad, se tiene

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) \cdot f_2(x_2|x_1) \cdot f_3(x_3|x_1, x_2) \cdots f_d(x_d|x_1, x_2, \ldots, x_{d-1})$$

y, como consecuencia, puede darse el siguiente algoritmo general:

1. `Generar` X_1 `con densidad` f_1.
2. `Desde` $i = 2$ `hasta` d `generar` X_i `con densidad` $f_i(\bullet|X_1, X_2, \ldots, X_{i-1})$.
3. `Devolver` $\overrightarrow{X} = (X_1, X_2, \ldots, X_d)$.

Es inmediato comprobar que el método anteriormente expuesto es igualmente válido si las variables X_i son discretas o, incluso, si algunas son discretas y otras continuas. En tal caso se sustituiría la densidad por la masa de probabilidad. Así pues, lo realmente importante para poder aplicar el método de las distribuciones condicionadas es conocer y saber simular la distribución marginal de X_1 y las distribuciones condicionadas del tipo $X_i|X_1, X_2, \ldots, X_{i-1}$ para $i = 2, 3, \ldots, d$.

Ejemplo 23 *Trataremos de dar un algoritmo para simular la distribución normal bidimensional por el método de las distribuciones condicionadas. Consi-*

deremos una

$$N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right),$$

por las propiedades de la distribución normal, bastará saber simular la distribución

$$N_2\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right)$$

y luego sumarle el vector $\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$.

Dado que $X_1 \overset{d}{=} N(0,\sigma_1)$, *se tiene que*

$$f_1(x_1) = \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{x_1^2}{2\sigma_1^2}\right).$$

Además

$$f(x_1,x_2) = f(\overrightarrow{x}) = \frac{1}{2\pi\sqrt{\det(\Sigma)}}\exp\left(-\frac{1}{2}\overrightarrow{x}^t\Sigma^{-1}\overrightarrow{x}\right).$$

Como

$$\Sigma^{-1} = \frac{1}{\det(\Sigma)}\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix} = \frac{1}{\sigma_1^2\sigma_2^2-\sigma_{12}^2}\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix},$$

se tiene que

$$\frac{1}{2}\overrightarrow{x}^t\Sigma^{-1}\overrightarrow{x} = \frac{\sigma_2^2x_1^2 - 2\sigma_{12}x_1x_2 + \sigma_1^2x_2^2}{2(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}$$

y, por tanto,

$$f_2(x_2|x_1) = \frac{f(x_1,x_2)}{f_1(x_1)}$$

$$= \frac{\sigma_1\sqrt{2\pi}}{2\pi\sqrt{\sigma_1^2\sigma_2^2-\sigma_{12}^2}}\exp\left(-\left(\frac{\sigma_2^2x_1^2 - 2\sigma_{12}x_1x_2+\sigma_1^2x_2^2}{2(\sigma_1^2\sigma_2^2-\sigma_{12}^2)} - \frac{x_1^2}{2\sigma_1^2}\right)\right)$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}}\exp\left(-\frac{\sigma_1^2\sigma_2^2x_1^2 - 2\sigma_1^2\sigma_{12}x_1x_2 + \sigma_1^4x_2^2 - (\sigma_1^2\sigma_2^2-\sigma_{12}^2)x_1^2}{2\sigma_1^2(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}\right)$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}}\exp\left(-\frac{-2\sigma_1^2\sigma_{12}x_1x_2 + \sigma_1^4x_2^2 + \sigma_{12}^2x_1^2}{2\sigma_1^2(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}\right)$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}}\exp\left(-\frac{-\frac{2\sigma_{12}x_1x_2}{\sigma_1^2} + x_2^2 + \frac{\sigma_{12}^2x_1^2}{\sigma_1^4}}{2\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}\right)$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}}\exp\left(-\frac{\left(x_2-\frac{\sigma_{12}x_1}{\sigma_1^2}\right)^2}{2\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}\right),$$

que es la densidad de una $N\left(\frac{\sigma_{12}}{\sigma_1^2}x_1, \sqrt{\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}\right)$.

En resumen, se tiene que si

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{d}{=} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right),$$

entonces $X_1 \overset{d}{=} N(0, \sigma_1)$ *y* $X_2|X_1 \overset{d}{=} N\left(\frac{\sigma_{12}}{\sigma_1^2}X_1, \sqrt{\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}\right)$. *Así, el algoritmo de simulación consistiría en los siguientes pasos:*

1. *Simular* $Z_1, Z_2 \sim N(0,1)$ *independientes.*
2. *Hacer* $Y_1 = \sigma_1 Z_1$.
3. *Hacer* $Y_2 = \frac{\sigma_{12}}{\sigma_1^2}Y_1 + Z_2\sqrt{\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2}}$.
4. *Hacer* $X_1 = Y_1 + \mu_1$, $X_2 = Y_2 + \mu_2$.
5. *Devolver* $\overrightarrow{X} = (X_1, X_2)^t$.

Ejemplo 24 (La distribución uniforme en el círculo unitario) *Se trata de la distribución bidimensional continua cuya densidad es constante en dicho círculo*

$$C = \{(x_1, x_2) \in \mathbb{R}^2 / x_1^2 + x_2^2 \leq 1\}.$$

Su función de densidad viene dada, por tanto, por

$$f(x_1, x_2) = \begin{cases} \frac{1}{\pi} & si \quad (x_1, x_2) \in C \\ 0 & si \quad (x_1, x_2) \notin C \end{cases}$$

La densidad marginal de la primera variable resulta

$$f_1(x_1) = \int_{-\sqrt{1-x_1^2}}^{+\sqrt{1-x_1^2}} \frac{1}{\pi} dx_2 = \frac{2\sqrt{1-x_1^2}}{\pi} \; si \; x_1 \in [-1,1],$$

es decir,

$$f_1(x_1) = \begin{cases} \frac{2}{\pi}\sqrt{1-x_1^2} & si \quad x_1 \in [-1,1] \\ 0 & si \quad x_1 \notin [-1,1] \end{cases}$$

Además

$$f_2(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{\frac{1}{\pi}}{\frac{2\sqrt{1-x_1^2}}{\pi}} = \frac{1}{2\sqrt{1-x_1^2}}, \; si \; x_2 \in \left[-\sqrt{1-x_1^2}, \sqrt{1-x_1^2}\right]$$

valiendo cero en otro caso. Se tiene entonces que

$$X_2|X_1 \overset{d}{=} U\left[-\sqrt{1-X_1^2}, \sqrt{1-X_1^2}\right],$$

siempre que $X_1 \in [-1,1]$.

Finalmente, el algoritmo resulta:

1. `Simular` X_1 `con densidad` $f_1(x_1) = \frac{2}{\pi}\sqrt{1-x_1^2}1_{\{|x_1|\leq 1\}}$.
2. `Simular` X_2 `con densidad` $U\left[-\sqrt{1-X_1^2}, \sqrt{1-X_1^2}\right]$.
3. `Devolver` $\overrightarrow{X} = (X_1, X_2)^t$.

Para el paso 1 puede utilizarse, por ejemplo, el método de aceptación/rechazo, pues se trata de una densidad acotada definida en un intervalo acotado.

6.2 El método de aceptación/rechazo

La idea general del método de aceptación/rechazo es aplicable para simular variables aleatorias definidas en cualquier espacio (no sólo en $\mathbb{R}$). En particular puede usarse para simular vectores aleatorios de $\mathbb{R}^d$. Sin embargo, en este contexto, resulta mucho más difícil encontrar una densidad auxiliar adecuada y, especialmente, conseguir que el número medio de comparaciones del método se mantenga dentro de unos límites de eficiencia razonables cuando la dimensión es elevada. Veamos un ejemplo

Ejemplo 25 *Simulación de puntos uniformemente distribuídos sobre la "esfera" unitaria d-dimensional*

$$C_d = \left\{(x_1, x_2, \ldots, x_d)\,/\, x_1^2 + x_2^2 + \cdots + x_d^2 \leq 1\right\}.$$

Denotando por $V_d(1)$, el "volumen" (la medida) de la esfera d-dimensional de radio 1 (en general, la de radio r verifica $V_d(r) = r^d V_d(1)$), se tiene:

$$f(x_1, x_2, \ldots, x_d) = \begin{cases} \frac{1}{V_d(1)} & si \quad (x_1, x_2, \ldots, x_d) \in C_d \\ 0 & si \quad (x_1, x_2, \ldots, x_d) \notin C_d \end{cases}$$

Para simular valores en $\mathbb{R}^d$, con densidad f, podemos utilizar como densidad auxiliar la de una $U\left([-1,1]\times[-1,1]\times\overset{d}{\cdots}\times[-1,1]\right) = U\left([-1,1]^d\right)$, dada por

$$g(x_1, x_2, \ldots, x_d) = \begin{cases} \frac{1}{2^d} & si \quad x_i \in [-1,1], \text{ para todo } i = 1, 2, \ldots, d \\ 0 & \text{en otro caso} \end{cases}$$

La constante c óptima para la utilización del método de aceptación/rechazo es

$$c = \max_{\overrightarrow{x}/g(\overrightarrow{x})>0} \frac{f(\overrightarrow{x})}{g(\overrightarrow{x})} = \frac{\frac{1}{V_d(1)}}{\frac{1}{2^d}} = \frac{2^d}{V_d(1)}$$

y la condición de aceptación $cUg\left(\overrightarrow{T}\right) \leq f\left(\overrightarrow{T}\right)$ se convierte en

$$\frac{2^d}{V_d(1)} U \frac{1}{2^d} 1_{[-1,1]^d}\left(\overrightarrow{T}\right) \leq \frac{1}{V_d(1)} 1_{C_d}\left(\overrightarrow{T}\right),$$

o, lo que es lo mismo, $U1_{[-1,1]^d}\left(\overrightarrow{T}\right) \leq 1_{C_d}\left(\overrightarrow{T}\right)$*. Dado que el número aleatorio* U *está en el intervalo* $[0,1)$ *y que las funciones indicadoras valen* 0 *ó* 1*, esta condición equivale a que* $1_{[-1,1]^d}\left(\overrightarrow{T}\right) = 1_{C_d}\left(\overrightarrow{T}\right)$*, es decir, a que* $\overrightarrow{T} \in C_d$*, es decir, que se verifique*

$$T_1^2 + T_2^2 + \cdots + T_d^2 \leq 1.$$

Por otra parte, la simulación de $T \sim U\left([-1,1]^d\right)$ *puede hacerse trivialmente mediante* $T_i \sim U([-1,1])$ *para cada* $i = 1, 2, \ldots, d$*, ya que las componentes son independientes. Como el valor de* U *es superfluo en este caso, el algoritmo queda:*

1. **Simular** $V_1, V_2, \ldots, V_d \sim U(0,1)$ **independientes.**
2. **Para** $i = 1, 2, \ldots, d$ **hacer** $T_i = 2V_i - 1$**.**
3. **Si** $T_1^2 + T_2^2 + \cdots + T_d^2 > 1$ **entonces volver al paso 1.**
4. **Devolver** $\overrightarrow{X} = (T_1, T_2, \ldots, T_d)^t$**.**

Usando las fórmulas del "volumen" de una "esfera" d-dimensional:

$$V_d(r) = \begin{cases} \dfrac{\pi^{d/2} r^d}{(d/2)!} & si \quad d \text{ es par} \\ \dfrac{2^{\lceil \frac{d}{2} \rceil + 1} \pi^{\lceil \frac{d}{2} \rceil} r^d}{1 \cdot 3 \cdot 5 \cdots d} & si \quad d \text{ es impar} \end{cases}$$

puede verse que el número medio de repeticiones de los pasos 1-3 del algoritmo, que viene dado por la constante $c = \frac{2^d}{V_d(1)}$*, puede hacerse enórmemente grande. Así, si* $d = 2$ *se tiene* $c = 1.27$*, si* $d = 3$ *se tiene* $c = 1.91$*, si* $d = 4$ *entonces* $c = 3.24$ *y para* $d = 10$ *resulta* $c = 401.5$ *que es un valor que hace que el algoritmo sea tremendamente lento en dimensión* 10*.*

6.3 Métodos de codificación o etiquetado

En el caso de que la función de distribución d-dimensional sea discreta existen métodos que permiten reducir la simulación de dicha variable al contexto de simular una variable aleatoria discreta unidimensional. Estos métodos son conocidos como métodos de etiquetado o codificación y la idea básica consiste en construir una función h que codifique las posibles d-tuplas del conjunto donde toma valores la variable discreta, haciendo corresponder a cada uno un número entero no negativo diferente.

Supongamos que tenemos una variable bidimensional discreta (X_1, X_2) cada una de cuyas componentes toma valores enteros no negativos. El subconjunto de $\mathbb{R}^2$ en el que toma valores el vector aleatorio es

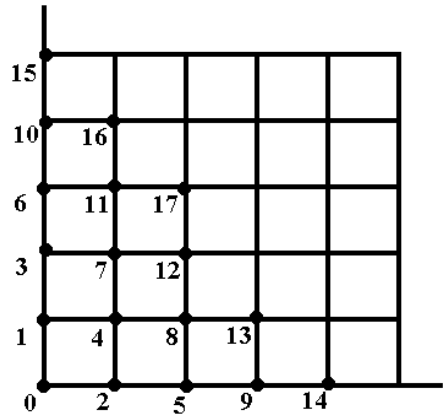
$$\mathbb{Z}^+ \times \mathbb{Z}^+ = \left(\mathbb{Z}^+\right)^2 = \{(i,j)\,/ i,j \in \{0,1,2,\ldots\}\}.$$

Se tratará de definir una función biyectiva, $h : \mathbb{Z}^+ \times \mathbb{Z}^+ \longrightarrow \mathbb{Z}^+$, que permita etiquetar los pares de enteros.

Una posibilidad sencilla consiste en utilizar

$$h(i,j) = \frac{(i+j)(i+j+1)}{2} + i,$$

que corresponde con un etiquetado de los vectores de $(\mathbb{Z}^+)^2$ de la forma dada por la siguiente figura.



Etiquetado dado por h

De esta forma, h induce sobre la variable transformada, $C = h(X_1, X_2)$, una masa de probabilidad

$$p_k^{(C)} := P(C = k) = P(h(X_1, X_2) = k) = P\left((X_1, X_2) = h^{-1}(k)\right) =: p_{h^{-1}(k)}^{(X_1,X_2)}.$$

Resulta inmediato, por tanto, obtener la masa de probabilidad de la variable discreta unidimensional C, a partir de la masa de probabilidad de la variable original (X_1, X_2). De todas formas, debemos tener en cuenta que para que esto sea calculable en la práctica en un tiempo razonable, la función h debe poder invertirse de forma rápida.

Así pues para simular la variable (X_1, X_2) podemos proceder mediante uno de los algoritmos posibles para simular C calculando en tantos pasos como sean necesarios los valores de la forma $h^{-1}(k)$. Con la función h comentada anteriormente podemos calcular de forma rápida su inversa. Veámoslo.

Consideremos $k \in \mathbb{Z}^+$, el valor $(i,j) = h^{-1}(k)$ debe verificar

$$h(i,j) = k \Leftrightarrow \frac{(i+j)(i+j+1)}{2} + i = k.$$

Denotando ahora $n = i + j$, para encontrar $(i, j) = h^{-1}(k)$ basta con hallar n e i, enteros positivos, con $n \geq i$, tales que

$$\frac{n(n+1)}{2} + i = k.$$

Debemos entonces encontrar el único n que cumple

$$\frac{n(n+1)}{2} \leq k \leq \frac{n(n+1)}{2} + n < \frac{n(n+1) + 2(n+1)}{2} = \frac{(n+1)(n+2)}{2}.$$

Como además $n^2 < n(n+1)$ y $(n+1)(n+2) < (n+2)^2$, se tiene que ese valor n ha de verificar

$$n^2 < 2k < (n+2)^2,$$

es decir

$$\left\lceil \sqrt{2k} \right\rceil - 2 < n \leq \left\lceil \sqrt{2k} \right\rceil.$$

Dicho de otro modo, se tiene que n ha de ser igual a $\left\lceil \sqrt{2k} \right\rceil - 1$ ó $\left\lceil \sqrt{2k} \right\rceil$. Basta entonces calcular la expresión $\frac{n(n+1)}{2}$ para esos posibles valores de n. Así, si $\left\lceil \sqrt{2k} \right\rceil \left(\left\lceil \sqrt{2k} \right\rceil + 1 \right) > 2k$ entonces $n = \left\lceil \sqrt{2k} \right\rceil - 1$ y, en caso contrario, $n = \left\lceil \sqrt{2k} \right\rceil$. Finalmente se calcula

$$i = k - \frac{n(n+1)}{2} \text{ y } j = n - i.$$

Ejemplo 26 *Calculemos, por el procedimiento descrito anteriormente, el valor* $h^{-1}(16)$. *Calculamos primeramente* $n = \lceil \sqrt{2 \cdot 16} \rceil = \lceil \sqrt{2 \cdot 16} \rceil = \lceil \sqrt{32} \rceil = \lceil 5.6568542 \rceil = 5$. *Luego calculamos* $5(5+1) = 30 \leq 32 = 2 \cdot 16$, *con lo cual* $n = 5$. *Además* $i = 16 - \frac{5 \cdot 6}{2} = 1$ *y* $j = 5 - 1 = 4$. *Así pues se obtiene* $h^{-1}(16) = (1, 4)$, *que coincide con lo que se puede observar en la figura anterior.*

Se observa pues como el cálculo de $h^{-1}(k)$ es muy rápido. El resto del algoritmo se reduce a la simulación de la variable unidimensional C.

Aunque no entraremos con detalle en ello, conviene resaltar que es posible generalizar este tipo de funciones de codificación a $(\mathbb{Z}^+)^d$. También es factible encontrar la inversa de tal función generalizada (llamada función de decodificación) que se puede calcular eficientemente.

Cuando la variable aleatoria X_2 toma un número finito de valores (supongamos comprendidos entre 0 y M), otra posible función de codificación, más sencilla es

$$h(i, j) = (M+1)\, i + j,$$

cuya inversa viene dada por

$$h^{-1}(k) = \left(\left\lceil \frac{k}{M+1} \right\rceil, k \bmod (M+1) \right).$$

Estas funciones de codificación y decodificación son generalizables a $(\mathbb{Z}^+)^d$ y aplicables al caso en que el vector aleatorio $\overrightarrow{X}$ tome un número finito de valores.

6.4 Métodos específicos para simular la distribución normal multivariante

Dado un vector $\overrightarrow{\mu} = (\mu_1, \mu_1, \ldots, \mu_d)^t \in \mathbb{R}^d$ y una matriz definida positiva

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix}$$

la distribución normal d-dimensional de esos parámetros, abreviadamente $N_d(\overrightarrow{\mu}, \Sigma)$, es la que tiene densidad dada por

$$f(\overrightarrow{x}) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2} (\overrightarrow{x} - \overrightarrow{\mu})^t \Sigma^{-1} (\overrightarrow{x} - \overrightarrow{\mu}) \right).$$

Cuando la matriz Σ es diagonal:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{pmatrix},$$

se obtiene fácilmente

$$\begin{aligned} f(\overrightarrow{x}) &= (2\pi)^{-d/2} \left(\prod_{i=1}^{d} \sigma_i^2 \right)^{-1/2} \\ &\quad \times \exp\left(-\frac{1}{2} (\overrightarrow{x} - \overrightarrow{\mu})^t \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_d^2} \end{pmatrix} (\overrightarrow{x} - \overrightarrow{\mu}) \right) \\ &= (2\pi)^{-d/2} \left(\prod_{i=1}^{d} \sigma_i^{-1} \right) \exp\left(-\frac{1}{2} \sum_{i=}^{d} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right) \\ &= \prod_{i=1}^{d} \left(\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right) \right) = \prod_{i=1}^{d} \phi_{\mu_i, \sigma_i}(x_i), \end{aligned}$$

siendo ϕ_{μ_i,σ_i} la función de densidad de una $N(\mu_i, \sigma_i)$. De esta forma, cuando la matriz Σ (que resulta ser la matriz de varianzas-covarianzas de la normal d-dimensional) es diagonal, entonces las componentes son independientes y resulta trivial simular la $N_d(\overrightarrow{\mu}, \Sigma)$ mediante el siguiente algoritmo:

1. `Simular` $Z_1, Z_2, \ldots, Z_d \sim N(0,1)$ `independientes.`
2. `Para` $i = 1, 2, \ldots, d$ `hacer` $X_i = \mu_i + \sigma_i Z_i$.
3. `Devolver` $\overrightarrow{X} = (X_1, X_2, \ldots, X_d)^t$.

Una propiedad que resulta muy útil para simular la distribución $N_d(\overrightarrow{\mu}, \Sigma)$ con Σ arbitraria es la siguiente.

Proposición 27 *Si* $\overrightarrow{X} \stackrel{d}{=} N_d(\overrightarrow{\mu}, \Sigma)$ *y* A *es una matriz de dimensión* $p \times d$*, de rango máximo, con* $p \leq d$*, entonces* $\overrightarrow{Y} = A \cdot \overrightarrow{X} \stackrel{d}{=} N_p(A \cdot \overrightarrow{\mu}, A \cdot \Sigma \cdot A^t)$.

Dada una variable aleatoria $\overrightarrow{X} \stackrel{\mathrm{d}}{=} N_d(\overrightarrow{\mu}, \Sigma)$, entonces $\overrightarrow{X'} = \overrightarrow{X} - \overrightarrow{\mu} \stackrel{\mathrm{d}}{=} N_d\left(\overrightarrow{0}, \Sigma\right)$. Además, dada una matriz definida positiva, Σ, existe una matriz ortogonal H (es decir, tal que $H^{-1} = H^t$) de forma que la matriz $\Lambda = H^t \Sigma H$ es diagonal. De hecho, H es la matriz de cambio de base para que la matriz asociada a la correspondiente aplicación lineal sea la matriz diagonal Λ (en lugar de la matriz de partida Σ).

Las columnas de la matriz H son precisamente los autovectores linealmente independientes (y de módulo unitario) de la matriz Σ, es decir, d vectores linealmente independientes, $\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_d}$, tales que $\overrightarrow{x_i}^t \overrightarrow{x_i} = 1$ para todo $i = 1, 2, \ldots, n$ y con $\overrightarrow{x_i}^t \overrightarrow{x_j} = 0$ si $i \neq j$, verificando además que $\exists \lambda_i \in \mathbb{R}$ tal que $\Sigma \cdot \overrightarrow{x_i} = \lambda_i \overrightarrow{x_i}$ (condición de ser un autovector). Además los autovalores $\lambda_1, \lambda_2, \ldots, \lambda_d$ (que son todos positivos) son precisamente los elementos de la diagonal de la matriz Λ.

Partiendo de una variable $\overrightarrow{Z} \stackrel{\mathrm{d}}{=} N_d\left(\overrightarrow{0}, I_d\right)$ (fácilmente simulable a partir de $Z_1, Z_2, \ldots, Z_d \sim N(0,1)$ independientes), se tiene que $\Lambda^{1/2} \overrightarrow{Z} \stackrel{\mathrm{d}}{=} N_d\left(\overrightarrow{0}, \Lambda\right)$, siendo

$$\Lambda^{1/2} = \begin{pmatrix} \lambda_1^{1/2} & 0 & \cdots & 0 \\ 0 & \lambda_2^{1/2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d^{1/2} \end{pmatrix}.$$

Ahora, multiplicando por la izquierda por la matriz H, se tiene

$$H\Lambda^{1/2} \overrightarrow{Z} \stackrel{\mathrm{d}}{=} N_d\left(\overrightarrow{0}, H\Lambda H^t\right) \stackrel{\mathrm{d}}{=} N_d\left(\overrightarrow{0}, \Sigma\right).$$

Finalmente, basta sumar el vector $\overrightarrow{\mu}$ para obtener

$$\overrightarrow{X} = \overrightarrow{\mu} + H\Lambda^{1/2} \overrightarrow{Z} \stackrel{\mathrm{d}}{=} N_d(\overrightarrow{\mu}, \Sigma).$$

Una vez diagonalizada la matriz Σ, con autovalores $\lambda_1, \lambda_2, \ldots, \lambda_d$ y autovectores asociados dados por las columnas de la matriz H, el algoritmo procedería como sigue:

1. `Simular` $Z_1, Z_2, \ldots, Z_d \sim N(0,1)$ `independientes.`
2. `Para` $i = 1, 2, \ldots, d$ `hacer` $Y_i = \sqrt{\lambda_i} Z_i$.
3. `Devolver` $\overrightarrow{X} = \overrightarrow{\mu} + H\overrightarrow{Y}$.

Ejemplo 28 *Dar un algoritmo para simular la distribución*

$$N_2\left(\begin{pmatrix} 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 2.36 & -0.48 \\ -0.48 & 2.64 \end{pmatrix}\right).$$

Para encontrar los autovalores y autovectores de Σ *resolvemos* $\det(\Sigma - \lambda I) = 0$*, es decir,*

$$\begin{vmatrix} 2.36-\lambda & -0.48 \\ -0.48 & 2.64-\lambda \end{vmatrix} = 0 \Leftrightarrow (2.36-\lambda)(2.64-\lambda) - (-0.48)^2 = 0$$

$$\Leftrightarrow \lambda^2 - 5\lambda + 6 = 0 \Leftrightarrow \lambda = \frac{5 \pm \sqrt{5^2 - 6 \cdot 4}}{2},$$

que ofrece como soluciones $\lambda_1 = 3$ *y* $\lambda_2 = 2$*. Para encontrar autovalores de módulo* 1 *correspondientes a esos autovalores no tenemos más que resolver los sistemas* $(\Sigma - \lambda_i I)\,\overrightarrow{x} = \overrightarrow{0}$ *para* $i = 1, 2$ *imponiendo la condición de módulo igual a* 1*, es decir* $x_1^2 + x_2^2 = 1$*. Así, resulta*

$$\Sigma - \lambda_1 I = \begin{pmatrix} -0.64 & -0.48 \\ -0.48 & -0.36 \end{pmatrix} = -0.04 \begin{pmatrix} 16 & 12 \\ 12 & 9 \end{pmatrix}, \text{ luego}$$

$$(\Sigma - \lambda_1 I)\,\overrightarrow{x} = \overrightarrow{0} \Leftrightarrow x_2 = -\frac{4}{3}x_1, \text{ pero como } x_1^2 + x_2^2 = 1, \text{ se tiene}$$

$$\frac{25}{9}x_1^2 = 1, \text{ luego } x_1 = \frac{3}{5} \text{ y } x_2 = -\frac{4}{5}$$

(también es solución si cambiamos ambos de signo);

$$\Sigma - \lambda_2 I = \begin{pmatrix} 0.36 & -0.48 \\ -0.48 & 0.64 \end{pmatrix} = 0.04 \begin{pmatrix} 9 & -12 \\ -12 & 16 \end{pmatrix}, \text{ luego}$$

$$(\Sigma - \lambda_2 I)\,\overrightarrow{x} = \overrightarrow{0} \Leftrightarrow x_2 = \frac{3}{4}x_1, \text{ pero como } x_1^2 + x_2^2 = 1, \text{ se tiene}$$

$$\frac{25}{16}x_1^2 = 1, \text{ luego } x_1 = \frac{4}{5} \text{ y } x_2 = \frac{3}{5}.$$

De esta forma, la matriz H *resulta, entre otras posibilidades,*

$$H = \begin{pmatrix} \frac{3}{5} & \frac{4}{5} \\ -\frac{4}{5} & \frac{3}{5} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.8 \\ -0.8 & 0.6 \end{pmatrix}.$$

Ahora

$$\overrightarrow{Y} = \Lambda^{1/2}\overrightarrow{Z} = \begin{pmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \sqrt{3}Z_1 \\ \sqrt{2}Z_2 \end{pmatrix}$$

y finalmente,

$$\overrightarrow{X} = \overrightarrow{\mu} + H\overrightarrow{Y} = \begin{pmatrix} 1 + 0.6Y_1 + 0.8Y_2 \\ 3 - 0.8Y_1 + 0.6Y_2 \end{pmatrix}.$$

Así, el algoritmo resultaría

1. *Simular* $Z_1, Z_2 \sim N(0,1)$ *independientes.*
2. *Hacer* $Y_1 = \sqrt{3}Z_1$ *e* $Y_2 = \sqrt{2}Z_2$.
3. *Obtener* $X_1 = 1 + 0.6Y_1 + 0.8Y_2$ *y* $X_2 = 3 - 0.8Y_1 + 0.6Y_2$.
4. *Devolver* $\overrightarrow{X} = (X_1, X_2)^t$.

6.5 Ejercicios propuestos

1. La diferencia en precio (X) y en tiempo de vida (Y) de un monitor de 17 pulgadas, expresados en tanto por uno con respecto a la media de mercado, puede suponerse una variable aleatoria, (X, Y), con función de densidad dada por

$$f(x,y) = \begin{cases} \dfrac{3}{16}\left[2 - (x^2 + y^2)\right] & \text{si} \quad x \in [-1,1] \text{ e } y \in [-1,1] \\ 0 & \text{en otro caso} \end{cases}$$

   Simular dicha distribución bidimensional mediante el método de aceptación-rechazo utilizando como densidad auxiliar una de muy sencilla simulación. ¿Cuál es la eficiencia del algoritmo? ¿Existe algún otro método alternativo para simular esta distribución? Dar, muy brevemente, la idea en que consiste alguno de ellos.

2. Dar una formulación general de un algoritmo de inversión para una distribución bidimensional.

3. La variable aleatoria bidimensional (X, Y), que representa el número de reinicios de un sistema informático y el número de utilizaciones del SAI (sistema de alimentación ininterrumpida), tiene la siguiente masa de probabilidad:

$P(X = i, Y = j)$	$i = 0$	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$j = 0$	0.15	0.23	0.06	0.03	0.01
$j = 1$	0.09	0.18	0.05	0.02	0
$j = 2$	0.01	0.05	0.11	0.01	0

   Dar un algoritmo para simular el par de variables (X, Y). Comentar la eficiencia del método. ¿Podría mejorarse dicha eficiencia si deseásemos simular solamente la variable X?

7

Diseño de experimentos de simulación

En el presente capítulo se abordarán algunas de las cuestiones más importantes a la hora de diseñar un estudio de simulación. En primer lugar se pondrán de manifiesto las similitudes y diferencias entre la simulación y la experimentación sobre el sistema real. También se comentarán algunas generalidades sobre la simulación estática y dinámica, diferenciando, dentro de ésta última, entre la simulación por eventos y por cuantos. Se abordarán algunas técnicas de reducción de la varianza y se indicará como poder resolver los problemas de estabilización y dependencia que habitualmente aparecen en la simulación dinámica.

7.1 Diferencias y similitudes con la experimentación real

Como ya se definió en el primer tema, la simulación es la técnica consistente en la realización de experimentos de muestreo sobre un modelo construído a partir de un sistema real. A partir de esta definición es obvio que la simulación necesitará de gran cantidad de técnicas estadísticas para obtener las muestras (muestreo) y para analizar los resultados obtenidos por la experimentación artificial (estimación, intervalos de confianza, contrastes de hipótesis, etc.). Por todo ello, puede afirmarse que, en general, en cuanto a la utilización de técnicas estadísticas es muy similar a la propia experimentación sobre el sistema real.

Entre las diferencias caben destacar las siguientes:

1. La utilización de técnicas de estimación puntual, construcción de intervalos de confianza y contrastes de hipótesis es algo menos frecuente en la simulación que en la experimentación real. La razón es que algunos de los parámetros (los de control) ya son conocidos en la simulación y, por tanto, no es necesario hacer inferencia sobre ellos, aunque sí sobre los de salida, que miden, de alguna forma, el comportamiento del sistema.

2. La simulación suele hacer un uso mucho más intensivo de técnicas de ordenación y optimización. Esto es debido a que, en el contexto de la simulación, es factible comparar un grandísismo número de escenarios (entre los que se desea optimizar, por ejemplo) en muy poco tiempo, cosa que se da muy raramente en la experimentación real.

3. Una peculiaridad de la simulación es que casi siempre es posible comparar distintas estrategias sobre las mismas muestras simuladas (simplemente

utilizando la misma semilla en la simulación, convenientemente planificada). Esto implica que se podrán usar muestras de datos apareados (positivamente correlados) en lugar de datos independientes, con la consiguiente ganacia en eficiencia (como se verá al estudiar la utilización de números aleatorios comunes, dentro de la sección de técnicas de reducción de la varianza).

7.2 Simulación estática y dinámica

La simulación se dice estática si en el modelo no juega ningún papel el transcurso del tiempo mientras que es dinámica si el tiempo es una de las variables importantes del modelo. En la simulación estática resulta muy sencillo comparar distintas estrategias ante las mismas condiciones del azar, mientras que esto es más complicado en la simulación dinámica, exigiendo un trabajo mayor de planificación. Además, el coste computacional de la simulación estática es bastante más moderado.

La simulación estática se usa muy frecuentemente en estadística para comprobar el comportamiento comparativo de diversos métodos estadísticos alternativos para tamaños muestrales finitos (complementando los estudios teóricos, casi siempre asintóticos).

En la simulación dinámica, normalmente se trata de ir analizando los distintos estados por los que va pasando un sistema que evoluciona en el tiempo. Esto provoca, en general, un mayor coste computacional y problemas de estabilización y dependencia. Existen dos grandes tipos de simulación dinámica: la simulación continua, en la que se supone que el sistema cambia de estado constantemente y la simulación discreta, para la cual los cambios se producen en ciertos instantes de tiempo singulares. La razón de sus nombres viene de que en el primer caso el conjunto de estados es continuo, mientras que en el segundo es discreto. Dentro de la simulación discreta distinguiremos la simulación por eventos y la simulación por cuantos.

7.3 Simulación por eventos y por cuantos

Con el nombre de simulación por eventos, o asíncrona, designamos el tipo de simulación dinámica discreta en la cual se controla la variable tiempo moviéndola hasta la ocurrencia del siguiente suceso (o evento). Esto implica la necesidad de controlar minuciosamente cuál es dicho próximo suceso: saber cuáles son los posibles sucesos en un futuro inmediato y cuál de ellos es el más inmediato.

La simulación por cuantos, o asíncrona, responde a una filisofía totalmente diferente. Se trata de examinar el sistema (que evoluciona en el tiempo) dejando pasar pequeños intervalos de tiempo de longitud δ, fija, (llamada cuanto) en los cuales se supone que, a lo sumo, un sólo suceso puede producirse.

En general, la simulación por eventos es exacta y de más difícil implementación, pero de mucha más rápida ejecución que la simulación por cuantos. Sin embargo esta última es muchas veces la única posibilidad factible en la simulación dinámica continua.

7.4 Técnicas de reducción de la varianza

Existe un sinfín de técnicas encaminadas a reducir la varianza en un estudio de simulación o bien a tratar de estimarla. Algunas de ellas son el uso de números aleatorios comunes, la utilización de variables antitéticas, la estratificación, el uso de variables de control, el método Jackknife, los métodos de remuestreo (destacando entre ellos el método bootstrap), etc. En lo que sigue estudiaremos los tres primeros métodos reseñados. En general conviene tener en cuenta que si uno de los objetivos de la simulación es precisamente estimar la variabilidad, no conviene utilizar estas técnicas de reducción de la varianza. Éstas son aplicables normalmente cuando la simulación pretende ofrecer respuestas, lo más precisas posibles, sólo sobre cantidades medias.

7.4.1 Números aleatorios comunes

Supóngase que se desea comparar dos estrategias distintas mediante N repeticiones de un experimento de simulación, de las cuales se han obtenido los valores numéricos de salida: $X_1, X_2, \ldots, X_N$, para la primera e $Y_1, Y_2, \ldots, Y_N$, para la segunda. Usando los mismos números aleatorios (es decir repitiendo los cálculos con la misma semilla) en las variables de entrada de la simulación, se tiene que $Cov(X_i, Y_i) > 0$. Si se quiere estimar la diferencia de las medias de esta variable de salida para ambas estrategias: $E(X) - E(Y) = E(X - Y)$ puede usarse $\overline{X} - \overline{Y} = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)$, cuya varianza viene dada por

$$\begin{aligned} Var\left(\overline{X} - \overline{Y}\right) &= \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i - Y_i) = \frac{1}{N} Var(X_1 - Y_1) \\ &= \frac{1}{N}\left(Var(X_1) + Var(Y_1) - 2Cov(X_1, Y_1)\right) \\ &\leq \frac{1}{N}\left(Var(X_1) + Var(Y_1)\right), \end{aligned}$$

que es la varianza que tendría $\overline{X}-\overline{Y}$ en caso de haber usado muestras independientes para cada estrategia. Esto demuestra que es más ventajoso, a efectos de reducir la varianza de la estimación, el uso de números aleatorios comunes. Este hecho es bien conocido en estadística cuando se diseña un experimento de comparación de dos poblaciones mediante la obtención de las llamadas muestras apareadas, en lugar de la utilización de muestras independientes.

7.4.2 Variables antitéticas

Supóngase ahora que se desea evaluar el resultado de una única estrategia (sin compararla con ninguna otra alternativa). Despues de N repeticiones de la simulación, tendremos N valores numéricos $X_1,\ X_2,\ \ldots,\ X_N$, procediendo a estimar la media $E(X)$ teórica mediante $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Como es sabido, dado que éste es un estimador insesgado, su precisión puede medirse mediante $Var\left(\overline{X}\right)$. Si las variables son independientes se tiene que $Var\left(\overline{X}\right) = \frac{1}{N^2}\sum_{i=1}^{N} Var\left(X_i\right)$, mientras que, en general, se tiene

$$Var\left(\overline{X}\right) = \frac{1}{N^2}\left(\sum_{i=1}^{N} Var\left(X_i\right) + 2\sum_{i,j=1,i<j}^{n} Cov\left(X_i, X_j\right)\right).$$

Una forma de utilizar esta última expresión para reducir la varianza del estimador consiste en hacer que cada variable con índice impar sea negativamente correlada con la variable de índice siguiente (siendo independientes de las demás). La forma más sencilla de conseguir esto cuando se utiliza el método de inversión para simular las X_i consiste en tomar un valor $U \sim U(0,1)$ para simular X_{2i-1} y el valor $1-U$ para simular X_{2i}, su variable antitética, para $i = 1, 2, \ldots, \frac{N}{2}$ (si N es par). El procedimiento es algo más complicado con otros métodos de simulación distintos del de inversión.

7.4.3 Estratificación

En ocasiones conviene dividir la población en estratos obteniendo, del total de la muestra, cierto número de observaciones de cada estrato (proporcional a la probabilidad de cada uno). Veamos un ejemplo.

Ejemplo 29 (muestreo estratificado de una exponencial) *Supóngase el siguiente problema (absolutamente artificial pero ilustrativo para comprender esta técnica). Dada una muestra de tamaño 10 de una población con distribución* $\exp(1)$ *se desea estimar la media poblacional (como se acaba de comentar el problema es artificial, pues es sobradamente conocido que dicha media es 1). Si pretendemos evitar que, por puro azar, exista alguna zona, en la que*

la exponencial toma valores, no representada en la muestra de 10 datos podemos proceder de la siguiente forma. Tomemos tres estratos, por ejemplo, el del 40% de valores menores, el siguiente 50% de valores (intermedios) y el 10% de valores más grandes para esta distribución.

Como el algoritmo de inversión (optimizado) para simular la $\exp(1)$ *es*

1. **Generar** $U \sim U(0,1)$.
2. **Hacer** $X = -\ln U$.

la forma de garantizar el que obtengamos 5, 4 y 1 valores, repectivamente, en cada uno de los tres estratos consiste en elegir $U \in [0.6, 1)$*, en el primer caso,* $U \in [0.1, 0.6)$*, en el segundo y* $U \in [0, 0.1)$ *para el tercer estrato. Dado que, en principio, simulando diez valores* $U_1, U_2, \ldots, U_{10} \sim U(0,1)$*, no hay nada que nos garantice que las proporciones de los estratos son las deseadas (aunque sí lo sean en media) una forma de proceder consiste en rechazar valores de* U *que caigan en uno de esos tres intervalos cuando el cupo de ese estrato esté ya lleno. Esto es lo mismo que simular 4 valores de* $U|_{U\in[0.6,1)} \stackrel{d}{=} U[0.6,1)$ *para el primer estrato, 5 valores de* $U|_{U\in[0.1,0.6)} \stackrel{d}{=} U[0.1,0.6)$ *para el segundo y uno de* $U|_{U\in[0,0.1)} \stackrel{d}{=} U[0,0.1)$ *para el tercero. El algoritmo con esta estratificación sería como sigue:*

1. **Generar** $U_i \sim U(0,1)$ **para** $i = 1, 2, \ldots, 10$.
2. **Si** $i \leq 4$ **entonces hacer** $U_i = 0.4 \cdot U_i + 0.6$.
3. **Si** $4 < i \leq 9$ **entonces hacer** $U_i = 0.5 \cdot U_i + 0.1$.
4. **Si** $i = 10$ **entonces hacer** $U_i = 0.1 \cdot U_i$.
5. **Desde** $i = 1$ **hasta** 10 **devolver** $X_i = -\ln U_i$.

No es difícil probar que $Var(X_i) = 0.0214644$ *si* $i = 1, 2, 3, 4$, $Var(X_i) = 0.229504$ *si* $i = 5, 6, 7, 8, 9$ *y* $Var(X_{10}) = 1$*. Como consecuencia,*

$$Var\left(\overline{X}\right) = \frac{1}{10^2}\sum_{i=1}^{10} Var(X_i) = 0.022338$$

que es bastante menor que 1, que es la varianza en el caso de muestreo aleatorio simple no estratificado.

7.5 Problemas de estabilización y dependencia

Ambas cuestiones suelen plantearse en la simulación dinámica. Los problemas de estabilización están relacionados con el hecho de que, en ocasiones, el sistema evoluciona en el tiempo de tal forma que tiene una distribución estacionaria que se supone de partida pero que puede ser muy sensible a las condiciones iniciales con las que se comience la simulación. En tal caso resulta conveniente el transcurso de cierto período de tiempo (denominado período de estabilización) durante el cual los resultados obtenidos para las variables

de salida son ignorados y cuyo único objeto es conseguir que se estabilice la distribución de probabilidad.

Ejemplo 30 *Supongamos el siguiente modelo de simulación:*

$$X_t = 10 + 0.7 \cdot (X_{t-1} - 10) + \varepsilon_t$$

para explicar la temperatura, X_t, tomada a las 12 a.m. en el día t, donde ε_t es un error aleatorio con distribución $N(0,1)$. Parece evidente que, en un modelo como éste, es crucial el valor de la condición inicial X_0 correspondiente al origen de tiempos. En otras palabras, tomando para X_0 un valor muy lejano a aquellos más probables bajo la distribución estacionaria (por ejemplo $X_0 = 100$), es intutitivo que se necesitaría de una etapa inicial considerable para llegar a alcanzar valores estacionarios. Por ejemplo, suponiendo que los ε_t fuesen cero (que aunque no es cierto, realmente son bastante pequeños en relación con el valor 100), se obtendría la siguiente sucesión de valores: $X_0 = 100$, $X_1 = 73$, $X_2 = 54.1$, $X_3 = 40.87$, $X_4 = 31.7$, $X_5 = 25.4$, ... El período de estabilización sería mucho menor si se partiese de un valor inicial más cercano a 10.

Los problemas de dependencia son aquellos derivados del hecho de que frecuentemente (de nuevo en modelos de simulación dinámica) las distintas variables de salida de la simulación son dependientes. En el ejemplo anterior es obvio que cada valor X_t depende de X_{t-1} (incluso de X_{t-2} y de otras anteriores, aunque cada vez en menor medida). Esto afecta fundamentalmente a la precisión de los estimadores construídos con observaciones de las mismas. Una forma de atenuar este efecto sería considerar observaciones de las mismas en instantes temporalmente lejanos (donde se supone que la dependencia es mucho más débil). En ocasiones, más que atenuar este efecto se trata de estimar la precisión del estimador resultante. Obviamente, para ello ha de tenerse en cuenta la dependencia.

7.6 Ejercicios propuestos

1. Considérese el algoritmo encontrado en el ejercicio 3, propuesto en el tema 3. Si posteriormente se quiere proceder de forma que los impactos cercanos y lejanos al centro estén equilibrados (es decir, si por cada uno cercano queremos tener otro lejano), ¿cómo se podrían simular ahora las distancias al centro de 10 impactos?

2. Una nueva componente llega cada 40 segundos a una cadena de ensamblado. El tiempo necesario para ensamblar dicha componente a la pieza matriz se supone una variable aleatoria con media de 30 segundos

(resultaría muy importante haber estimado su distribución: $N(30,1)$, $N(30,5)$, $\Gamma(2,60)$, etc.). Si un tiempo de ensamblado es superior a 40 segundos, las componentes que van llegando se acumulan hasta que se les vaya dando salida. Dar un algoritmo que permita responder a las siguientes preguntas ¿cuál es la probabilidad de que una componente que llega tenga que esperar?, ¿cuál es la probabilidad de que estén más de tres piezas esperando?

3. El número de toneladas de pan producidas diariamente por una empresa panificadora tiene distribución de Pareto con parámetros 3 y 2. Suponiendo que la demanda diaria de pan (en toneladas) tiene distribución $N(3,0.5)$ y que es independiente de la producción, calcular la probabilidad de que un día concreto la demanda no sea satisfecha. Suponiendo una pérdida de 0.05 euros por cada kilo de pan no vendido y una penalización de 0.2 euros por cada kilo de pan demandado por clientes y no entregado (por falta de existencias), dar un algoritmo para calcular las pérdidas medias diarias en concepto de excedentes o demandas insatisfechas.

4. En una cadena productiva, el tiempo que tarda una componente en llegar a una máquina ensambladora desde que llegó la pieza anterior sigue una distribución exponencial de media 10 segundos. El tiempo (en segundos) que emplea la máquina ensambladora con cada componente se ajusta a una distribución $\Gamma(3,20)$. Cada vez que una nueva componente llega a la fase de ensamblaje y la máquina está ocupada, dicha pieza se desvía a otra linea de producción distinta. Dar un algoritmo, lo más preciso posible, para poder aproximar mediante simulación el porcentaje de tiempo que la máquina ensambladora perderá esperando la llegada de una nueva pieza. Justificar el funcionamiento de dicho algoritmo.

5. Una máquina produce tiras de goma de longitud aleatoria con distribución $\exp(2)$ (en metros). Después de fabricadas, las tiras de goma se pasan a otra máquina que las estira hasta que rompen en dos (para comprobar su elasticidad). Suponiendo que el lugar por donde se rompe cada tira es aleatorio y con distribución uniforme a lo largo de toda su longitud, describir detalladamente un algoritmo que simule las longitudes de los dos trozos en los que se rompe cada goma.

PARTE II

TEORÍA DE COLAS

8

Introducción a la teoría de colas

La teoría de colas es una disciplina, dentro de la Investigación Operativa, que tiene por objeto el estudio y análisis de situaciones en las que existen entes que demandan cierto servicio, de tal forma que dicho servicio no puede ser satisfecho instantáneamente, por lo cual se provocan esperas.

Tal y como queda patente en la definición anterior, el ámbito de aplicación de la teoría de colas es enorme: desde las esperas para ser atendidos en establecimientos comerciales, esperas para ser procesados determinados programas informáticos, esperas para poder atravesar un cruce los vehículos que circulan por una ciudad o esperas para establecer comunicación o recibir información de un servidor web, a través de internet, entre muchas otras.

8.1 Reseña histórica

Históricamente, los primeros trabajos que comenzaron a dar cuerpo a la teoría de colas (también llamada Teletráfico en Ingeniería de Telecomunicaciones) son los debidos al matemático-ingeniero danés A.K. Erlang, quien en 1909 publicó *La teoría de probabilidades y las conversaciones telefónicas*. Erlang era por 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*, seguido, en 1928, de *Probabilidad y sus usos en Ingeniería*, por T.C. Fry. 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.

A pesar de que a comienzos del estudio de la teoría, 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 Informática, las Telecomunicaciones y, en general, las nuevas tecnologías abren un aún mayor porvenir a esta teoría matemática.

8.2 Contenidos de la parte de teoría de colas

Los contenidos de teoría de colas que se tratarán en los siguientes capítulos son los siguientes:

1. Estudio de las herramientas probabilísticas más importantes para el tratamiento de las colas con llegadas y servicios exponenciales: nociones generales sobre procesos estocásticos, el proceso de Poisson y procesos de nacimiento y muerte.

2. Introducción a la teoría de colas: describiendo los elementos del sistema de una cola, estableciendo la terminología habitual, estudiando algunas propiedades de interés de las distribuciones exponencial y gamma y dando las fórmulas de Little.

3. Los modelos de colas más importantes con tasas de llegada y de servicio exponenciales: M/M/1, M/M/s, M/M/1/K, M/M/s/K, M/M/1/∞/H, M/M/S/∞/H (sin y con repuestos) y M/M/∞.

4. Las redes de colas tanto, abiertas como cerradas, con distribución exponencial para los servicios y para las llegadas (si es el caso).

5. Algunas indicaciones sobre cómo resolver problemas de colas cuando la hipótesis de exponencialidad deja de ser cierta. Esto incluye el estudio analítico de modelos sencillos con un sólo servidor, como el M/G/1, y algunas nociones sobre cómo aproximar los valores de interés mediante simulación.

9

Nociones básicas sobre procesos estocásticos

9.1 Noción de proceso estocástico

Definición 31 *Un proceso estocástico no es más que una colección de variables aleatorias que denotaremos por* $\{X(t)/t \in T\}$*, o bien,* $\{X_t/t \in T\}$*, o simplemente* $\{X_t\}_{t\in T}$*. Es decir, cada* X_t *es una variable aletoria* $X_t : \Omega \to \mathbb{R}$*, todas ellas definidas sobre el mismo espacio de probabilidad* Ω*. El subconjunto,* E*, de* $\mathbb{R}$ *en el que toman valores todas las variables se conoce como espacio de estados, mientras que el conjunto de índices,* T*, se denomina espacio de tiempos. Según los conjuntos* E *y* T *sean finitos (o infinitos numerables) o contengan, al menos, un intervalo se hablará de procesos estocásticos con espacio de estados discreto o continuo y en tiempo discreto o continuo.*

Algunos ejemplos teóricos de procesos estocásticos son: una variable aleatoria (aquí el conjunto de índices tiene un único elemento), un vector aleatorio -o variable aleatoria multidimensional- (en el cual el espacio de tiempos es finito), o una sucesión de variables aleatorias (cuando $T = \mathbb{N}$), entre otros. Como es obvio, la mayor novedad del concepto se da cuanto el espacio de tiempos es un conjunto infinito.

Ejemplos de situaciones reales que podrían modelizarse mediante procesos estocásticos son las siguientes:

1. El tiempo de respuesta en un sistema informático multitusuario, según la hora del día. Aquí el espacio de estados podría ser $E = (0, \infty)$ y el de tiempos $T = [0, 24)$, por tanto se trataría de un proceso estocástico en tiempo continuo y con espacio de estados continuo.

2. El número de terminales conectadas a un servidor según el instante del día. De nuevo, $T = [0, 24)$, mientras que ahora $E = \{0, 1, 2, \dots\}$. Se trata pues de un proceso estocástico en tiempo continuo y con espacio de estados discreto.

3. El tiempo de CPU de un servidor dedicado a usuarios según el día del año. Se trata de un proceso estocástico en tiempo discreto y con espacio de estados continuo, pues $T = \{1, 2, \dots, 365\}$ y $E = [0, \infty)$. También, obviamente, es una variable aletoria continua 365-dimensional.

Fijado un elemento del espacio muestral, $\omega \in \Omega$, la función $g_\omega : T \to E$, que a cada $t \in T$ le hace corresponder $g_\omega(t) = X_t(\omega)$ se denomina trayectoria, o realización muestral del proceso, para la elección del azar (o suceso elemental) ω. En este sentido, un proceso estocástico también puede definirse, de forma

equivalente, como una única "variable aletoria" pero que toma valores en el conjunto de las funciones de T en E, $\mathcal{G} = \{g/g : T \to E\}$.

9.2 Características y propiedades que puede verificar un proceso estocástico

Definición 32 *Se denomina función de medias de un proceso estocástico* $\{X_t\}_{t\in T}$ *a la aplicación definida por*

$$m(t) = E(X_t), \forall t \in T.$$

Definición 33 *La función de autocovarianzas de un proceso* $\{X_t\}_{t\in T}$ *es una aplicación, de* $T \times T$ *en* $\mathbb{R}$*, dada por*

$$c(s,t) = Cov(X_s, X_t) = E[(X_s - m(s))(X_t - m(t))], \text{ para todos } s,t \in T.$$

Debido a la simetría de la covarianza, esta función es simétrica, es decir, $c(s,t) = c(t,s)$ para todos $s,t \in T$.

Definición 34 *Un proceso estocástico* $\{X_t\}_{t\in T}$ *se dice débilmente estacionario (o estacionario en media y covarianza) si verifica*

1. Su función de medias es constante, i.e., $m(t) = m$, $\forall t \in T$.
2. La función de autocovarianzas es función solamente de la diferencia de instantes de tiempo, es decir, $c(s,t) = \gamma(t-s)$.

Definición 35 *Un proceso estocástico* $\{X_t\}_{t\in T}$ *se dice fuertemente estacionario (o estacionario en sentido estricto) si para cualquier* $n \in \mathbb{N}$*, cualesquiera instantes verificando* $t_1 < t_2 < \cdots < t_n$ *y cualquier valor real* h*, tal que* $t_i + h \in T$ *para* $i = 1, 2, \ldots, n$*, se verifica que la variable* n*-dimensional*

$$(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h})$$

tiene la misma distribución de probabilidad que

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_n}).$$

A partir de esta definición se tiene que, en particular, todo proceso estocástico fuertemente estacionario verifica $X_s \overset{\mathrm{d}}{=} X_t$ para todos $s,t \in T$ (es decir las distribuciones marginales de las variables del proceso son iguales). Lo anterior implica que

$$m(s) = E(X_s) = E(X_t) = m(t), \text{ para todos } s,t \in T.$$

Además, aplicando para $n = 2$ y $h = -s$, la propiedad de estacionariedad fuerte, se tiene que

$$(X_s, X_t) \overset{\mathrm{d}}{=} (X_0, X_{t-s})$$

y, por tanto $c(s,t) = c(0, t-s) = \gamma(t-s)$. En virtud de todo lo anterior, se tiene que todo proceso fuertemente estacionario también es débilmente estacionario.

Una propiedad muy importante que puede verificar un proceso estocástico es la llamada propiedad de Markov.

Definición 36 *Un proceso estocástico $\{X_t\}_{t\in T}$, se dice markoviano si, para cualquier $n \in \mathbb{N}$ y cualesquiera instantes $t_1 < t_2 < \cdots < t_n$ en T, se verifica la siguiente igualdad en distribución:*

$$X_{t_n}|_{X_{t_1}, X_{t_2}, \ldots, X_{t_{n-1}}} \overset{d}{=} X_{t_n}|_{X_{t_{n-1}}}$$

En términos intuitivos, la propiedad de Markov se expresa diciendo que, conocido el presente, la distribución de probabilidad de posibles valores futuros del proceso depende sólamente del valor del presente y no de valores del proceso en el pasado.

En lo que sigue nos centraremos en el estudio de algunos procesos estocásticos, en tiempo continuo y con espacio de estados discreto, que serán de especial interés por su utilidad como herramientas para la teoría de colas.

9.3 Procesos de contar: el proceso de Poisson

Definición 37 *Un proceso estocástico en tiempo continuo y con espacio de estados discreto, $\{N_t\}_{t\in[0,\infty)}$, se dice de contar si verifica los siguientes tres axiomas:*

1. $N_0 = 0$.
2. *N_t toma únicamente valores enteros no negativos.*
3. *Si $s < t$ entonces $N_s(\omega) \leq N_t(\omega)$ para todo $\omega \in \Omega$.*

Algunas situaciones prácticas que pueden modelizarse mediante procesos de contar son

1. El número de procesos enviados a un servidor hasta un tiempo t.
2. El número de llamadas telefónicas recibidas por una oficina de reserva de billetes hasta un tiempo t.

3. El número de errores de software habidos en un sistema informático hasta un instante t.

En general, puede afirmarse que los procesos de contar sirven para modelizar el número de ocurrencias de un cierto fenómeno habidas desde un inicio de tiempos (fijado en $t = 0$) hasta un instante t.

Definición 38 *Dado un proceso de contar, $\{N_t / t \geq 0\}$, se dice que es un proceso de Poisson de parámetro (o intensidad) $\lambda > 0$, si verifica las siguientes propiedades:*

1. *El proceso tiene incrementos independientes, es decir, si se tienen instantes $0 \leq t_0 < t_1 < \cdots < t_n$, entonces, las variables aleatorias $N_{t_1} - N_{t_0}$, $N_{t_2} - N_{t_1}, \cdots, N_{t_n} - N_{t_{n-1}}$ (que representan los números de ocurrencias en los intervalos $(t_0, t_1]$, $(t_1, t_2], \cdots, (t_{n-1}, t_n]$) son independientes.*

2. *El proceso tiene incrementos estacionarios, es decir, $N_{t+h} - N_{s+h} \stackrel{d}{=} N_t - N_s$ para todos $s < t$ y h arbitrario.*

3. *La probabilidad de que en el intervalo $[0, h]$ se de exactamente una ocurrencia del fenómeno es $P(N_h = 1) = \lambda h + o(h)$. En virtud del axioma anterior esto mismo es válido para cualquier intervalo de longitud h.*

4. *La probabilidad de que haya dos o más ocurrencias en ese mismo intervalo $[0, h]$ es $P(N_h \geq 2) = o(h)$. De nuevo, por el axioma 2, esta propiedad es válida para cualquier intervalo de longitud h.*

Como consecuencia de los dos últimos axiomas de la definición, se tiene $P(N_h = 0) = 1 - \lambda h + o(h)$. Por otra parte, el número medio de ocurrencias del fenómeno en un intervalo de longitud h, viene dado por

$$\begin{aligned} E(N_{t+h} - N_t) &= E(N_h) = 0 \cdot (1 - \lambda h + o(h)) + 1 \cdot (\lambda h + o(h)) + o(h) \\ &= \lambda h + o(h), \end{aligned}$$

con lo que el número medio de ocurrencias por unidad de tiempo es

$$\frac{E(N_{t+h} - N_t)}{h} = \frac{\lambda h + o(h)}{h} = \lambda + \frac{o(h)}{h},$$

cuyo límite, cuando $h \to 0^+$ es λ. Así, el parámetro del proceso de Poisson, λ, puede interpretarse como el número medio de ocurrencias del fenómeno por unidad de tiempo (de ahí el nombre de intensidad).

Seguidamente se presentan algunos resultados teóricos interesantes para el proceso de Poisson. Como consecuencia de los mismos se tiene, entre otras cosas, que la distribución de un proceso de Poisson queda caracterizada de forma única por los axiomas anteriores.

Teorema 39 *Sea $\{N_t/t \geq 0\}$ un proceso de Poisson de parámetro λ. Entonces la variable aleatoria N_t tiene distribución de Poisson de parámetro λt. En otras palabras, las distribuciones marginales del proceso son de Poisson.*

Teorema 40 *Sea $\{N_t/t \geq 0\}$ un proceso de Poisson de parámetro λ y sean $0 = t_0 < t_1 < t_2 < \cdots$ los tiempos (aleatorios) de ocurrencia del fenómeno en cuestión (es decir, los puntos $t \geq 0$ que verifican que $N_s < N_t$ para todo $s < t$) y defínanse los tiempos entre ocurrencias $\tau_1 = t_1 - t_0$, $\tau_2 = t_2 - t_1$, $\cdots$; entonces las variables aleatorias τ_k son mutuamente independientes e idénticamente distribuídas, con distribución exponencial de parámetro λ. Recíprocamente, si $\{N_t/t \geq 0\}$ es un proceso de contar y los tiempos entre ocurrencias, τ_k, definidos como antes, son variables independientes e idénticamente distribuídas, con distribución exponencial de parámetro λ, entonces $\{N_t/t \geq 0\}$ es un proceso de Poisson de parámetro λ.*

Este resultado es muy útil para simular el proceso de Poisson sin más que utilizar un algoritmo (como el de inversión) que permita simular la distribución exponencial. También, junto con el resultado anterior, es de mucho interés teórico pues simplifica notablemente el cálculo de diversas probabilidades relativas a procesos de Poisson. Un ejemplo de esto se puede ver en la demostración del teorema siguiente.

Teorema 41 *Considérese $\{N_t/t \geq 0\}$, un proceso de Poisson de parámetro λ, para el cual se sabe que en el intervalo $[a, b]$ (con $b > a$) se ha dado exactamente una ocurrencia. Entonces, la variable aleatoria Y, definida como el instante concreto del intervalo $[a, b]$ en el que se ha dado esa ocurrencia, sigue una distribución $U[a, b]$.*

Demostración: Sea $t \in [a, b]$. Calculemos $G(t)$, la función de distribución de la variable Y en este punto t:

$$\begin{aligned}
G(t) &= P(Y \leq t) = P\left(N_t - N_a = 1|_{N_b - N_a = 1}\right) = P\left(N_{t-a} = 1|_{N_{b-a}=1}\right) \\
&= \frac{P(N_{t-a} = 1, N_{b-a} = 1)}{P(N_{b-a} = 1)} = \frac{P(N_{t-a} = 1, N_{b-a} - N_{t-a} = 0)}{P(N_{b-a} = 1)} \\
&= \frac{P(N_{t-a} = 1) \cdot P(N_{b-a} - N_{t-a} = 0)}{P(N_{b-a} = 1)} \\
&= \frac{P(N_{t-a} = 1) \cdot P\left(N_{b-a-(t-a)} = 0\right)}{P(N_{b-a} = 1)} \\
&= \frac{\frac{e^{-\lambda(t-a)}(\lambda(t-a))^1}{1!} \cdot \frac{e^{-\lambda(b-t)}(\lambda(b-t))^0}{0!}}{\frac{e^{-\lambda(b-a)}(\lambda(b-a))^1}{1!}} = \frac{e^{-\lambda(t-a+b-t)}\lambda(t-a)}{e^{-\lambda(b-a)}\lambda(b-a)} = \frac{t-a}{b-a},
\end{aligned}$$

que es la fórmula de la función de distribución de una $U[a, b]$.

9.4 Procesos de nacimiento y muerte

El proceso de Poisson (y, en general, los procesos de contar) es útil para modelizar situaciones en las que el objetivo es contabilizar el número de ocurrencias de cierto fenómeno (o nacimientos en una población) hasta un instante t. Existen otros procesos más generales, llamados de nacimiento y muerte, que contemplan la posibilidad de que dicho número pueda disminuir (por ejemplo, si se pretende contabilizar el número de individuos en una población, cada vez que se produce una muerte). Los procesos de nacimiento y muerte, además de generalizar el proceso de Poisson en el sentido recién comentado, también son más generales por el hecho de permitir que las tasas de nacimientos y muertes puedan depender del número de individuos de la población (cosa que efectivamente ocurre en las poblaciones de seres vivos).

Definición 42 *Considérese un proceso estocástico* $\{N_t/t \geq 0\}$ *con espacio de estados discreto:* $E = \{0, 1, 2, \dots\}$ *y supóngase que el proceso describe un sistema que diremos que se encuentra en 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* $\{\lambda_n/n = 0, 1, \dots\}$ *y* $\{\mu_n/n = 1, 2, \dots\}$ *(llamadas tasas de nacimiento y de 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* $n \geq 1$.

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

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

Como ya se comentó antes, es digno de mención que el proceso de Poisson es un caso particular de la definición anterior sin más que considerar que las tasas de muerte son todas cero y que las tasas de nacimiento son constantemente iguales a λ.

A continuación, trataremos de encontrar algún modo de determinar la distribución de probabilidad de un proceso de nacimiento y muerte (o, al menos, sus distribuciones marginales). Denotemos por $P_n(t)$ la probabilidad de que el sistema se encuentre en el estado E_n en el instante t, matemáticamente, $P_n(t) = P(N_t = n)$. Gracias a la regla de las probabilidades totales, podemos expresar fácilmente la probabilidad de que el sistema esté en el estado E_n en el instante $t+h$, en términos de probabilidades de que el sistema esté en distintos

posibles estados en el instante t. Así, si $n \geq 1$, se tiene

$$\begin{aligned} P_n(t+h) &= P(N_{t+h}=n) = P\left(N_{t+h}=n|_{N_t=n-1}\right) P(N_t=n-1) \\ &\quad +P\left(N_{t+h}=n|_{N_t=n}\right) P(N_t=n) \\ &\quad +P\left(N_{t+h}=n|_{N_t=n+1}\right) P(N_t=n+1) \\ &\quad +\sum_{m=0,m\notin\{n-1,n,n+1\}}^{\infty} P\left(N_{t+h}=n|_{N_t=m}\right) P(N_t=m) \\ &= P_{n-1}(t)(\lambda_{n-1}h+o(h)) \\ &\quad +P_n(t)(1-\lambda_n h+o(h))(1-\mu_n h+o(h)) \\ &\quad +P_{n+1}(t)\left(\mu_{n+1}h+o(h)\right)+\sum_{m=0,m\notin\{n-1,n,n+1\}}^{\infty} P_m(t)\,o(h) \\ &= P_{n-1}(t)\lambda_{n-1}h+P_n(t)(1-\lambda_n h-\mu_n h)+P_{n+1}(t)\mu_{n+1}h+o(h). \end{aligned}$$

Como consecuencia se tiene,

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

y, así,

$$\begin{aligned} P_n'(t) &= \frac{dP_n(t)}{dt} = \lim_{h\to 0}\frac{P_n(t+h)-P_n(t)}{h} \\ &= \lambda_{n-1}P_{n-1}(t)-(\lambda_n+\mu_n)P_n(t)+\mu_{n+1}P_{n+1}(t)+\lim_{h\to 0}\frac{o(h)}{h} \\ &= \lambda_{n-1}P_{n-1}(t)-(\lambda_n+\mu_n)P_n(t)+\mu_{n+1}P_{n+1}(t). \end{aligned}$$

De forma totalmente análoga puede tratarse el caso $n=0$, obteniendo las llamadas ecuaciones diferenciales de balance:

$$\begin{aligned} P_n'(t) &= \lambda_{n-1}P_{n-1}(t)-(\lambda_n+\mu_n)P_n(t)+\mu_{n+1}P_{n+1}(t) \text{ si } n\geq 1 \text{ y} \\ P_0'(t) &= -\lambda_0 P_0(t)+\mu_1 P_1(t). \end{aligned}$$

Suponiendo que en el origen de tiempos el sistema se encuentra en estado E_0 (es decir, en $t=0$, no hay individuos en la población), se tienen las condiciones iniciales $P_0(0)=1$ y $P_n(0)=0$ para todo $n\geq 1$.

En general, las ecuaciones de balance (que no es más que un sistema de infinitas ecuaciones diferenciales lineales) son difíciles de resolver. De todas formas hay algunos casos particulares en los que la resolución es más sencilla. Así, si $\mu_n=0$ para todo $n=1,2,\ldots$ y $\lambda_n=\lambda$ para todo $n=0,1,\ldots$ (es decir, para el proceso de Poisson), las ecuaciones de balance resultan especialmente sencillas:

$$\begin{aligned} P_n'(t) &= \lambda P_{n-1}(t)-\lambda P_n(t) \text{ si } n\geq 1 \text{ y} \\ P_0'(t) &= -\lambda P_0(t) \end{aligned}$$

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 $P_0'(t) = -\lambda P_0(t)$ es muy fácil de resolver, siendo su solución general de la forma $P_0(t) = Ce^{-\lambda t}$ (determinando la constante C a partir de las condiciones iniciales, en este caso $P_0(0) = 1$) 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 diferenciales de balance se convierte en un sistema de infinitas ecuaciones lineales:

$$\begin{aligned} 0 &= \lambda_{n-1}p_{n-1} - (\lambda_n + \mu_n)p_n + \mu_{n+1}p_{n+1} \text{ si } n \geq 1 \text{ y} \\ 0 &= -\lambda_0 p_0 + \mu_1 p_1, \end{aligned}$$

que también pueden expresarse como

$$\begin{aligned} (\lambda_n + \mu_n)p_n &= \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1} \text{ si } n \geq 1 \text{ y} \\ \lambda_0 p_0 &= \mu_1 p_1. \end{aligned}$$

Una forma intuitiva de interpretar estas ecuaciones (mediante la cual pueden entenderse como un balance, según su nombre indica) 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 formas de salir de este estado hacia otro distinto. Los términos de la derecha de cada ecuación de balance expresan la suma de la probabilidades de aquellos estados desde los cuales se puede llegar al estado n en un única transición, multiplicadas por las tasas correspondientes a dicha transición.

Así, si $n \geq 1$, en el término de la izquierda se multiplica 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 estado n (pasando, por tanto, a $n+1$) y μ_n, correspondiente a que se produzca una muerte cuando hay n individuos en la población (pasando, consiguientemente, a una población con $n-1$ individuos). Cuando $n = 0$ el razonamiento anterior es válido salvo en lo tocante a μ_0, que no aparece pues no puede haber muertes si ya hay 0 individuos en el sistema. Para los términos de la derecha, cuando $n \geq 1$, se puede llegar al 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 en el que 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 $n = 1$. Supongámosla cierta para n y probémosla para $n+1$. Utilizando la ecuación de balance n-ésima, se tiene

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

que, gracias a la hipótesis de inducción, puede escribirse como

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

o, lo que es lo mismo,

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

lo cual, simplificando los términos μ_n, y despejando p_{n+1} da lugar a

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

que es la expresión que se quería demostrar.

De esta forma se tiene

$$\begin{aligned} p_1 &= \frac{\lambda_0}{\mu_1} p_0, \\ p_2 &= \frac{\lambda_1}{\mu_2} p_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0, \\ &\text{y, en general} \\ p_n &= \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} p_0, \text{ para todo } n = 1, 2, \ldots \end{aligned}$$

Dicho de otra forma, definiendo

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

se tiene que $p_n = c_n \cdot p_0$ para todo $n = 1, 2, \ldots$ Ahora bien, como las p_n son una masa de probabilidad, se verifica

$$\begin{aligned} 1 &= \sum_{n=0}^{\infty} p_n = \left(1 + \sum_{n=1}^{\infty} c_n\right) p_0, \\ \text{con lo cual } p_0 &= \frac{1}{1 + \sum_{n=1}^{\infty} c_n}, \end{aligned}$$

siempre que $\sum_{n=1}^{\infty} c_n < \infty$. Esta última es la llamada condición de estado estacionario, que viene a querer decir que para que exista un proceso de nacimiento y muerte estacionario con tasas de nacimiento $\{\lambda_n\}_{n\geq 0}$ y tasas de muerte $\{\mu_n\}_{n\geq 1}$ ha de cumplirse que $\sum_{n=1}^{\infty} c_n < \infty$. Además puede demostrarse que esta condición es también suficiente (no sólo necesaria) para que el proceso sea estacionario.

10

Conceptos generales de la teoría de colas

Hasta el momento hemos estudiado diversas herramientas de procesos estocásticos que serán utiles para el estudio de colas con tiempos entre llegadas y tiempos de servicio poissonianos. En este capítulo daremos una introducción al concepto de cola, propiamente dicho, estableciendo la terminología básica que utilizaremos en el siguiente capítulo e indicando algunas propiedades importantes, tanto de las distribuciones exponencial y gamma (de mucha importancia en los modelos de colas antedichos) como para diversos conceptos asociados con algunas variables de interés en una cola (las llamadas fórmulas de Little).

10.1 Descripción del sistema de una cola

En muchas ocasiones en la vida real, un fenómeno muy común es la formación de colas o líneas de espera. 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: los cruces de dos vías de circulación, los semáforos, el peaje de una autopista, los cajeros automáticos, la atención a clientes en un establecimiento comercial, la avería de electrodomésticos u otro tipo de aparatos que deben ser reparados por un servicio técnico, etc.

Todavía más frecuentes, si cabe, son las situaciones de espera en el contexto de la informática, las telecomunicaciones y, en general, las nuevas tecnologías. Así, por ejemplo, los procesos enviados a un servidor para ejecución forman colas de espera mientras no son atendidos, la información solicitada, a través de internet, a un servidor web puede recibirse con demora debido a congestión en la red o en el servidor propiamente dicho, podemos recibir la señal de líneas ocupadas si la central de la que depende nuestro teléfono móvil está colapsada en ese momento, etc.

Fenómenos como los que se acaban de relatar (y muchísimos otros) tienen ciertas características comunes que dan lugar al modelo de sistema de una cola. En dicho modelo distinguiremos los siguientes elementos.

Fuente de entrada o población potencial: Es un conjunto de individuos (no necesariamente seres vivos) que pueden llegar a solicitar el servicio en cuestión. 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.

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.

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.

Disciplina de la cola: Es el modo en el que los clientes son seleccionados para ser servidos. Las disciplinas más habituales son:

1. La disciplina FIFO (first in first out), también llamada FCFS (first come first served): según la cual se atiende primero al cliente que antes haya llegado.

2. La disciplina LIFO (last in first out), también conocida como LCFS (last come first served) o pila: que consiste en atender primero al cliente que ha llegado el último.

3. La RSS (random selection of service), o SIRO (service in random order), que selecciona a los clientes de forma aleatoria.

4. La disciplina RR (round robin), según la cual se otorga un pequeño cuanto de tiempo de servicio a cada cliente de forma secuencial. Esto viene a equivaler a repartir los recursos de forma igualitaria entre todos los clientes en espera y, por supuesto sólo tiene sentido en algunas circunstancias (como el ámbito de la informática).

Mecanismo de servicio: Es el procedimiento por el cual se da servicio a los clientes que lo solicitan. 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 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.
La cola, propiamente dicha, es el conjunto de clientes que hacen espera, es decir los clientes que ya han solicitado el servicio pero que aún no han pasado al mecanismo de servicio.
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.

10.2 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.
$\mathbf{N}(t)$**:** 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.
$\mathbf{N}_q(t)$**:** Representa el número de clientes en la cola en el instante t.
$\mathbf{P}_n(t)$**:** Es la probabilidad de que, en el instante t, se encuentren n clientes en el sistema. A estos efectos se supone conocido el número de clientes en el instante cero (usualmente dicho número es cero).
s: Denota el número de servidores del mecanismo de servicio.
$\boldsymbol{\lambda_n}$**:** 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 λ dicho valor constante.
$\boldsymbol{\mu_n}$**:** 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 $n = 1, 2, \ldots, s$ y $\mu_n = s\mu$ para $n \geq s$.
$\boldsymbol{\rho}$**:** Es la llamada constante de utilización del sistema o intensidad de tráfico. Se define, como

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

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 $s\mu$ es el número medio de clientes a los que pueden dar servicio los s

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 que 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.
$\mathbf{N}_q$: Denota la variable aleatoria número de clientes en la cola.
$\mathbf{p}_n$: Es la probabilidad de que se encuentren n clientes en el sistema ($n = 0, 1, \dots$).
L: Representa el número medio de clientes en el sistema, es decir $L = E(N)$.
$\mathbf{L}_q$: Que no es más que el número medio de clientes en la cola, o lo que es lo mismo, $L_q = E(N_q)$.
$\mathcal{W}$: Es la variable aleatoria que describe el tiempo que un cliente pasa en el sistema.
$\mathcal{W}_q$: Representa el tiempo que un cliente espera en la cola.
W: Es el tiempo medio que un cliente está en el sistema, o simplemente, $W = E(\mathcal{W})$.
$\mathbf{W}_q$: Denota el tiempo medio de espera en la cola para un cliente genérico. Matemáticamente, $W_q = E(\mathcal{W}_q)$.

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 $A/B/s$ para indicar que la distribución del tiempo entre llegadas es de del tipo A, que B es la distribución del tiempo de servicio y que s 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 $A/B/s/K/H/Z$, donde:
A es la distribución del tiempo entre llegadas. Algunas de las abreviaturas más usadas para las distribuciones entre llegadas son: M (exponencial), D (determinística), E_k (Erlang con segundo parámetro k), U (uniforme), Γ (gamma) o G (distribución genérica), entre otras.
B es la distribución del tiempo de servicio. Se usan las mismas abreviaturas que las mencionadas para A.
s es el número de servidores del sistema. Puede ser un número entero positivo ($s = 1, 2, \dots$) o bien $s = \infty$.
K es la capacidad de la cola (o longitud máxima de la misma). También K puede ser un número entero mayor o igual que cero, o bien $K = \infty$, si no hay límite para la cola. El valor de K puede omitirse, tomándose por defecto $K = \infty$.

H es el tamaño de la población potencial. También puede ser finito o infinito. Este último valor es el que se toma por defecto cuando se omite el H.
Z es la disciplina en la cola. Algunas abreviaturas para Z son FIFO, LIFO, RSS, PR (disciplina con prioridades) o GD (disciplina general). Su valor por defecto (en caso de omitirse Z) es FIFO.

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 atentidos según una disciplina FIFO. Como los tres últimos valores (∞, ∞ y 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 $A = E_2$, $B = U$, $s = 3$ y $Z = FIFO$, nunca sabríamos si con él pretendemos indicar $K = 4$ y $H = \infty$ o bien $H = 4$ y $K = \infty$.

10.3 Fórmulas de Little

En los modelos con distribución del tiempo 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 clientes 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 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 las λ_n no son constantes (¿qué sería λ entonces?), pero sí pueden generalizarse a esa situación mediante:

$$\begin{aligned} L &= \overline{\lambda} \cdot W \text{ y} \\ L_q &= \overline{\lambda} \cdot W_q, \\ \text{siendo } \overline{\lambda} &= \sum_{n=0}^{\infty} \lambda_n p_n. \end{aligned}$$

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

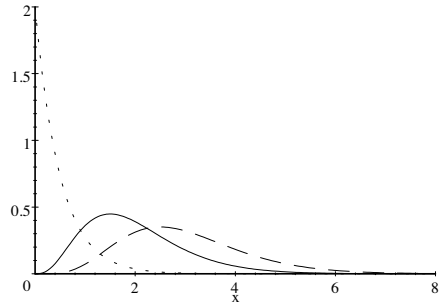
$$W = W_q + \frac{1}{\mu}.$$

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 ($1/\mu$, ya que μ es el número medio de clientes que un servidor puede atender por unidad de tiempo).

10.4 Algunas propiedades importantes de las distribuciones exponencial y gamma

La distribución $\Gamma(a, p)$ depende de dos parámetros: $a > 0$, parámetro de escala, y $p > 0$, parámetro de forma. Esta distribución, ya introducida con detalle en el capítulo 5, generaliza la exponencial, puesto que $\Gamma(a, 1) \overset{\mathrm{d}}{=} \exp(a)$.

En la siguiente gráfica se muestran las densidades de una $\Gamma(2, 4)$, una $\Gamma(2, 6)$ y una $\Gamma(2, 1) = \exp(2)$.



Densidades de una $\Gamma(2,4)$ (trazo continuo), una $\Gamma(2,6)$ (trazo discontinuo) y una $\Gamma(2,1)$ (trazo a puntos)

En general, la media y la varianza de una variable $X \overset{\mathrm{d}}{=} \Gamma(a,p)$ son $E(X) = \frac{p}{a}$ y $Var(X) = \frac{p}{a^2}$. A continuación veremos algunas propiedades interesantes de estas distribuciones en relación con los modelos de colas de tipo $M/M/\cdots$.

Proposición 43 *Si* $T \overset{d}{=} \exp(\alpha)$, *entonces*

$$P(0 \leq T \leq \Delta t) > P(t \leq T \leq t + \Delta t),$$

para todo $t > 0$.

Esta propiedad puede deducirse fácilmente del hecho de que la función de densidad de la exponencial sea estrictamente decreciente en $[0, \infty)$.

Proposición 44 (carencia de memoria de la exponencial) *Si* $T \overset{d}{=} \exp(\alpha)$, *entonces*

$$P\left(T > t + \Delta t|_{T>\Delta t}\right) = P(T > t), \text{ para todo } t > 0.$$

Esto implica que, si T *tiene distribución exponencial y* $\Delta t > 0$, *entonces*

$$T - \Delta t|_{T>\Delta t} \overset{d}{=} T.$$

Demostración: Gracias a la definición de probabilidad condicionada, para cada $t > 0$, se tiene

$$\begin{aligned} P\left(T > t + \Delta t|_{T>\Delta t}\right) &= \frac{P(T > t + \Delta t, T > \Delta t)}{P(T > \Delta t)} = \frac{P(T > t + \Delta t)}{P(T > \Delta t)} \\ &= \frac{e^{-\alpha(t+\Delta t)}}{e^{-\alpha \Delta t}} = e^{-\alpha t} = P(T > t), \end{aligned}$$

con lo que queda demostrada la primera afirmación. Otra manera alternativa de escribir la igualdad probada es

$$P\left(T - \Delta t > t|_{T>\Delta t}\right) = P(T > t),$$

lo cual implica inmediatamente que $T - \Delta t|_{T>\Delta t} \stackrel{\mathrm{d}}{=} T$.

La propiedad de carencia de memoria de la exponencial (también conocida como la propiedad de Markov) es muy útil en modelos de colas con distribución exponencial para el tiempo entre llegadas o para el tiempo de servicio. Así, si sabemos que el tiempo entre la llegada de dos clientes consecutivos al sistema es exponencial, aunque ya haya pasado cierto tiempo desde que llegó al sistema el primero de ellos, el tiempo que resta para que llegue el siguiente sigue teniendo distribución exponencial con el mismo parámetro. Un comentario semejante sería válido para un tiempo de servicio (si tiene distribución exponencial).

Proposición 45 *Si consideramos variables aleatorias* $T_1 \stackrel{d}{=} \exp(\alpha_1)$, $T_2 \stackrel{d}{=} \exp(\alpha_2)$, $\ldots$, $T_n \stackrel{d}{=} \exp(\alpha_n)$, *independientes, entonces, la variable aleatoria definida como* $T = \min\{T_1, T_2, \ldots, T_n\}$ *tiene distribución* $\exp(\sum_{i=1}^n \alpha_i)$.

Demostración: La función de distribución, F, de la variable T viene dada por

$$\begin{aligned} F(t) &= 1 - P(T > t) = 1 - P(\min\{T_1, T_2, \ldots, T_n\} > t) = \\ &= 1 - P(T_1 > t, T_2 > t, \ldots, T_n > t) = 1 - \prod_{i=1}^n P(T_i > t) \\ &= 1 - \prod_{i=1}^n e^{-\alpha_i t} = 1 - e^{-\left(\sum_{i=1}^n \alpha_i\right)t}, \end{aligned}$$

que, efectivamente, es la función de distribución de una $\exp(\sum_{i=i}^n \alpha_i)$.

Esta propiedad recién demostrada también se utiliza con mucha frecuencia en modelos de colas del tipo $M/M/\cdots$. Así, si en un preciso momento hay s servidores dando servicio a otros tantos clientes y la distribución del tiempo de servicio es $\exp(\mu)$, en virtud de la carencia de memoria de la exponencial, los tiempos de servicio que restan para cada cliente son igualmente de distribución $\exp(\mu)$ y, debido al resultado relativo al mínimo de exponenciales, el tiempo que resta hasta que el próximo cliente salga servido (es decir, el de menor tiempo de servicio) sigue una distribución $\exp(s\mu)$.

Proposición 46 *La distribución* $\Gamma(a, p)$ *es reproductiva con respecto a su segundo parámetro. En otras palabras, dadas variables independientes* $X_1 \stackrel{d}{=} \Gamma(a, p_1)$, $X_2 \stackrel{d}{=} \Gamma(a, p_2)$, $\ldots$, $X_n \stackrel{d}{=} \Gamma(a, p_n)$, *la variable aleatoria* $X = \sum_{i=1}^n X_i$ *tiene distribución* $\Gamma(a, \sum_{i=1}^n p_i)$.

La utilidad de este resultado en el contexto de las colas con llegadas y servicios exponenciales tiene que ver con el hecho de que la propiedad anterior, en particular, permite deducir la distribución de la suma de variables exponenciales independientes con el mismo parámetro. Particularizando el resultado anterior para $p_1 = p_2 = \cdots = p_n = 1$ y $a = \alpha$, se tiene que si X_1, X_2, $\ldots$,

X_n son variables aleatorias independientes y con distribución idéntica, $\exp(\alpha)$, entonces $X = \sum_{i=1}^{n} X_i$ tiene distribución $\Gamma(\alpha, n)$. Las variables X_i podrían representar los tiempos desde que llega al sistema el cliente i-ésimo hasta que llega el $(i+1)$-ésimo. En ese caso X representaría el tiempo que transcurre entre la llegada de un cliente y la del que llega n clientes después que él.

Otra aplicación interesante de este resultado en teoría de colas es la siguiente. Supóngase que en un modelo $M/M/s/\cdots$ un cliente llega al sistema cuando ya hay n clientes dentro del sistema $(n > s)$. Si deseamos saber la distribución de la variable tiempo que el cliente pasará en la cola, no hay más que razonar de la siguiente forma. El tiempo que tardará en salir servido el próximo cliente tiene distribución $\exp(s\mu)$, según lo razonado en uno de los comentarios hechos anteriormente. Ahora bien, para que el cliente en cuestión comience a ser servido han de salir servidos $n-s+1$ clientes del sistema, siendo las duraciones de todos esos tiempos entre dos salidas de clientes del sistema variables independientes y con distribución $\exp(s\mu)$. Así la última propiedad nos permite afirmar que la variable de nuestro interés tiene distribución $\Gamma(s\mu, n-s+1)$.

10.5 Ejercicios propuestos

1. Utilizando las fórmulas de Little calcular, para un modelo $M/M/s$, el número medio de servidores ocupados.

2. En un modelo $M/M/2$ con $\mu = 2$ clientes por minuto, llega un cliente en un determinado momento en el que hay otros dos haciendo cola para entrar al mecanismo de servicio. ¿Cuál es el tiempo medio que transcurrirá hasta que el cliente recién llegado salga servido del sistema?

11

Modelos con tasas de llegada y servicio de tipo Poisson

En este capítulo estudiaremos los modelos más habituales para el sistema de una cola en la que se supone que la distribución del tiempo entre llegadas y la del tiempo de servicio son exponenciales. Concretamente se hará un recorrido por los modelos $M/M/1$, $M/M/s$, $M/M/1/K$, $M/M/s/K$, $M/M/1/\infty/H$, $M/M/s/\infty/H$ (sin y con repuestos) y el $M/M/\infty$.

11.1 El modelo $M/M/1$

Como su nombre indica es aquel en el que la distribución del tiempo entre dos llegadas consecutivas de clientes al sistema es una $\exp(\lambda)$, independientemente del número de clientes que haya dentro del mismo, la distribución del tiempo de servicio es $\exp(\mu)$ y sólo hay un servidor. Los valores por defecto de los tres últimos parámetros indican que no hay restricción respecto al número de clientes que pueden estar en la cola, que la población potencial es infinita (de ahí que la distribución del tiempo entre llegadas sea siempre la misma exponencial) y que la disciplina de la cola es FIFO.

Teniendo en cuenta todo esto, resulta inmediato deducir las tasas de llegada

$$\lambda_n = \lambda, \text{ para todo } n = 0, 1, \ldots$$

y las tasas de servicio

$$\mu_n = \mu, \text{ para todo } n = 1, 2, \ldots$$

Como consecuencia se tiene que

$$c_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1} = \frac{\lambda^n}{\mu^n} = \left(\frac{\lambda}{\mu}\right)^n = \rho^n, \text{ para } n = 1, 2, \ldots$$

Como $\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} \rho^n$ es una serie geométrica, será convergente si y sólo sí $|\rho| < 1$, que dado que $\rho > 0$, equivale a $\rho < 1$. Esta condición, $\rho < 1$, 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 $\rho < 1$). Lo primero que debemos calcular es la suma de la serie de las c_n:

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} \rho^n = \frac{\rho}{1-\rho}$$

y, por lo tanto,

$$p_0 = \frac{1}{1+\sum_{n=1}^{\infty} c_n} = \frac{1}{1+\frac{\rho}{1-\rho}} = 1-\rho.$$

Además, para cualquier $n \geq 1$, se tiene $p_n = c_n \cdot p_0 = (1-\rho)\,\rho^n$ (valiendo esta fórmula final incluso para $n = 0$). Así pues, la masa de probabilidad de la variable "número de clientes en el sistema" es

$$P(N = n) = p_n = (1-\rho)\,\rho^n,\ n = 0, 1, \ldots$$

y así, N tiene distribución geométrica de parámetro $1-\rho$ (probabilidad de éxito, entendiendo la geométrica como el número de fracasos antes del primer éxito). Suponiendo conocida la fórmula para la media de una distribución geométrica, se calcularía de forma inmediata el valor de $L = E(N)$. También podemos hacer el cálculo directamente:

$$L = E(N) = \sum_{n=0}^{\infty} n \cdot p_n = \sum_{n=1}^{\infty} n \cdot (1-\rho)\,\rho^n = (1-\rho)\,\rho \sum_{n=1}^{\infty} n \cdot \rho^{n-1}.$$

Esta última serie es del tipo que se suele denominar convertible en geométrica. Una forma de calcular su suma es la siguiente. Considérese la función $f(x) = \sum_{n=1}^{\infty} x^n$, definida para los $x \in (-1, 1)$. Dado que para cada $x \in (-1, 1)$, $f(x)$ está definida a través de una serie geométrica convergente, se tiene que $f(x) = \frac{x}{1-x}$. Además, la función $f(x)$ puede derivarse, tanto directamente a partir de esta expresión encontrada, como término a término en su definición como serie de potencias. Así, se obtiene:

$$\sum_{n=1}^{\infty} n x^{n-1} = f'(x) = \frac{1\,(1-x) - x\,(-1)}{(1-x)^2} = \frac{1}{(1-x)^2}.$$

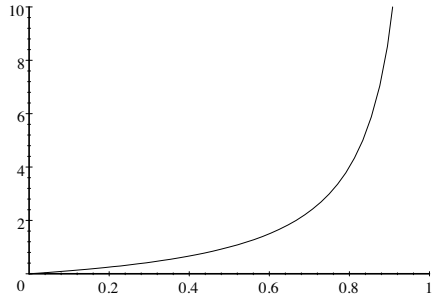
De este modo, en particular,

$$\sum_{n=1}^{\infty} n \cdot \rho^{n-1} = \frac{1}{(1-\rho)^2}$$

y, por tanto,

$$L = (1-\rho)\,\rho \frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.$$

Se trata, pues, de una función de ρ con una asíntota vertical en $\rho \to 1^-$.



L como función de ρ

Aunque L_q puede calcularse de forma similar, también puede obtenerse a partir de L. Veámoslo,

$$\begin{aligned} L_q &= E(N_q) = \sum_{n=0}^{\infty} n \cdot P(N_q = n) = \sum_{n=1}^{\infty} n \cdot P(N_q = n) \\ &= \sum_{n=1}^{\infty} n \cdot P(N = n+1) = \sum_{n=1}^{\infty} n \cdot p_{n+1} = \sum_{n=0}^{\infty} n \cdot p_{n+1} \\ &= \sum_{m=1}^{\infty} (m-1) \cdot p_m = \sum_{m=1}^{\infty} m \cdot p_m - \sum_{m=1}^{\infty} p_m \\ &= \sum_{m=0}^{\infty} m \cdot p_m - (1-p_0) = L - (1-p_0), \end{aligned}$$

fórmula que es válida para cualquier modelo de la forma $G/G/1/\cdots$. En nuestro caso, de esta fórmula se deduce

$$\begin{aligned} L_q &= L - (1-p_0) = \frac{\rho}{1-\rho} - (1-(1-\rho)) = \frac{\rho}{1-\rho} - \rho \\ &= \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}. \end{aligned}$$

Aplicando las fórmulas de Little pueden obtenerse expresiones para W y W_q:

$$\begin{aligned} W &= \frac{L}{\lambda} = \frac{1}{\mu-\lambda}, \\ W_q &= \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}. \end{aligned}$$

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

$\mathcal{W}$ y $\mathcal{W}_q$. 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 $\mathcal{W}$, que denotaremos por $W(t)$. 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 $\mathcal{W}|_{N=n} \overset{\mathrm{d}}{=} \Gamma(\mu, n+1)$. De esta forma, para cada $t \geq 0$, se tiene

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

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

$$\mathcal{W} \overset{\mathrm{d}}{=} \exp(\mu - \lambda).$$

Es obvio, por tanto, que como conclusión de esto también se puede volver a obtener $W = \frac{1}{\mu - \lambda}$.

La función de distribución del tiempo que un cliente está en la cola puede obtenerse de forma análoga. Basta descomponer dicha función en dos términos de la forma:

$$W_q(t) = P(\mathcal{W}_q \leq t) = P(\mathcal{W}_q = 0) + P(0 < \mathcal{W}_q \leq t)$$

y aplicar al segundo de ellos los mismos argumentos usados para el cálculo de $W(t)$. De esa forma se obtendría:

$$W_q(t) = \begin{cases} 1 - \dfrac{\lambda}{\mu} e^{-(\mu-\lambda)t} & \text{si } t \geq 0 \\ 0 & \text{si } t < 0 \end{cases}$$

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

$$W_q(0) - W_q(0^-) = 1 - \frac{\lambda}{\mu} = 1 - \rho = p_0.$$

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

Ejemplo 47 *En una estación de trabajo con un único procesador se ejecutan programas (que se supone prácticamente su única carga de trabajo) con tiempo de CPU de distribución exponencial de media 3 minutos. Los programas se atienden según una disciplina FIFO. Sabiendo que las llegadas de programas a la estación se producen según un proceso de Poisson con una intensidad de 15 programas cada hora, por término medio, se pide:*

1. *¿Cuál es la probabilidad de que haya más de dos programas en espera de ejecución (además del que se está ejecutando)?*
2. *Calcular el tiempo medio que transcurre desde que se envía un programa al servidor hasta que se termina su ejecución. ¿Cuál es la relación entre este tiempo y el tiempo medio de CPU?*
3. *Calcular la probabilidad de que el programa esté en el servidor (esperando o ejecutándose) más de 10 minutos.*
4. *¿Cuál es el número medio de programas que están a la espera de comenzar a ejecutarse?*
5. *Obtener las respuestas a los apartados anteriores suponiendo que ahora se ha incrementado la llegada de programas hasta 18 a la hora, por término medio.*

Solución: La exponencialidad de los tiempos de servicio y de los tiempos entre llegadas nos indica que se trata de un modelo $M/M/1$. De hecho, estamos suponiendo (al no especificar ningún valor) que tanto el límite para la cola como el número de clientes de la población potencial son infinitos. En virtud de los datos del problema y trabajando en horas se tiene

$$\lambda = 15,$$
$$\frac{1}{\mu} = \frac{3}{60}, \text{ con lo cual } \mu = 20.$$

De este modo $\rho = \frac{15}{20} = 0.75 < 1$ y el modelo es estacionario.

1. *En este apartado se nos pide*

$$\begin{aligned} P(N_q > 2) = P(N_q \geq 3) = P(N \geq 4) &= 1 - P(N \leq 3) \\ &= 1 - (p_0 + p_1 + p_2 + p_3) \\ &= 1 - \left(1 - \rho + (1-\rho)\rho + (1-\rho)\rho^2 + (1-\rho)\rho^3\right) \\ &= 1 - \frac{(1-\rho)\rho^3\rho - (1-\rho)}{\rho - 1} = 1 - (1-\rho)\frac{\rho^4 - 1}{\rho - 1} \\ &= 1 + \rho^4 - 1 = \rho^4 = 0.75^4 = 0.3164 \end{aligned}$$

2. *Aquí se pide*

$$W = \frac{1}{\mu - \lambda} = \frac{1}{20 - 15} = 0.2 \text{ horas} = 12 \text{ minutos.}$$

Su relación con el tiempo medio de servicio, $W_s = \frac{1}{\mu}$*, es*

$$\frac{W}{W_s} = \frac{\frac{1}{5}}{\frac{1}{20}} = 4,$$

es decir cada proceso está en la estación un tiempo equivalente a cuatro veces su tiempo de CPU.

3. *Como* $\mathcal{W}$ *tiene una distribución* $\exp(\mu - \lambda) = \exp(5)$ *entonces*

$$P\left(\mathcal{W} > \frac{10}{60}\right) = e^{-5\frac{10}{60}} = e^{-\frac{5}{6}} = 0.4346.$$

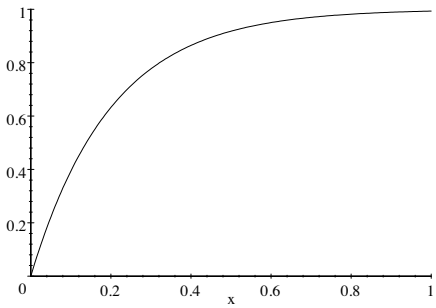
4. *Se pide* L_q*. Como* $W_q = W - \frac{1}{\mu} = \frac{1}{5} - \frac{1}{20} = \frac{3}{20}$ *y* $L_q = \lambda W_q$*, entonces*

$$L_q = 15\frac{3}{20} = \frac{9}{4} = 2.25 \text{ procesos en espera.}$$

5. *La única diferencia es que ahora* $\lambda = 18$*, con lo cual* $\rho = \frac{18}{20} = 0.9 < 1$*,*

$$\begin{aligned} P(N_q > 2) &= 0.9^4 = 0.6561, \\ W &= \frac{1}{20 - 18} = 0.5 \text{ horas} = 30 \text{ minutos,} \\ P\left(\mathcal{W} > \frac{10}{60}\right) &= e^{-2\frac{10}{60}} = e^{-\frac{1}{3}} = 0.7165 \text{ y} \\ W_q &= \frac{1}{2} - \frac{1}{20} = \frac{9}{20} \Rightarrow L_q = 18\frac{9}{20} = 8.1 \text{ procesos.} \end{aligned}$$

Como puede verse claramente, el sistema está bastante más congestionado ahora.



Distribución del tiempo (en horas) de espera en el sistema para el caso $\lambda = 15$

11.2 El modelo $M/M/s$

Es una generalización del modelo anterior al caso en que haya s servidores. Se trata pues de una cola en la que la distribución del tiempo entre llegadas consecutivas es una $\exp(\lambda)$, la distribución del tiempo de servicio es $\exp(\mu)$ y hay s servidores. Como en el caso anterior, la población potencial y la capacidad de la cola son infinitas y la disciplina de la cola es FIFO.

Las tasas de llegada vienen dadas por

$$\lambda_n = \lambda, \text{ para todo } n = 0, 1, \ldots$$

y las tasas de servicio

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

Con lo cual

$$c_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1} = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} & \text{si } n = 1, 2, \ldots, s \\ \dfrac{\lambda^n}{s!s^{n-s}\mu^n} & \text{si } n = s+1, s+2, \ldots \end{cases}$$

Para analizar cuando el sistema es estacionario basta estudiar la convergencia de la serie

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{s-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=s}^{\infty} \frac{\lambda^n}{s!s^{n-s}\mu^n} = \sum_{n=1}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{s!\mu^s} \sum_{n=s}^{\infty} \rho^{n-s},$$

que, a partir del término s-ésimo, es geométrica de razón ρ y, por tanto, convergente siempre que $\rho < 1$. Así pues el sistema es estacionario siempre que $\rho < 1$, o lo que es lo mismo, siempre que $\lambda < s\mu$. En tal caso, la suma de dicha serie es

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

con lo cual p_0 viene dado por

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} c_n} = \frac{1}{1 + \sum_{n=1}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{(s-1)!\mu^{s-1}(s\mu-\lambda)}}.$$

A la hora de implementar el cálculo de p_0 debe tenerse precaución con la fórmula anterior pues, si s es elevado, los términos con $n!$ en el denominador

pueden producir errores de tipo "overflow". Además, la implementación directa de esas expresiones sería muy ineficiente por estar repitiendo muchos cálculos sin reutilizarlos. Así, por ejemplo, si $s = 100$, en el denominador de c_{98} aparecería 98!, que, en lugar de calcularlo directamente podría obtenerse mucho más eficientemente multipicando por 98 el término 97! que aparece en c_{97}. Este mismo argumento también puede aplicarse a las potencias λ^n y μ^n. En resumen, parece más eficiente definir $c_0 = 1$ y utilizar el cálculo recursivo

$$c_n = c_{n-1}\frac{\lambda_{n-1}}{\mu_n} = c_{n-1}\frac{\lambda}{n\mu}, \text{ para } n = 1, 2, \dots, s-1,$$

que se deduce de la definición de los c_n. Además el término $\frac{\lambda^s}{(s-1)!\mu^{s-1}(s\mu-\lambda)}$ que representa $\sum_{n=s}^{\infty} c_n$ y que denotaremos por $c_{\geq s}$, puede también calcularse fácilmente a partir de c_{s-1} mediante $c_{\geq s} = c_{s-1}\frac{\lambda}{s\mu-\lambda}$.

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

$$\begin{aligned} p_0 &= \frac{1}{\sum_{n=0}^{s-1} c_n + c_{\geq s}}, \text{ donde} \\ c_0 &= 1,\ c_n = c_{n-1}\frac{\lambda}{n\mu}, \text{ para } n = 1, 2, \dots, s-1 \\ \text{y } c_{\geq s} &= c_{s-1}\frac{\lambda}{s\mu-\lambda}. \end{aligned}$$

Resulta fácil obtener una expresión explícita para las p_n:

$$p_n = c_n \cdot p_0 = \begin{cases} \frac{\lambda^n}{n!\mu^n}p_0 & \text{si } n = 1, 2, \dots, s \\ \frac{\lambda^s}{s!\mu^s}\rho^{n-s}p_0 & \text{si } n = s+1, s+2, \dots \end{cases}$$

aunque, de nuevo, estas fórmulas implementadas directamente son de muy escasa eficiencia computacional. Como para el cálculo de p_0, también aquí podemos proceder más eficientemente.

Una vez calculado p_0 y si se mantienen en memoria los valores de los c_n, el cálculo de los p_n sería muy sencillo. Así, si $n = 1, 2, \dots, s-1$, se tiene $p_n = c_n \cdot p_0$ (fórmula que, aunque también es válida para $n \geq s$, no es muy útil pues no se han calculado los c_n para $n \geq s$). Si $n \geq s$ los p_n se pueden calcular recursivamente ya que p_{s-1} está calculado y

$$p_n = c_n \cdot p_0 = \frac{\lambda_{n-1}}{\mu_n}c_{n-1} \cdot p_0 = \frac{\lambda_{n-1}}{\mu_n}p_{n-1} = \frac{\lambda}{s\mu}p_{n-1} = \rho \cdot p_{n-1}$$

para todo $n = s, s+1, \dots$ Así, de forma resumida, se procedería a

$$\begin{aligned} \text{calcular directamente } p_n &= c_n \cdot p_0 \text{ para } n = 1, 2, \dots, s-1 \text{ y} \\ \text{recursivamente } p_n &= \rho \cdot p_{n-1} \text{ para } n = s, s+1, \dots \\ &\quad \text{hasta el índice que se desee.} \end{aligned}$$

Gracias a la expresión explícita obtenida para las p_n, puede encontrarse fácilmente el valor de L_q:

$$\begin{aligned} L_q &= E(N_q) = 0 \cdot (p_0 + p_1 + \cdots + p_s) + \sum_{n=s+1}^{\infty} (n-s) \cdot p_n \\ &= \sum_{n=s}^{\infty} (n-s) \frac{\lambda^n}{s! s^{n-s} \mu^n} p_0 = \frac{\lambda^s}{s! \mu^s} p_0 \sum_{n=s}^{\infty} (n-s) \rho^{n-s} = \frac{\lambda^s}{s! \mu^s} p_0 \sum_{k=0}^{\infty} k \rho^k \\ &= \frac{\lambda^s p_0}{s! \mu^s} \cdot \frac{\rho}{(1-\rho)^2} = \frac{\lambda^{s+1} p_0}{(s-1)! \mu^{s-1} (s\mu - \lambda)^2}. \end{aligned}$$

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

$$L_q = c_{\geq s} \frac{\lambda p_0}{s\mu - \lambda}.$$

A partir de L_q puede obtenerse el valor de $W_q = \frac{L_q}{\lambda}$, mediante la segunda fórmula de Little. Ahora, se obtendría $W = W_q + \frac{1}{\mu}$ y luego $L = \lambda W$. De todas formas, es posible obtener expresiones explícitas de estas cantidades en términos de λ, μ, s y p_0:

$$\begin{aligned} W_q &= \frac{\lambda^s p_0}{(s-1)! \mu^{s-1} (s\mu - \lambda)^2}, \\ W &= \frac{\lambda^s p_0}{(s-1)! \mu^{s-1} (s\mu - \lambda)^2} + \frac{1}{\mu}, \\ L &= \frac{\lambda^{s+1} p_0}{(s-1)! \mu^{s-1} (s\mu - \lambda)^2} + \frac{\lambda}{\mu}, \end{aligned}$$

que no es recomendable usar en la práctica por su mayor lentitud de cálculo con respecto a las propuestas anteriores.

De forma análoga al caso del modelo $M/M/1$, se puede llegar a expresiones para las funciones de distribución del tiempo que un cliente está en la cola y

del tiempo que un cliente está en el sistema. Son las siguientes:

$$W_q(t) = \begin{cases} 1 - \dfrac{\lambda^s p_0}{(s-1)!\mu^{s-1}(s\mu-\lambda)} e^{-(s\mu-\lambda)t} & \text{si } t \geq 0 \\ 0 & \text{si } t < 0 \end{cases}$$

En el caso $\dfrac{\lambda}{\mu} \neq s-1$ se tiene

$$W(t) = 1 + \frac{\lambda - s\mu + \mu W_q(0)}{s\mu - \lambda - \mu} e^{-\mu t} + \frac{\lambda^s p_0}{(s-1)!\mu^{s-2}(s\mu-\lambda)(s\mu-\lambda-\mu)} e^{-(s\mu-\lambda)t},$$

si $t \geq 0$ (0 en otro caso)

mientras que si $\dfrac{\lambda}{\mu} = s-1$ su expresión es

$$W(t) = \begin{cases} 1 - \left(1 + \dfrac{\lambda^s p_0 t}{(s-1)!\mu^{s-2}(s\mu-\lambda)}\right) e^{-\mu t} & \text{si } t \geq 0 \\ 0 & \text{si } t < 0 \end{cases}$$

Las constantes que aparecen en las expresiones anteriores son eficientemente calculables a partir de lo comentado anteriormente. Así, por ejemplo,

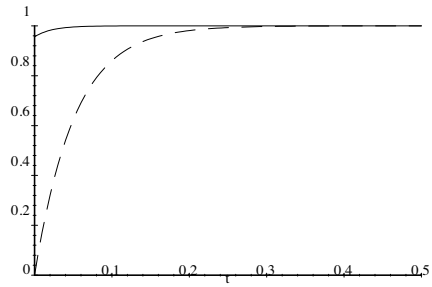
$$W_q(t) = 1 - c_{\geq s} \cdot p_0 \cdot e^{-(s\mu-\lambda)t} \text{ si } t \geq 0.$$

Gracias a esta última expresión resulta muy sencillo calcular la probabilidad de que un cliente no tenga que esperar en la cola (es decir, que pase directamente a ser servido):

$$P(\mathcal{W}_q = 0) = P(\mathcal{W}_q \leq 0) = W_q(0) = 1 - c_{\geq s} \cdot p_0 = 1 - \sum_{n=s}^{\infty} c_n \cdot p_0$$

$$= 1 - \sum_{n=s}^{\infty} p_n = \sum_{n=0}^{s-1} p_n = P(N \leq s-1),$$

que, por otra parte, resulta totalmente intuitivo, pues corresponde con el hecho de que cuando llegue el cliente en cuestión haya algún servidor libre. Como puede observarse en la siguiente figura, es obvio que la gráfica de la función $W_q(t)$ está sobre la de $W(t)$.



Funciones de distribución del tiempo en la cola (trazo continuo) y del tiempo en el sistema (trazo discontinuo)

Ejemplo 48 *Calcular las cantidades pedidas en los apartados 1 y 4 del último ejemplo suponiendo los mismos datos pero considerando una estación de trabajo con tres servidores idénticos al de dicho ejemplo. Hallar también el número medio total de procesos en la estación.*

Solución: Obviamente se trata de un modelo $M/M/s$ *con* $s = 3$, $\lambda = 15$ *y* $\mu = 20$, *con lo cual,* $\rho = \frac{15}{3 \cdot 20} = \frac{1}{4} < 1$, *siendo, por tanto, el modelo estacionario. Las dos primeras cantidades pedidas siguen siendo, obviamente,* $P(N_q > 2)$ *y* L_q. *En primer lugar debemos hallar* p_0*:*

$$\text{Como } c_0 = 1,\ c_1 = c_0\frac{15}{20} = \frac{3}{4},\ c_2 = c_1\frac{15}{2 \cdot 20} = \frac{9}{32}$$
$$\text{y } c_{\geq 3} = c_2\frac{15}{3 \cdot 20 - 15} = \frac{3}{32},\ \text{resulta:}$$
$$p_0 = \frac{1}{\sum_{n=0}^{2} c_n + c_{\geq 3}} = \frac{1}{\frac{32+24+9+3}{32}} = \frac{32}{68} = \frac{8}{17} = 0.4706.$$

Ahora,

$$\begin{aligned} P(N_q > 2) &= 1 - (p_0 + p_1 + p_2 + p_3 + p_4 + p_5) \\ &= 1 - p_0(1 + c_1 + c_2 + c_3 + c_4 + c_5) \\ &= 1 - \frac{8}{17}\left(1 + \frac{3}{4} + \frac{9}{32} + \frac{9}{128} + \frac{9}{512} + \frac{9}{2048}\right) \\ &= 1 - \frac{8}{17} \cdot \frac{4349}{2048} = 1 - \frac{4349}{4352} = \frac{3}{4352} = 0.000689. \end{aligned}$$

Además,

$$\begin{aligned} L_q = \frac{\lambda^{s+1} p_0}{(s-1)!\mu^{s-1}(s\mu - \lambda)^2} &= \frac{15^4 \frac{8}{17}}{2 \cdot 20^2 \cdot (3 \cdot 20 - 15)^2} \\ &= \frac{1}{68} = 0.0147 \text{ procesos.} \end{aligned}$$

Se observa pues como al pasar de uno a tres procesadores el sistema experimenta una descongestión más que notable. Lo último que se pide es

$$\begin{aligned} L = \lambda W = \lambda\left(W_q + \frac{1}{\mu}\right) = L_q + \frac{\lambda}{\mu} &= \frac{1}{68} + \frac{15}{20} \\ &= \frac{13}{17} = 0.7647 \text{ procesos.} \end{aligned}$$

que indica también poca carga en la estación.

11.3 El modelo $M/M/1/K$

Se trata de un modelo como el $M/M/1$, ya estudiado, pero con límitación K para el tamaño de la cola. Es decir, 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:

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

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

$$\mu_n = \mu, \text{ para todo } n = 1, 2, \ldots$$

Haciendo uso de estas tasas se obtienen inmediatamente las c_n:

$$c_n = \begin{cases} \rho^n & \text{si } n = 1, 2, \ldots, K+1 \\ 0 & \text{si } n = K+2, K+3, \ldots \end{cases}$$

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 como 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 estuviese el sistema saturado (es decir $P(N = K+1) = 1$).

La suma de la serie de los c_n es realmente una suma finita y vale

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{K+1} \rho^n = \begin{cases} \dfrac{\rho^{K+2} - \rho}{\rho - 1} & \text{si } \rho \neq 1 \\ K+1 & \text{si } \rho = 1 \end{cases}$$

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

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

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} c_n} = \frac{1}{1 + \dfrac{\rho^{K+2} - \rho}{\rho - 1}} = \frac{\rho - 1}{\rho^{K+2} - 1}.$$

Ahora las p_n se obtienen fácilmente,

$$p_n = \begin{cases} \dfrac{\rho - 1}{\rho^{K+2} - 1}\rho^n & \text{si } n = 0, 1, \ldots, K+1 \\ 0 & \text{si } n = K+2, K+3, \ldots \end{cases}$$

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

$$L = \sum_{n=0}^{\infty} n \cdot p_n = \sum_{n=0}^{K+1} n \cdot \frac{\rho - 1}{\rho^{K+2} - 1}\rho^n = \frac{(\rho - 1)\,\rho}{\rho^{K+2} - 1} \sum_{n=0}^{K+1} n\rho^{n-1}.$$

Ahora bien, siguiendo las mismas pautas que las usadas en el $M/M/1$ 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

$$f(x) = \sum_{n=0}^{K+1} x^n = \frac{x^{K+2} - 1}{x - 1}, \text{ para } x \neq 1,$$

su derivada vale

$$\begin{aligned} f'(x) &= \sum_{n=0}^{K+1} n x^{n-1} = \frac{(K+2)\,x^{K+1}(x-1) - \left(x^{K+2} - 1\right)}{(x-1)^2} \\ &= \frac{(K+1)\,x^{K+2} - (K+2)\,x^{K+1} + 1}{(x-1)^2}. \end{aligned}$$

De esta forma

$$\sum_{n=0}^{K+1} n\rho^{n-1} = \frac{(K+1)\,\rho^{K+2} - (K+2)\,\rho^{K+1} + 1}{(\rho - 1)^2}$$

y, por tanto,

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

En resumen

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

cuyo primer sumando es precisamente el valor de L para un $M/M/1$.

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:

$$\begin{aligned}\overline{\lambda} &= \sum_{n=0}^{\infty}\lambda_n\cdot p_n=\sum_{n=0}^{K}\lambda\cdot p_n=\lambda\sum_{n=0}^{K}p_n=\lambda\left(1-p_{K+1}\right)\\ &= \lambda\left(1-\frac{\rho-1}{\rho^{K+2}-1}\rho^{K+1}\right)=\lambda\frac{\rho^{K+2}-1-(\rho-1)\rho^{K+1}}{\rho^{K+2}-1}=\frac{\lambda\left(\rho^{K+1}-1\right)}{\rho^{K+2}-1}.\end{aligned}$$

A partir de esta expresión se tiene

$$\begin{aligned}W &= \frac{L}{\overline{\lambda}}=\frac{\frac{1-\rho}{1-\rho^{K+2}}\cdot\frac{(K+1)\rho^{K+3}-(K+2)\rho^{K+2}+\rho}{(1-\rho)^2}}{\frac{\lambda\left(\rho^{K+1}-1\right)}{\rho^{K+2}-1}}\\ &= \frac{1-\rho}{1-\rho^{K+1}}\cdot\frac{(K+1)\rho^{K+3}-(K+2)\rho^{K+2}+\rho}{\lambda(1-\rho)^2}\\ &= \frac{1-\rho}{1-\rho^{K+1}}\cdot\frac{\rho-\rho^{K+2}-(K+1)(1-\rho)\rho^{K+2}}{\lambda(1-\rho)^2}\\ &= \frac{1-\rho}{1-\rho^{K+1}}\cdot\left[\frac{\rho\left(1-\rho^{K+1}\right)}{\lambda(1-\rho)^2}-\frac{(K+1)(1-\rho)\rho^{K+2}}{\lambda(1-\rho)^2}\right]\\ &= \frac{\rho}{\lambda(1-\rho)}-\frac{(K+1)\rho^{K+2}}{\lambda\left(1-\rho^{K+1}\right)}=\frac{1}{\mu-\lambda}-\frac{(K+1)\rho^{K+2}}{\lambda\left(1-\rho^{K+1}\right)},\end{aligned}$$

luego

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

que, de nuevo, es muy parecida a su análoga para el modelo $M/M/1$. A partir de esta última expresión y teniendo en cuenta que $W_q=W-\frac{1}{\mu}$, se tiene

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

Finalmente,

$$\begin{aligned}
L_q &= \overline{\lambda} W_q = \overline{\lambda}\left(W - \frac{1}{\mu}\right) = L - \frac{\overline{\lambda}}{\mu} = \frac{\rho}{1-\rho} - \frac{(K+2)\,\rho^{K+2}}{1-\rho^{K+2}} - \frac{\dfrac{\lambda\left(\rho^{K+1}-1\right)}{\rho^{K+2}-1}}{\mu} \\
&= \frac{\rho}{1-\rho} - \frac{(K+2)\,\rho^{K+2}}{1-\rho^{K+2}} - \frac{\rho\left(1-\rho^{K+1}\right)}{1-\rho^{K+2}} \\
&= \frac{\rho}{1-\rho} - \frac{(K+2)\,\rho^{K+2}}{1-\rho^{K+2}} - \frac{\rho - \rho^{K+2} + \rho^{K+3} - \rho^{K+3}}{1-\rho^{K+2}} \\
&= \frac{\rho}{1-\rho} - \frac{(K+2)\,\rho^{K+2}}{1-\rho^{K+2}} - \frac{\rho^{K+2}(\rho-1)}{1-\rho^{K+2}} - \frac{\rho\left(1-\rho^{K+2}\right)}{1-\rho^{K+2}} \\
&= \frac{\rho}{1-\rho} - \rho - \frac{(K+1+\rho)\,\rho^{K+2}}{1-\rho^{K+2}} = \frac{\rho^2}{1-\rho} - \frac{(K+1+\rho)\,\rho^{K+2}}{1-\rho^{K+2}},
\end{aligned}$$

luego, en resumen,

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

expresión, esta última, 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, L_q, W y W_q) 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 $K \to \infty$ (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 formulas anteriores para los p_n, L, L_q, W y W_q coinciden con los que aparecen en el $M/M/1$. Así, a título de ejemplo,

$$\begin{aligned}
\lim_{K\to\infty} L_{M/M/1/K} &= \frac{\rho}{1-\rho} - \lim_{K\to\infty} \frac{(K+2)\,\rho^{K+2}}{1-\rho^{K+2}} \\
&= \frac{\rho}{1-\rho} - \frac{\lim_{K\to\infty}(K+2)\,\rho^{K+2}}{1} = \frac{\rho}{1-\rho} = L_{M/M/1},
\end{aligned}$$

ya que el límite del numerador es cero (utilizando, por ejemplo, la regla de L'Hôpital).

Caso $\boldsymbol{\rho = 1}$: En este caso los cálculos resultan bastante más sencillos. El valor de p_0 viene dado por

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} c_n} = \frac{1}{1+K+1} = \frac{1}{K+2}.$$

Como además $c_n = \rho^n = 1$, para $n = 1, 2, \ldots, K+1$, entonces

$$p_n = \frac{1}{K+2}, \text{ para todo } n = 0, 1, \ldots, K+1,$$

es decir, N tiene distribución uniforme discreta sobre su conjunto de valores posibles.

En este caso,

$$\begin{aligned}
L &= \sum_{n=0}^{K+1} n \cdot p_n = \sum_{n=0}^{K+1} n \cdot \frac{1}{K+2} = \frac{1}{K+2}\frac{(K+1)(K+2)}{2} = \frac{K+1}{2}, \\
\overline{\lambda} &= \sum_{n=0}^{\infty} \lambda_n \cdot p_n = \lambda \sum_{n=0}^{K} p_n = \lambda (1 - p_{K+1}) = \lambda \left(1 - \frac{1}{K+2}\right) = \frac{\lambda (K+1)}{K+2}, \\
W &= \frac{L}{\overline{\lambda}} = \frac{\frac{K+1}{2}}{\frac{\lambda(K+1)}{K+2}} = \frac{K+2}{2\lambda}, \\
W_q &= W - \frac{1}{\mu} = \frac{K+2}{2\lambda} - \frac{1}{\lambda} = \frac{K}{2\lambda}, \\
L_q &= \overline{\lambda} W_q = \frac{\lambda (K+1)}{K+2} \cdot \frac{K}{2\lambda} = \frac{K (K+1)}{2 (K+2)}.
\end{aligned}$$

Para cualquier valor de ρ (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 está 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:

$$\begin{aligned}
q_n &= P(N = n|_{T=0}) = \frac{f(0|_{N=n})\, p_n}{\sum_{m=0}^{K} f(0|_{N=m})\, p_m} = \frac{\lambda_n e^{-\lambda_n 0} p_n}{\sum_{m=0}^{K} \lambda_m e^{-\lambda_m 0} p_m} \\
&= \frac{\lambda \cdot p_n}{\lambda \cdot \sum_{m=0}^{K} p_m} = \frac{p_n}{1 - p_{K+1}}, \text{ para } n = 0, 1, \ldots, K, \text{ mientras que} \\
q_n &= 0 \text{ para } n = K+1, K+2, \ldots
\end{aligned}$$

Es digno de mención que para modelos como el $M/M/s$ (incluyendo el caso $s = 1$), se verifica que las variables T y N son independientes y, por tanto

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

Las funciones de distribución de las variables $\mathcal{W}$ y $\mathcal{W}_q$ vienen dadas por

$$\begin{aligned}
W(t) &= 1 - e^{-\mu t} \sum_{n=0}^{K} q_n \sum_{r=0}^{n} \frac{(\mu t)^r}{r!}, \text{ si } t \geq 0 \text{ (y } W(t) = 0 \text{ en otro caso)} \\
W_q(t) &= 1 - e^{-\mu t} \sum_{n=1}^{K} q_n \sum_{r=0}^{n-1} \frac{(\mu t)^r}{r!}, \text{ si } t \geq 0 \text{ (y } W_q(t) = 0 \text{ si } t < 0),
\end{aligned}$$

entendiendo, en las expresiones anteriores que cuando $t = 0$ y $r = 0$, en el sumatorio más interno, la indeterminación 0^0 se resuelve como 1. Puede probarse de forma muy sencilla que $W_q(0) = P(\mathcal{W}_q = 0) = q_0 = P(N = 0|_{T=0})$.

Una implementación directa de las funciones $W(t)$ y $W_q(t)$ puede resultar muy lenta (sobre todo si se evalúan en muchos valores de t y K es elevado). Así, para $K = 100$, el doble bucle que aparecería en la implementación directa de $W(t)$ requeriría $1 + 2 + \cdots + 101 = \frac{101 \cdot 102}{2} = 5151$ iteraciones para cada evaluación de la función en un t. En realidad, esta implementación puede ser mejorada muchísimo en términos de eficiencia. Denotando $A_r(t) = \frac{(\mu t)^r}{r!}$ y $S_n(t) = \sum_{r=0}^{n} \frac{(\mu t)^r}{r!} = \sum_{r=0}^{n} A_r(t)$, se tiene que $S_0(t) = A_0(t) = 1$ y que se verifican las siguientes relaciones recursivas

$$\begin{aligned} A_r(t) &= \frac{\mu t}{r} A_{r-1}(t), \text{ para } r = 1, 2, \ldots, K \\ S_n(t) &= S_{n-1}(t) + A_n(t), \text{ para } n = 1, 2, \ldots, K. \end{aligned}$$

De esta forma el cálculo de $W(t)$ podría programarse con un único bucle, procediendo como en el siguiente algoritmo:

1. `Hacer` $A = 1$, $S = 1$ `y` $B = q_0$.
2. `Desde` $n = 1$ `hasta` K `repetir`
2.1. `Hacer` $A = A \cdot \frac{\mu t}{n}$ `y` $S = S + A$.
2.2. `Hacer` $B = B + q_n \cdot S$.
3. `Devolver` $1 - B \cdot e^{-\mu t}$.

Análogamente se puede proceder para el cálculo eficiente de $W_q(t)$.

Es obvio que el ahorro computacional del algortimo, frente a la implementación directa, es muy grande. Así, para el ejemplo anterior con $K = 100$, esta implementación requeriría, tan sólo, 100 iteraciones del bucle (unas 51 veces más rápido que las 5151 del método directo).

En 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 número medio de llegadas de clientes al sistema por unidad de tiempo y el número medio de clientes a los que el servidor tendría capacidad de dar servicio por unidad de tiempo, sinó más bien como un cociente semejante, pero donde el númerador 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 a través de

$$\overline{\rho} = \frac{\overline{\lambda}}{\mu} = \begin{cases} \frac{\frac{\lambda(K+1)}{K+2}}{\mu} = \frac{K+1}{K+2} & \text{si } \rho = 1 \\ \frac{\frac{\lambda\left(\rho^{K+1}-1\right)}{\rho^{K+2}-1}}{\mu} = \frac{\rho^{K+2}-\rho}{\rho^{K+2}-1} & \text{si } \rho \neq 1 \end{cases}$$

que efectivamente sí es siempre menor que 1.

11.4 El modelo $M/M/s/K$

Se trata de la extensión del modelo anterior al caso en que exista un número entero cualquiera, s, de servidores. Las tasas de llegada son casi idénticas a las del modelo $M/M/1$, mientras que las de servicio son exáctamente iguales a las de un $M/M/s$:

$$\lambda_n = \begin{cases} \lambda & \text{si } n = 0, 1, \ldots, K+s-1 \\ 0 & \text{si } n = K+s, K+s+1, \ldots \end{cases}$$

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

Como consecuencia se obtiene:

$$c_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} & \text{si } n = 1, 2, \ldots, s \\ \dfrac{\lambda^n}{s!s^{n-s}\mu^n} & \text{si } n = s+1, s+2, \ldots, K+s \\ 0 & \text{si } n = K+s+1, K+s+2, \ldots \end{cases}$$

De nuevo 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:

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{K+s} c_n = \sum_{n=1}^{s-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=s}^{K+s} \frac{\lambda^n}{s!s^{n-s}\mu^n} = \sum_{n=1}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{s!\mu^s} \sum_{n=s}^{K+s} \rho^{n-s}$$

$$= \begin{cases} \displaystyle\sum_{n=1}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{s!\mu^s} \cdot \frac{1-\rho^{K+1}}{1-\rho} & \text{si } \rho \neq 1 \\ \displaystyle\sum_{n=1}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s\,(K+1)}{s!\mu^s} & \text{si } \rho = 1 \end{cases}$$

Como consecuencia,

$$p_0 = \begin{cases} \dfrac{1}{\displaystyle\sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s}{s!\mu^s} \cdot \frac{1-\rho^{K+1}}{1-\rho}} & \text{si } \rho \neq 1 \\ \dfrac{1}{\displaystyle\sum_{n=0}^{s-1} \frac{\lambda^n}{n!\mu^n} + \frac{\lambda^s\,(K+1)}{s!\mu^s}} & \text{si } \rho = 1 \end{cases}$$

Tal y como ya se comentó para el modelo $M/M/s$, la implementación directa de las formulas anteriores no es precisamente la manera más eficiente de

calcular p_0. Así, empleando fórmulas recursivas puede calcularse p_0 mediante

$$\begin{aligned} p_0 &= \frac{1}{\sum_{n=0}^{s-1} c_n + c_{\geq s}}, \text{ donde} \\ c_0 &= 1,\ c_n = c_{n-1}\frac{\lambda}{n\mu}, \text{ para } n = 1, 2, \ldots, s-1 \\ \text{siendo } c_{\geq s} &= c_{s-1}\frac{\rho - \rho^{K+2}}{1-\rho} \text{ si } \rho \neq 1 \text{ y} \\ c_{\geq s} &= c_{s-1}\,(K+1) \text{ si } \rho = 1. \end{aligned}$$

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

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

Nuevamente, las fórmulas implementadas directamente a partir de estas expresiones son muy poco eficientes y resulta preferible proceder de forma análoga a lo hecho para el $M/M/s$.

Reutilizando los valores c_n, necesarios para la implementación del cálculo eficiente de p_0, los p_n se obtienen de manera muy sencilla: $p_n = c_n \cdot p_0$, para $n = 1, 2, \ldots, s-1$. Por otra parte, si $n \geq s$, el término p_n se calcula recursivamente comenzando en p_{s-1} que ya está calculado:

$$p_n = \frac{\lambda_{n-1}}{\mu_n} p_{n-1} = \rho \cdot p_{n-1}, \text{ para todo } n = s, s+1, \ldots, K+s$$

Fórmula que, aún siendo válida para todo valor de ρ, se trivializa si $\rho = 1$, dando como resultado $p_{K+s} = p_{K+s-1} = \cdots = p_s = p_{s-1}$ para $\rho = 1$.

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

$$\sum_{n=0}^{K+1} n x^{n-1} = \frac{(K+1)\,x^{K+2} - (K+2)\,x^{K+1} + 1}{(x-1)^2}$$

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

$$\begin{aligned} L_q &= \sum_{n=s+1}^{s+K} (n-s)\cdot p_n = \sum_{n=s}^{s+K} (n-s)\cdot \frac{\lambda^n}{s!s^{n-s}\mu^n}\cdot p_0 \\ &= \frac{\lambda^s}{s!\mu^s}\cdot p_0 \cdot \rho \sum_{n=s}^{s+K} (n-s)\cdot \rho^{n-s-1} = \frac{\lambda^s}{s!\mu^s}\cdot p_0 \cdot \rho \sum_{j=0}^{K} j\cdot \rho^{j-1} \\ &= \frac{\lambda^s}{s!\mu^s}\cdot p_0 \cdot \rho \cdot \frac{K\rho^{K+1} - (K+1)\rho^K + 1}{(\rho-1)^2} \\ &= \frac{\lambda^s \cdot p_0 \cdot \rho\left[1 + K\rho^{K+1} - (K+1)\rho^K\right]}{s!\mu^s (1-\rho)^2}. \end{aligned}$$

Por otra parte, si $\rho = 1$ entonces

$$\begin{aligned} L_q &= \sum_{n=s+1}^{s+K} (n-s)\cdot p_n = p_{s-1} \sum_{n=s+1}^{s+K} (n-s) \\ &= p_{s-1}\frac{K(K+1)}{2} = \frac{\lambda^{s-1}K(K+1)\cdot p_0}{(s-1)!\cdot 2\cdot \mu^{s-1}}. \end{aligned}$$

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

$$L_q = \begin{cases} \dfrac{\lambda^s \cdot p_0 \cdot \rho\left[1 + K\rho^{K+1} - (K+1)\rho^K\right]}{s!\mu^s (1-\rho)^2} & \text{si } \rho \neq 1 \\ \dfrac{\lambda^{s-1}K(K+1)\cdot p_0}{(s-1)!\cdot 2\cdot \mu^{s-1}} & \text{si } \rho = 1 \end{cases}$$

o bien, para un cálculo mucho más eficiente (supuestos realizados los cálculos eficientes comentados más arriba),

$$L_q = \begin{cases} \dfrac{\left[1 + K\rho^{K+1} - (K+1)\rho^K\right]\rho^2}{(1-\rho)^2}\cdot p_{s-1} & \text{si } \rho \neq 1 \\ \dfrac{K(K+1)}{2}\cdot p_{s-1} & \text{si } \rho = 1 \end{cases}$$

Para poder obtener el resto de cantidades medias deberemos usar las fórmulas de Little y calcular, primeramente, $\overline{\lambda}$:

$$\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n \cdot p_n = \lambda \sum_{n=0}^{K+s-1} p_n = \lambda\left(1 - p_{K+s}\right),$$

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

$$\overline{\lambda} = \lambda\left(1 - \frac{\lambda^s}{s!\mu^s}\rho^K p_0\right),$$

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

$$\overline{\rho} = \frac{\overline{\lambda}}{s\mu} = \rho\left(1 - \frac{\lambda^s}{s!\mu^s}\rho^K p_0\right) = \rho - \frac{\lambda^s}{s!\mu^s}\rho^{K+1}p_0.$$

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

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

En este modelo resulta bastante complicada la distribución del tiempo que un cliente está en el sistema, aunque no lo es tanto la del tiempo que un cliente pasa en la cola:

$$W_q(t) = 1 - e^{-s\mu t}\sum_{n=s}^{K+s-1} q_n \sum_{r=0}^{n-s}\frac{(s\mu t)^r}{r!}, \text{ si } t \geq 0 \text{ (y } W_q(t) = 0 \text{ si } t < 0),$$

donde las q_n tienen el mismo significado que en el modelo $M/M/1/K$, es decir, probabilidad de que haya n clientes en el sistema justo cuando una llegada se está produciendo, y vienen dadas por

$$q_n = \frac{p_n}{1 - p_{K+s}},\ n = 0, 1, \ldots, K+s-1.$$

Los comentarios hechos sobre la forma de implementar las funciones $W(t)$ y $W_q(t)$ en un modelo $M/M/1/K$, son válidos para la implementación de $W_q(t)$ en este modelo. Así el aparente doble bucle (en n y en r), en la expresión para $W_q(t)$, puede reducirse a un bucle simple utilizando una técnica totalmente análoga a la de entonces.

Un caso particular de este modelo es el $M/M/s/0$, que es uno de los primeros modelos que estudió Erlang, en los albores de la Teoría de Colas, para analizar la probabilidad de n líneas ocupadas en una central telefónica que dispone de s líneas. En el modelo la capacidad de la cola es $K = 0$, pues se supone que, si una llamada llega a la central cuando todas las líneas están ocupadas, no permanece en espera sino que recibe la señal de saturación en la red abortando la posibilidad de comunicación en ese momento. Como caso particular de lo visto en el modelo $M/M/s/K$, se tiene que la probabilidad de

n líneas ocupadas en la central viene dada por

$$p_n = \frac{\frac{\lambda^n}{n!\mu^n}}{\sum_{j=0}^{s} \frac{\lambda^j}{j!\mu^j}}, \text{ para } n = 0, 1, \ldots, s,$$

que es la famosa fórmula de Erlang, quien probó que también es válida para cualquier distribución del tiempo de servicio, entendiendo entonces que μ es la inversa de la media de dicha distribución.

Ejemplo 49 *Considérese la estación de trabajo analizada en los dos últimos ejemplos y supóngase que aún teniendo tres procesadores, sólo es posible mantener un único proceso en cola de espera. Calcular las tres cantidades pedidas en el último ejemplo.*

Solución: El modelo pasa a ser ahora un $M/M/s/K$ *con* $s = 3$, $K = 1$, $\lambda = 15$ *y* $\mu = 20$. *En primer lugar, resulta obvio que* $P(N_q > 2) = 0$, *ya que el límite para la cola es* $K = 1$. *Por otra parte, aunque podríamos utilizar todas las fórmulas obtenidas anteriormente para el cálculo de* L *y* L_q, *dado que sólo hay cinco estados posibles en el sistema* ($N = 0, 1, 2, 3, 4$) *quizá sea más rápido calcularlas directamente:*

$$\begin{aligned}
c_0 &= 1,\ c_1 = c_0\frac{15}{20} = \frac{3}{4},\ c_2 = c_1\frac{15}{2\cdot 20} = \frac{9}{32},\\
c_3 &= c_2\frac{15}{3\cdot 20} = \frac{9}{128} \text{ y } c_4 = c_3\frac{15}{3\cdot 20} = \frac{9}{512},\\
\text{siendo } c_n &= 0 \text{ para } n = 5, 6, \ldots \text{ De esta forma,}\\
p_0 &= \frac{1}{1 + \frac{3}{4} + \frac{9}{32} + \frac{9}{128} + \frac{9}{512}} = \frac{1}{\frac{1085}{512}} = \frac{512}{1085} = 0.4719.
\end{aligned}$$

Utilizando todo lo anterior:

$$\begin{aligned}
L &= \sum_{n=0}^{4} np_n = p_0\left(c_1 + 2c_2 + 3c_3 + 4c_4\right)\\
&= \frac{512}{1085}\left(\frac{3}{4} + \frac{2\cdot 9}{32} + \frac{3\cdot 9}{128} + \frac{4\cdot 9}{512}\right) = \frac{512}{1085}\frac{51}{32}\\
&= \frac{816}{1085} = 0.7521 \text{ procesos.}
\end{aligned}$$

Asimismo,

$$\begin{aligned}
L_q &= 0\cdot(p_0 + p_1 + p_2 + p_3) + 1\cdot p_4 = \frac{9}{1085} =\\
&= 0.008295 \text{ procesos.}
\end{aligned}$$

Se observa entonces que el sistema está todavía menos cargado que antes. La razón es que la limitación en el tamaño de la cola impide un número elevado de clientes en el sistema (aunque antes dicho número ya era muy improbable).

11.5 El modelo $M/M/1/\infty/H$

Estudiaremos ahora un modelo en el que la población potencial es finita, formada por H individuos. Dado que ahora es obvio que la distribución del tiempo entre dos llegadas de clientes consecutivos al sistema dependerá de cuántos clientes estén en el sistema, una forma razonable de tener esto en cuenta es considerar que, para cada cliente que acaba de salir del sistema, la distribución de probabilidad del tiempo que falta para que vuelva a entrar en él es una $\exp(\lambda)$. Tal como indica la nomenclatura introducida, se supondrá un único servidor (que atiende los clientes según una disciplina FIFO) y que no hay límite para el tamaño de la cola.

En virtud de la carencia de memoria de la distribución exponencial y de la probabilidad que nos daba la distribución del mínimo de varias exponenciales independientes, podemos razonar cuál será la distribución del tiempo que falta para que el próximo cliente entre el sistema cuando ya hay n clientes dentro de él. Así, si $N = n$, el número de clientes en la población potencial será $H - n$ y el tiempo hasta la próxima entrada en el sistema será el mínimo de $H - n$ exponenciales de parámetro λ, independientes entre sí. Se trata por tanto de una $\exp((H-n)\lambda)$. Dado que los tiempos de servicio son también exponenciales, el sistema puede modelizarse mediante un proceso de nacimiento y muerte con tasas de llegada

$$\lambda_n = \begin{cases} (H-n)\lambda & \text{si } n = 0, 1, \ldots, H \\ 0 & \text{si } n = H, H+1, \ldots \end{cases}$$

y tasas de servicio

$$\mu_n = \mu, \text{ si } n = 1, 2, \ldots$$

Aunque puede encontrarse una expresión explícita para las c_n:

$$c_n = \begin{cases} \dfrac{H!}{(H-n)!}\rho^n & \text{si } n = 1, 2, \ldots, H \\ 0 & \text{si } n = H+1, H+2, \ldots \end{cases}$$

resulta mucho más eficiente obtener p_0 mediante un cálculo recursivo

$$\begin{aligned} p_0 &= \frac{1}{\sum_{n=0}^{H} c_n}, \text{ donde} \\ c_0 &= 1,\ c_n = (H-n+1)\cdot\rho\cdot c_{n-1}, \text{ para } n = 1, 2, \ldots, H \end{aligned}$$

Además, el resto de probabilidades se calculan a partir de $p_n = c_n \cdot p_0$, para $n = 1, 2, \ldots, H$, ya que $p_n = 0$ si $n > H$. Obviamente el modelo es siempre estacionario.

En este caso no resulta sencillo obtener una fórmula explícita de L ó L_q, pues la suma resultante no corresponde a una serie geométrica o convertible en geométrica. Por este motivo, debemos calcular directamente

$$L = \sum_{n=1}^{H} n \cdot p_n.$$

Además,

$$\begin{aligned} \overline{\lambda} &= \sum_{n=0}^{H} \lambda_n \cdot p_n = \sum_{n=0}^{H} (H-n)\,\lambda \cdot p_n = \lambda \cdot \left(\sum_{n=0}^{H} H \cdot p_n - \sum_{n=0}^{H} n \cdot p_n \right) \\ &= \lambda \cdot (H - L)\,. \end{aligned}$$

Teniendo esto en cuenta se llega a que

$$\begin{aligned} W &= \frac{L}{\overline{\lambda}} = \frac{L}{\lambda \cdot (H-L)}, \\ W_q &= W - \frac{1}{\mu}, \\ L_q &= \overline{\lambda} W_q. \end{aligned}$$

Además, la intensidad de tráfico efectiva es

$$\overline{\rho} = \frac{\overline{\lambda}}{\mu} = \frac{\lambda \cdot (H-L)}{\mu} = \rho \cdot (H - L)\,.$$

Definiendo las probabilidades $q_n = P\left(N = n|_{T=0}\right)$, como en modelos anteriores, se tiene

$$\begin{aligned} q_n &= P\left(N = n|_{T=0}\right) = \frac{f\left(0|_{N=n}\right) p_n}{\sum_{m=0}^{H} f\left(0|_{N=m}\right) p_m} = \frac{\lambda_n e^{-\lambda_n 0} p_n}{\sum_{m=0}^{H} \lambda_m e^{-\lambda_m 0} p_m} \\ &= \frac{(H-n)\,\lambda \cdot p_n}{\sum_{m=0}^{H} (H-m)\,\lambda \cdot p_m} = \frac{(H-n) \cdot p_n}{H \sum_{m=0}^{H} p_m - \sum_{m=0}^{H} m \cdot p_m} \\ &= \frac{(H-n) \cdot p_n}{H-L}, \text{ para } n = 0, 1, \ldots, H-1, \text{ mientras que} \\ q_n &= 0 \text{ para } n = H, H+1, \ldots \end{aligned}$$

Estas probabilidades aparecen en las expresiones para $W(t)$ y $W_q(t)$:

$$\begin{aligned} W(t) &= 1 - e^{-\mu t} \sum_{n=0}^{H-1} q_n \sum_{r=0}^{n} \frac{(\mu t)^r}{r!}, \text{ si } t \geq 0 \text{ (y } W(t) = 0 \text{ en otro caso)} \\ W_q(t) &= 1 - e^{-\mu t} \sum_{n=1}^{H-1} q_n \sum_{r=0}^{n-1} \frac{(\mu t)^r}{r!}, \text{ si } t \geq 0 \text{ (y } W_q(t) = 0 \text{ si } t < 0). \end{aligned}$$

De nuevo, estas fórmulas se pueden implementar de manera mucho más eficiente de lo que podría parecer a primera vista, gracias a un único bucle con cálculos recursivos.

11.6 El modelo $M/M/s/\infty/H$

Generaliza el anterior a un número entero, s, cualquiera de servidores. Al igual que en aquel, la población potencial es finita, con H elementos y, por tanto, la distribución del tiempo entre dos llegadas de clientes consecutivos al sistema dependerá de cuántos clientes estén en el sistema, suponiendo, de nuevo, que cada cliente que acaba de salir del sistema, vuelve a entrar en él al cabo de un tiempo aleatorio con distribución $\exp(\lambda)$. Igual que en el modelo anterior la disciplina es FIFO y la capacidad de la cola ilimitada. En este modelo supondremos $s \leq H$. Esta restricción, además de ser bastante realista, no supone ninguna pérdida de generalidad pues un modelo con $s > H$ puede resolverse haciendo el número de servidores igual a H. De esta forma, cantidades como L, L_q, W, W_q, o las p_n, calculadas a partir del modelo modificado, serían las mismas que para el modelo original. Tan sólo algunas cuestiones como el número medio de servidores ocupados necesitarían algún reajuste posterior para tener en cuenta el verdadero valor de $s > H$.

Los razonamientos realizados para el modelo $M/M/1/\infty/H$, sobre las tasas de llegada son perfectamente válidos ahora, obteniendo, como consecuencia,

$$\lambda_n = \begin{cases} (H-n)\,\lambda & \text{si } n = 0, 1, \ldots, H \\ 0 & \text{si } n = H, H+1, \ldots \end{cases}$$

Por su parte, las tasas de servicio son igual que para los modelos $M/M/s$ ó $M/M/s/K$,

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

Una expresión explícita para las constantes c_n es

$$c_n = \begin{cases} \dfrac{H!}{(H-n)!} \cdot \dfrac{\lambda^n}{n!\mu^n} & \text{si } n = 1, 2, \ldots, s \\ \dfrac{H!}{(H-n)!} \cdot \dfrac{\lambda^s}{s!\mu^s} \cdot \rho^{n-s} & \text{si } n = s, s+1, \ldots, H \\ 0 & \text{si } n = H+1, H+2, \ldots \end{cases}$$

Consecuentemente el modelo es estacionario siempre. De nuevo, resulta mucho más eficiente un cálculo recursivo de las c_n para obtener p_0:

$$\begin{aligned} p_0 &= \frac{1}{\sum_{n=0}^{H} c_n}, \text{ con} \\ c_0 &= 1,\ c_n = (H-n+1) \cdot \frac{\lambda}{n\mu} \cdot c_{n-1}, \text{ si } n = 1, 2, \ldots, s, \\ c_n &= (H-n+1) \cdot \rho \cdot c_{n-1}, \text{ para } n = s+1, s+2, \ldots, H. \end{aligned}$$

Como siempre las demás probabilidades se obtienen como $p_n = c_n \cdot p_0$, para $n = 1, 2, \ldots, H$, y $p_n = 0$ si $n > H$.

Como cabe esperar (pues ya ocurría en el modelo anterior con $s = 1$), no podemos obtener una fórmula explícita para L ó L_q. Así, se calcula directamente

$$L = \sum_{n=1}^{H} n \cdot p_n,$$

demostrando, igual que en el modelo anterior, que $\overline{\lambda} = \lambda \cdot (H - L)$. Las fórmulas de Little y la relación entre tiempos medios permiten ahora obtener W, W_q y L_q. También ahora la intensidad de tráfico efectiva es $\overline{\rho} = \rho \cdot (H - L)$.

La expresión para $W(t)$, en este modelo, es muy complicada. Seguidamente se incluye la de $W_q(t)$, que vuelve a depender de $q_n = P(N = n|_{T=0}) = \frac{(H-n)\cdot p_n}{H-L}$ $(n = 0, 1, \ldots, H-1)$,

$$W_q(t) = 1 - e^{-s\mu t} \sum_{n=s}^{H-1} q_n \sum_{r=0}^{n-s} \frac{(s\mu t)^r}{r!}, \text{ si } t \geq 0 \text{ (y } W_q(t) = 0 \text{ si } t < 0).$$

Como en ocasiones anteriores, el aparente doble bucle que conlleva la implementación de $W_q(t)$ es evitable y puede programarse por medio de un único bucle con cálculos recursivos.

Ejemplo 50 *Se dispone de un sistema informático con 5 estaciones que cada cierto tiempo deben realizar un back-up, impidendo durante ese tiempo su utilización por parte de los usuarios del sistema. Debido a las características del sistema los back-up no se producen a una hora fija sino cada vez que se alcanza cierto límite de capacidad. Se ha comprobado que el tiempo que transcurre desde que una de estas estaciones termina un back-up hasta que lo vuelve a realizar es aleatorio y con distribución exponencial de media de 2 horas. Por su parte, la duración del proceso de back-up también es aleatoria y con distribución exponencial de media 5 minutos. Además existen dos unidades de cinta contra las que puede realizarse el proceso de back-up, permaneciendo la estación en espera si ambas estuviesen siendo utilizadas por otras estaciones. Se pide:*

1. *Calcular la probabilidad de que una estación tenga que esperar para iniciar su back-up.*
2. *¿Cuál es el número medio de estaciones disponibles para los usuarios?*
3. *Encontrar el tiempo medio que cada estación está fuera de servicio con motivo de un back-up.*
4. *¿Durante qué porcentaje de tiempo es utilizada cada unidad de cinta?*

Solución: Se trata de un modelo $M/M/s/\infty/H$, con $s = 2$ (unidades de cinta), $H = 5$ (estaciones), $\frac{1}{\lambda} = 2$ y $\frac{1}{\mu} = \frac{5}{60}$, expresando el tiempo en horas.

De esto se obtiene que $\lambda = \frac{1}{2}$ *y* $\mu = 12$. *En primer lugar se calcularán las constantes* c_n *y el valor* p_0*:*

$$\begin{aligned}
c_0 &= 1,\ c_1 = c_0\frac{5\cdot\frac{1}{2}}{12} = \frac{5}{24},\\
c_2 &= c_1\frac{4\cdot\frac{1}{2}}{2\cdot 12} = \frac{5}{24\cdot 12} = \frac{5}{288},\\
c_3 &= c_2\frac{3\cdot\frac{1}{2}}{2\cdot 12} = \frac{5}{288}\cdot\frac{3}{48} = \frac{5}{4608},\\
c_4 &= c_3\frac{2\cdot\frac{1}{2}}{2\cdot 12} = \frac{5}{4608\cdot 24} = \frac{5}{110592},\\
c_5 &= c_4\frac{\frac{1}{2}}{2\cdot 12} = \frac{5}{110592\cdot 48} = \frac{5}{5308416}\ \textit{y así}\\
p_0 &= \frac{1}{c_0+c_1+c_2+c_3+c_4+c_5}\\
&= \frac{1}{1+\frac{5}{24}+\frac{5}{288}+\frac{5}{4608}+\frac{5}{110592}+\frac{5}{5308416}}\\
&= \frac{1}{\frac{6512501}{5308416}} = \frac{5308416}{6512501} = 0.81511
\end{aligned}$$

Resolvamos apartado por apartado:

1. *Se trata de* $q_2+q_3+q_4$. *Como* $q_n = \frac{(H-n)\cdot p_n}{H-L}$. *Primeramente debemos calcular*

$$\begin{aligned}
H-L &= 5p_0+4p_1+3p_2+2p_3+p_4 = p_0\left(5c_0+4c_1+3c_2+2c_3+c_4\right)\\
&= \frac{5308416}{6512501}\left(5\cdot 1+4\cdot\frac{5}{24}+3\cdot\frac{5}{288}+2\cdot\frac{5}{4608}+\frac{5}{110592}\right)\\
&= \frac{5308416}{6512501}\cdot\frac{651125}{110592} = \frac{31254000}{6512501} = 4.799,
\end{aligned}$$

con lo cual

$$\begin{aligned}
q_2 &= \frac{(5-2)\cdot\frac{5308416}{6512501}\cdot\frac{5}{288}}{\frac{31254000}{6512501}} = \frac{1152}{130225} = 0.008846,\\
q_3 &= \frac{(5-3)\cdot\frac{5308416}{6512501}\cdot\frac{5}{4608}}{\frac{31254000}{6512501}} = \frac{48}{130225} = 3.6859282\times 10^{-4},\\
q_4 &= \frac{(5-4)\cdot\frac{5308416}{6512501}\cdot\frac{5}{110592}}{\frac{31254000}{6512501}} = \frac{1}{130225} = 7.6790171\times 10^{-6}.
\end{aligned}$$

De forma que $q_2+q_3+q_4 = \frac{1152+48+1}{130225} = \frac{1201}{130225} = 0.009222$.

2. *Se pide* $H-L$, *que ya fue calculado previamente:*

$$H-L = 4.799 \textit{ estaciones.}$$

3. En primer lugar, obsérvese que

$$\overline{\lambda} = \lambda\,(H - L) = \frac{1}{2} \cdot \frac{31254000}{6512501} = \frac{15627000}{6512501} = 2.3995.$$

A partir de este valor se obtiene

$$\begin{aligned} W &= \frac{L}{\overline{\lambda}} = \frac{5 - \frac{31254000}{6512501}}{\frac{15627000}{6512501}} = \frac{5 - 4.799}{2.3995} = 0.0837 \text{ } horas \\ &\simeq 5 \text{ } minutos \text{ } y \text{ } 1.3 \text{ } segundos. \end{aligned}$$

4. La probabilidad de que una unidad de cinta esté desocupada es

$$\begin{aligned} p_{desocupada} &= 1 \cdot p_0 + \frac{1}{2} \cdot p_1 + 0 \cdot (p_2 + p_3 + p_4 + p_5) = p_0 \left(1 + \frac{c_1}{2}\right) \\ &= \frac{5308416}{6512501}\left(1 + \frac{\frac{5}{24}}{2}\right) = \frac{5861376}{6512501} = 0.900019 \text{ } y \text{ } así \\ p_{ocupada} &= 0.099981. \end{aligned}$$

Así, se tiene que el porcentaje de tiempo en el que cada unidad está ocupada es ligeramente inferior al 10% *(concretamente el* 9.998%*).*

11.7 El modelo $M/M/s/\infty/H$ con repuestos

Es una modificación del modelo anterior que responde a la idea de ocupar el espacio vacío que deja un cliente en la población potencial, al entrar en el sistema, con un nuevo cliente un tanto especial (llamado repuesto), que en un principio no se encontraba en ella. Esto se hará en tanto existan repuestos disponibles.

El modelo $M/M/s/\infty/H$ con Y repuestos suele utilizarse en situaciones donde los clientes están realizando cierto tipo de cometido cuando se hallan en la población potencial y entran en el sistema cuando se averían (representando, por tanto, el sistema el mecanismo de reparación). Ejemplos reales de esta situación pueden ser

- H máquinas de una factoría que están funcionando en una cadena de montaje (en la cual no podemos instalar más de H máquinas) para las cuales tenemos Y máquinas de repuesto por si se avería alguna de ellas.

- el propietario de H licencias de máquinas "tragaperras", que dispone de un stock de otras Y máquinas más, para no dejar de obtener recaudación cuando una de las H máquinas se avería y ha de pasar cierto tiempo siendo reparada.

De forma similar a lo que ocurría en el modelo anterior, supondremos $s \leq Y + H$, aunque el caso $s > Y + H$ podría cubrirse, en casi todo, con aquél en el que se da la igualdad.

Mientras existan repuestos suficientes, las tasas de llegadas de clientes de la población potencial serán constantemente iguales a $H \cdot \lambda$, pues habrá siempre H clientes fuera del sistema. A partir del caso $N = Y + 1$, las tasas decrecerán como ya lo hacían en el modelo sin repuestos. Como consecuencia de todo ello, las tasas de llegada son

$$\lambda_n = \begin{cases} H\lambda & \text{si } n = 0, 1, \ldots, Y \\ (H + Y - n)\,\lambda & \text{si } n = Y, Y+1, \ldots, Y+H \\ 0 & \text{si } n = Y+H, Y+H+1, \ldots \end{cases}$$

Las tasas de servicio son exactamente iguales que en el modelo sin repuestos:

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

Es inmediato probar que el modelo es estacionario siempre (ya que hay sólo un número finito de constantes c_n no nulas). Además, pueden obtenerse expresiones explícitas de las c_n distinguiendo dos casos: si $s \leq Y$,

$$c_n = \begin{cases} \dfrac{H^n \lambda^n}{n! \mu^n} & \text{si } n = 1, 2, \ldots, s \\ \dfrac{H^n \lambda^n}{s! s^{n-s} \mu^n} & \text{si } n = s, s+1, \ldots, Y \\ \dfrac{H^Y H! \lambda^n}{(H + Y - n)! s! s^{n-s} \mu^n} & \text{si } n = Y, Y+1, \ldots, Y+H \\ 0 & \text{si } n = Y+H+1, Y+H+2, \ldots \end{cases}$$

y si $Y \leq s \leq Y + H$,

$$c_n = \begin{cases} \dfrac{H^n \lambda^n}{n! \mu^n} & \text{si } n = 1, 2, \ldots, Y \\ \dfrac{H^Y H! \lambda^n}{(H + Y - n)! n! \mu^n} & \text{si } n = Y, Y+1, \ldots, s \\ \dfrac{H^Y H! \lambda^n}{(H + Y - n)! s! s^{n-s} \mu^n} & \text{si } n = s, s+1, \ldots, Y+H \\ 0 & \text{si } n = Y+H+1, Y+H+2, \ldots \end{cases}$$

Como en modelos anteriores, es mucho más eficiente calcular recursivamente

estas constantes. Para ello se hace $c_0 = 1$. Además,

$$\begin{aligned}
\text{Si } s &\leq Y, \text{ hacer } c_n = \frac{H\lambda}{n\mu} \cdot c_{n-1}, \text{ para } n = 1, 2, \dots, s, \\
c_n &= H \cdot \rho \cdot c_{n-1}, \text{ para } n = s+1, s+2, \dots, Y, \\
c_n &= (H + Y - n + 1) \cdot \rho \cdot c_{n-1}, \text{ si } n = Y+1, Y+2, \dots, Y+H. \\
\text{Si } Y &\leq s \leq Y + H, \text{ hacer } c_n = \frac{H\lambda}{n\mu} \cdot c_{n-1}, \text{ para } n = 1, 2, \dots, Y, \\
c_n &= \frac{(H + Y - n + 1)\,\lambda}{n\mu} \cdot c_{n-1}, \text{ para } n = Y+1, Y+2, \dots, s, \\
c_n &= (H + Y - n + 1) \cdot \rho \cdot c_{n-1}, \text{ si } n = s+1, s+2, \dots, Y+H.
\end{aligned}$$

Con estos valores se calcularía

$$p_0 = \frac{1}{\sum_{n=0}^{H+Y} c_n}$$

y el resto de probabilidades no nulas se obtendrían a partir de $p_n = c_n \cdot p_0$, para $n = 1, 2, \dots, Y + H$. El valor de L habrá de calcularse directamente de su definición, $L = \sum_{n=1}^{Y+H} n \cdot p_n$, mientras que

$$\begin{aligned}
\overline{\lambda} &= \sum_{n=0}^{Y+H} \lambda_n \cdot p_n = \sum_{n=0}^{Y-1} H\lambda \cdot p_n + \sum_{n=Y}^{Y+H} (H + Y - n)\,\lambda \cdot p_n \\
&= H\lambda \cdot \sum_{n=0}^{Y-1} p_n + H\lambda \cdot \sum_{n=Y}^{Y+H} p_n + \sum_{n=Y}^{Y+H} (Y - n)\,\lambda \cdot p_n \\
&= \lambda \left(H - \sum_{n=Y}^{Y+H} (n - Y) \cdot p_n \right).
\end{aligned}$$

Gracias a las fórmulas de Little y la relación entre los tiempos medios de espera en el sistema y en la cola, se obtienen fácilmente W, W_q y L_q. Además, la intensidad de tráfico efectiva es $\overline{\rho} = \frac{\overline{\lambda}}{s\mu} = \rho \cdot \left(H - \sum_{n=Y}^{Y+H} (n - Y) \cdot p_n \right)$.

La expresión para $W_q(t)$ viene dada por

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

siendo

$$q_n = P\left(N = n|_{T=0}\right) = \begin{cases} \dfrac{H \cdot p_n}{H - \sum_{m=Y}^{Y+H} (m - Y) \cdot p_m} & \text{si } n = 0, 1, \dots, Y-1 \\ \dfrac{(H + Y - n) \cdot p_n}{H - \sum_{m=Y}^{Y+H} (m - Y) \cdot p_m} & \text{si } n = Y, \dots, Y+H-1 \end{cases}$$

Al igual que en modelos precedentes $W_q(t)$ puede implementarse de manera mucho más eficiente que con el bucle anidado que sugiere su expresión. Por su parte la función de distribución del tiempo de un cliente en el sistema, $W(t)$, es de expresión mucho más complicada y la obviaremos.

11.8 El modelo $M/M/\infty$

Como su nombre indica, se trata de un modelo con infinitos servidores, distribución del tiempo entre llegadas $\exp(\lambda)$, distribución del tiempo del servicio $\exp(\mu)$, sin limitación para la cola, población potencial infinita y disciplina FIFO. En realidad, el hecho de que haya infinitos servidores implica que tanto la capacidad de la cola como la disciplina son parámetros irrelevantes. Una de las situaciones prácticas en la que puede tener sentido el considerar un número infinito de servidores es aquella en la que cada cliente se da servicio a sí mismo. Por ese motivo este modelo se denomina, en ocasiones, modelo de servicio amplio o de autoservicio.

Sin realizar cálculo alguno, es intuitivo que, para este modelo, N_q y $\mathcal{W}_q$ serán variables aleatorias que toman el valor cero con probabilidad 1, pues nunca habrá clientes que esperen en la cola (todos pasan directamente a ser servidos) y, como consecuencia, $L_q = 0$ y $W_q = 0$. Por el motivo antes citado, se tiene también que $\mathcal{W} \overset{\mathrm{d}}{=} \exp(\mu)$ y, en particular, $W = \frac{1}{\mu}$. Por último, la primera fórmula de Little nos lleva a que $L = \frac{\lambda}{\mu}$.

De manera un poco más precisa, las tasas de llegada son

$$\lambda_n = \lambda, \text{ para todo } n = 0, 1, \ldots$$

mientras que las tasas de servicio vienen dadas por

$$\mu_n = n\mu, \text{ para } n = 1, 2, \ldots$$

Como consecuencia se tiene,

$$c_n = \frac{\lambda^n}{n!\mu^n}, \text{ si } n = 1, 2, \ldots$$

Dado que la serie

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}$$

es siempre convergente, entonces el modelo siempre alcanza un estado estacionario. Además,

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} c_n} = \frac{1}{\sum_{n=0}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}} = \frac{1}{e^{\frac{\lambda}{\mu}}} = e^{-\frac{\lambda}{\mu}},$$

mientras que cualquier otra probabilidad viene dada por

$$p_n = c_n \cdot p_0 = \frac{e^{-\frac{\lambda}{\mu}} \left(\frac{\lambda}{\mu}\right)^n}{n!},\ n = 1, 2, \ldots$$

A la vista de esto se deduce que la variable aleatoria N tiene distribución de Poisson de parámetro $\frac{\lambda}{\mu}$. En particular su media es $L = \frac{\lambda}{\mu}$, tal y como ya habíamos anticipado. Por último, la intensidad de tráfico es, en sentido estricto, cero, pues la capacidad de servicio es infinitamente mayor que la llegada de clientes al sistema.

11.9 Ejercicios propuestos

1. El sistema de la cola de un surtidor de gasolina se adecúa a un modelo $M/M/1$ estacionario. Se sabe que es dos veces más probable que se halle tan sólo una persona esperando para hacer uso del surtidor que que estén dos esperando. El número medio de clientes que acuden por hora es de 40. Se pide:

   (a) ¿Cuál es la probabilidad de que el surtidor de gasolina esté vacío?

   (b) ¿Cuál es el tiempo medio que un cliente necesita para llenar el depósito de su vehículo desde que llega a la gasolinera?

2. Una sucursal bancaria dispone de 3 cajeros automáticos. De vez en cuando el papel de algún cajero se atasca y el aparato deja de funcionar hasta que uno de los empleados (especialmente adiestrado para llevar a cabo esta tarea) consigue arreglar la avería. Se sabe que el tiempo que utiliza dicho empleado sigue una distribución exponencial con media de 10 minutos, mientras que la distribución del tiempo que un cajero está funcionando hasta que se atasca el papel es también exponencial pero con media de 2 horas. Calcular:

   (a) La probabilidad de que funcionen los tres cajeros.

   (b) El número medio de cajeros averiados.

   (c) El tiempo medio que un cajero está averiado.

   (d) Si en un momento dado funcionan los tres cajeros, ¿cuál es el tiempo medio hasta la próxima avería?

3. A fin de evitar el envío masivo de procesos a una cola de ejecución, un centro de cálculo establece el límite máximo de tres procesos en espera de comenzar a ser ejecutados (a parte del proceso que se encuentre en ese

momento activo) bajo una disciplina FIFO. Los procesos se envían según un proceso de Poisson, a razón de 4 por minuto. El tiempo de CPU de un proceso sigue una distribución exponencial de media 1 minuto. Se pide:

(a) La probabilidad de que un proceso enviado a cola sea rechazado por estar ésta llena.

(b) El número medio de procesos en espera de comenzar a ejecutarse.

(c) El tiempo medio que transcurre entre la entrada en cola de un proceso y el fin de su ejecución.

(d) El porcentaje de tiempo en el que la CPU está inactiva.

4. A una centralita telefónica llegan llamadas, según un proceso de Poisson, a razón de 4 por minuto. El tiempo que emplea la única telefonista en direccionar cada llamada sigue una distribución exponencial con media de 10 segundos. Suponiendo que las llamadas recibidas mientras la operadora está ocupada se mantienen en espera para ser atendidas según su orden de llegada, calcular:

   (a) El número medio de llamadas en espera (sin tener en cuenta las que se están atendiendo).

   (b) La probabilidad de que una llamada tenga que esperar más de un minuto hasta que sea direccionada.

   (c) Asumiendo ahora que no es posible mantener llamadas en espera cuando la telefonista está ocupada, calcular la probabilidad de que una llamada realizada obtenga la señal de "comunicando".

5. Obtener una fórmula, lo más simple posible, para el cálculo de la probabilidad de que haya más o igual que k clientes en un modelo $M/M/1$.

6. Una estación de trabajo con dos procesadores, cada uno de los cuales es capaz de atender 60 procesos por hora, da servicio de forma que un mismo proceso sólo es atendido por un único servidor y según un riguroso orden de llegada. El tiempo de servicio de un proceso sigue un distribución exponencial y además, los procesos llegan según un proceso de Poisson, a razón de 100 cada hora. Se pide:

   (a) La probabilidad de que estén más de tres procesos en la estación.

   (b) La probabilidad de que haya algún procesador libre.

   (c) El número medio de procesadores ocupados.

7. Un laboratorio de informática consta de 5 estaciones de trabajo. Cada estación se avería, por término medio, una vez cada 30 días, siendo el tiempo hasta la próxima avería, de distribución exponencial. El laboratorio dispone de dos personas que, en caso de ser necesario, pueden

arreglar estas averías. El tiempo de reparación (para cada uno de los técnicos) es exponencial, con media de 3 días. Calcular:

(a) El número medio de estaciones funcionando.

(b) El porcentaje de tiempo que cada uno de los técnicos puede dedicar a otras tareas ajenas a la reparación de las estaciones.

8. A una CPU llegan procesos, a razón de 5 por minuto, que son atendidos sobre la base de que el primero que llega es el primero en ser ejecutado en su totalidad. El tiempo entre la llegada de dos procesos consecutivos es de distribución exponencial y el tiempo de ejecución de los procesos también es exponencial, teniendo como media 5 segundos.

(a) ¿Es el sistema estacionario? ¿Por qué?

(b) ¿Cuánto tiempo transcurre, por término medio, desde que llega un proceso a la CPU hasta que termina de ejecutarse? ¿Qué relación guarda con el tiempo medio de ejecución?

(c) Calcular la probabilidad de que un proceso que llega pueda ser ejecutado sin espera previa y el número medio de procesos en espera de ejecución o en ejecución propiamente dicha.

9. A una máquina perforadora de una cadena de producción llegan mecanismos de interruptores diferenciales según un proceso de Poisson, con media de 10 por minuto. El tiempo, en minutos, necesario para llevar a cabo la perforación del mecanismo es de distribución exponencial con parámetro 12. Cuando un nuevo mecanismo llega a la máquina perforadora y ésta está ocupada, aguarda, según el turno que le corresponda, hasta que pueda ser perforado. A tal efecto, se supone que la *zona de espera* en la que se van almacenando los mecanismos antes de ser perforados es lo suficientemente amplia para que no existan aglomeraciones que sobrepasen estas dimensiones.

(a) ¿Cuál es el porcentaje de tiempo durante el cual la perforadora está libre?

(b) ¿Cuál es el número medio de mecanismos en toda la zona de perforación (perforadora y zona de espera)?

(c) Calcular el tiempo medio que un mecanismo pasa en todo el proceso de perforación (desde que llega a esa zona hasta que sale perforado) y la probabilidad de que para un mecanismo se emplee más de un minuto en todo ese proceso.

(d) Si ahora se supone que la zona de espera tiene sólo capacidad para 3 mecanismos y que cuando un mecanismo que llega y se encuentra dicha zona completa, se desvía a otra rama de la cadena de producción, calcular la probabilidad de que se produzca dicho desvío.

(e) Bajo el supuesto del apartado anterior, ¿cuál es ahora el tiempo medio desde la llegada de un mecanismo hasta su perforación?

10. Un centro de cálculo dispone de 4 estaciones de trabajo. De vez en cuando alguna de las estaciones queda fuera de servicio y uno de los dos analístas del citado centro ha de subsanar el problema poniendo de nuevo en funcionamiento dicha estación. Se sabe que el tiempo que tarda cualquiera de los analistas en levantar el sistema del servidor sigue una distribución exponencial con media de 10 minutos, mientras que la distribución del tiempo que un servidor está funcionando, hasta que queda fuera de servicio, es también exponencial pero con media de 2 horas. Calcular:

    (a) La probabilidad de que funcionen las cuatro estaciones.
    (b) El número medio de estaciones averiadas.
    (c) El tiempo medio que una estación está averiada.
    (d) Si en un momento dado funcionan las cuatro estaciones, ¿cuál es el tiempo medio hasta la próxima avería?
    (e) Calcular el porcentaje de tiempo que cada analista dedica a subsanar averías.

11. Un sistema informático de una biblioteca dispone de 3 lectores de CD que funcionan ininterrumpidamente. No obstante, de vez en cuando se produce algún error de lectura en alguno de ellos y deja de funcionar hasta que uno de los encargados de la biblioteca (que es quien siempre lleva a cabo esta tarea) consigue arreglar la avería. Se sabe que el tiempo que esta persona utiliza en dicha reparación sigue una distribución exponencial con media de 5 minutos, mientras que la distribución del tiempo que un lector está funcionando hasta que se produce algún error de lectura es también exponencial pero con media de 1 hora. Calcular:

    (a) La probabilidad de que funcionen los tres lectores.
    (b) El número medio de lectores averiados.
    (c) El tiempo medio que un lector está averiado.
    (d) Si en un momento dado funcionan dos lectores, ¿cuál es el tiempo medio hasta la próxima avería?

12. Cinco estaciones de trabajo MOON se encuentran funcionando en un centro de cálculo. Por problemas con el sistema operativo Lunaris 3.2, las máquinas pueden quedar fuera de servicio temporalmente hasta que el técnico encargado vuelve a hacerlas funcionar. El tiempo que tarda una estación de trabajo en averiarse sigue una distribución exponencial con media de 10 días. El tiempo de reparación es también exponencial y con media de cinco horas. Las máquinas averiadas son atendidas según una disciplina FIFO. Calcular:

(a) La probabilidad de que alguna estación esté averiada.

(b) El número medio de estaciones funcionando.

(c) Cuando una máquina se avería, ¿cuántas horas pasan, por término medio, hasta que vuelve a funcionar?

(d) Si todas las estaciones están funcionando, ¿cuál es la probabilidad de que en el transcurso de las próximas 40 horas se averíe alguna?

13. Un computador de procesamiento en paralelo dispone de seis unidades de procesamiento idénticas. El tiempo que una de estas unidades permanece en funcionamiento hasta que se avería sigue una distribución exponencial de media 12 días. Las unidades averiadas son reparadas (por orden de avería) por un técnico, al cual le lleva un tiempo exponencial de media medio día realizar su tarea para cada unidad. Se pide:

(a) la probabilidad de que el técnico esté desocupado.

(b) el número medio de unidades de procesamiento en funcionamiento.

14. Por razones técnicas, una centralita con dos operadoras sólo permite mantener tres llamadas en espera (de tal forma que cualquier llamada producida cuando ya hay dos siendo atendidas por las operadoras y otras tres en espera, recibe el tono de "línea ocupada"). Las llamadas llegan según un proceso de Poisson, a razón de 6 por minuto, siendo 15 segundos la media del tiempo que tarda cada operadora en direccionar una llamada y dicho tiempo de distribución exponencial. Calcular:

(a) El porcentaje de tiempo en que cada operadora está ocupada.

(b) El número medio de llamadas en espera.

(c) La probabilidad de que una llamada obtenga la señal de "línea ocupada".

(d) Calcular las tres cantidades anteriores bajo el supuesto de que sólo hubiese una operadora.

15. Una factoría dispone de cuatro equipos de generación de corriente eléctrica que suministran gran parte de la energía que necesita dicha empresa. La distribución del tiempo que transcurre desde que un generador comienza a funcionar hasta que se avería es exponencial, con media de 40 días. El tiempo de reparación de un generador es una variable aleatoria de distribución exponencial y media 10 días. Sabiendo que existe un único técnico capaz de reparar los generadores, se pide:

(a) La probabilidad de que el técnico esté ocupado.

(b) El porcentaje medio de tiempo en el que todos los equipos de generación están averiados.

(c) El número medio de averías de equipos en un mes.

(d) El tiempo medio que transcurre desde la avería de un equipo hasta su reparación.

(e) El número medio de equipos funcionando.

16. Dar una fórmula, lo más sencilla posible, para el cálculo de la probabilidad de que cierto servidor esté ocupado en un modelo de colas $M/M/s$. Demostrar la validez de dicha expresión.

17. Una estación de trabajo con cuatro procesadores recibe procesos que son ejecutados en uno de los procesadores según una disciplina FIFO. La llegada de procesos obedece un proceso estocástico de Poisson con intensidad de 5 procesos por minuto. El tiempo de CPU de un proceso (tiempo durante el cual está procesándose, descontando el tiempo de espera previa) elegido al azar sigue una distribución exponencial con media de 30 segundos. Se pregunta:

    (a) ¿Es el proceso estacionario? ¿Por qué?

    (b) ¿Cuál es el número medio de procesos totales en la estación?

    (c) ¿Cuánto tiempo transcurre desde que se envía un proceso hasta que su ejecución termina (tiempo medio real)? Expresarlo como cociente entre el tiempo medio de CPU.

    (d) Calcular el número medio de procesadores ocupados.

    (e) ¿Cuál es la probabilidad de que, al enviar un proceso, éste comience a ser ejecutado sin demora?

18. A un cajero automático (con un único terminal) llegan usuarios según un proceso de Poisson a razón de 15 por minuto. El tiempo de uso del cajero, por parte de cada individuo, es una cantidad aleatoria con distribución exponencial y media de 3 minutos. Se pide:

    (a) ¿Es el sistema estacionario? ¿Por qué?

    (b) ¿Cuál es el número medio de usuarios haciendo cola fuera del cajero?

    (c) ¿Cuánto tiempo transcurre, por término medio, desde que un usuario se acerca al cajero hasta que termina su transacción en el mismo? Expresarlo en relación con el tiempo medio de uso del cajero por cada cliente.

    (d) Calcular la probabilidad de que el cajero esté desocupado.

    (e) ¿Cuánto tiempo pasará, por término medio, desde que un cliente llega al cajero hasta que termina de utilizarlo, sabiendo que en el momento de su llegada hay una persona haciendo uso del cajero y dos más esperando fuera?

19. La única línea telefónica de atención al cliente de una pequeña entidad recibe llamadas, según un proceso de Poisson, a razón de 5 por hora (en condiciones normales de trabajo). La duración de las llamadas es de distribución exponencial con una media de 2 minutos. Cuando se produce una llamada en el momento en que el operador está atendiendo otra, esta nueva llamada permanece en espera para ser atendida, salvo que ya exista otra llamada en espera, en cuyo caso obtiene el tono de línea ocupada. Calcular:

    (a) La probabilidad de que una llamada sea atendida sin demora.

    (b) La probabilidad de que una llamada necesite estar en espera.

    (c) La probabilidad de que una llamada se pierda por estar la línea ocupada.

    (d) ¿Cuánto valdría la probabilidad pedida en el apartado anterior si suponemos que estamos en un período punta correspondiente a 5 llamadas por minuto?

20. Para el sistema de una cola con llegadas y servicios exponenciales pero sin repuestos (modelos $M/M/s$, $M/M/s/K$, $M/M/s/\infty/H$ sin repuestos y $M/M/\infty$) encontrar una expresión que permita hallar el número medio de servidores ocupados en términos de L y L_q. En el caso concreto de un modelo $M/M/s/K$, expresar dicha cantidad en función del número de servidores y la intensidad de tráfico efectiva. Por último, a partir de la expresión encontrada, justificar que la probabilidad de bloqueo, p_{bloqueo}, de un sistema $M/M/s$ puede acotarse inferiormente mediante $p_{\text{bloqueo}} > 1 - \frac{1}{\rho}$.

12

Redes de colas

Se estudiarán aquí los modelos básicos de redes de colas abiertas y cerradas con distribución del tiempo de servicio y distribución del tiempo entre llegadas (si es el caso) exponencial. Además se impondrá la restricción de que los clientes que salen servidos de una de las colas que compone la red se mueven, instantáneamente y con ciertas probabilidades prefijadas, a cualquier otra posible cola de la red. Estos modelos dan lugar a las llamadas redes de Jackson (abiertas y cerradas). También se estudiarán en el capítulo otros modelos particulares de redes de colas como las colas en serie (sin y con bloqueo) y las redes circulares de colas.

12.1 Introducción a las redes de colas

Una red de colas no es más que una red en la que cada nodo está constituido por el sistema de una cola. Se trata, por tanto, de un grafo orientado en el que se pueden producir transiciones de clientes que salen servidos de un nodo (que es una cola) hacia otro nodo. La forma más habitual (aunque no la única) para modelizar el modo en que los clientes servidos en un nodo se dirigen a otro es considerando que lo hacen de acuerdo a una distribución de probabilidad discreta.

Al igual que en los modelos de una única cola, en las redes de colas también pueden producirse llegadas de clientes desde fuera del sistema (desde fuera de la red, en este caso) y salidas de clientes servidos hacia fuera de la red. A diferencia de aquél caso, en las redes de colas sí tiene sentido el plantear situaciones en las que no hay llegadas de clientes desde fuera de la red ni salidas hacia fuera de la red. Esto da lugar a las llamadas redes cerradas, que tienen la peculiaridad de que el número total de clientes en la red es fijo y lo único que desconocemos es dónde se encuentran (en qué nodos concretos) y en qué estado de servicio se hallan. Por su parte, las redes abiertas son aquellas en que sí se producen llegadas de clientes y salidas hacia fuera de la red.

Como se comentaba con anterioridad, denotando por $1, 2, \ldots, K$, los nodos que forman la red (es decir las etiquetas con las que denotamos cada cola) la manera más frecuente de modelizar las transiciones de clientes consiste en suponer que cuando un cliente sale servido de la cola del nodo i (siendo $i \in \{1, 2, \ldots, K\}$) se desplaza instantáneamente al sistema de la cola de cualquier otro nodo $j \in \{1, 2, \ldots, K\}$, con probabilidad p_{ij}. Evidentemente, en las colas abiertas también es posible que desde algunos nodos se pueda abandonar la red. Denotando con el índice 0 el exterior de la red, la probabilidad de que un

cliente abandone la red cuando sale servido del nodo i se denotará por p_{i0} y puede calcularse a partir de las anteriores mediante

$$p_{i0} = 1 - \sum_{j=1}^{K} p_{ij}.$$

Este tipo de esquema de transición de clientes se denotará por esquema de transiciones instantáneas aleatorias de clientes.

En una situación como la anterior, las probabilidades de transición de clientes de unos nodos a otros se pueden expresar de forma matricial:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots & p_{iK} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{Kj} & \cdots & p_{KK} \end{pmatrix}.$$

Esta es la llamada matriz de transición de la red.

Normalmente denotaremos por λ_i la tasa de entrada desde fuera del sistema a la cola del nodo $i \in \{1, 2, \ldots, K\}$. Asimismo μ_i denotará la tasa de servicio de cada uno de los servidores del subsistema del nodo i. De esta forma, bajo la hipótesis de que los tiempos entre llegadas desde fuera del sistema sean de distribución exponencial de parámetro λ_T y denotando por p_{0i} la probabilidad de que cuando un cliente entra al sistema lo haga a través del nodo i, se tiene que la distribución del tiempo entre dos entradas consecutivas de clientes al subsistema del nodo i es también exponencial y con parámetro $\lambda_i = \lambda_T \cdot p_{0i}$. Obviamente los valores λ_i también pueden interpretarse como el número medio de clientes que entran al sistema, por el nodo i, por unidad de tiempo. El número efectivo de llegadas de clientes al subsistema del nodo i (sean procedentes de fuera del sistema o de otro nodo del mismo) se denotará con la letra lambda mayúscula Λ_i. Para evitar la confusión entre las cantidades citadas en este párrafo y la tasas de llegadas y servicios en función del número de clientes que hay en el sistema (tradicionalmente denotadas por λ_n y μ_n), pasaremos a denotar estas últimas por $\Lambda_i^{(n)}$ y $\mu_i^{(n)}$, queriendo expresar éstas las tasas de llegadas y servicios al subsistema del nodo i, cuando hay n clientes en dicho subsistema.

12.2 Redes de Jackson abiertas

Una red de Jackson abierta no es más que una red de colas abierta (es decir, en la que hay llegadas de clientes desde fuera de la red y salidas de clientes a fuera de la red) que verifica las tres propiedades siguientes:

1. Cada nodo $i = 1, 2, \ldots, K$ tiene un mecanismo de servicio consistente en s_i servidores con tiempo de servicio de idéntica distribución, exponencial de parámetro μ_i.

2. Los clientes que llegan al nodo i desde fuera del sistema lo hacen según un proceso de Poisson de intensidad λ_i. Esto equivale a decir que los tiempos entre dos llegadas de clientes consecutivos desde fuera del sistema al nodo i, siguen una distribución exponencial de dicho parámetro.

3. El flujo de clientes sigue el esquema de transiciones instantáneas aleatorias. Es decir, cada cliente que sale servido del mecanismo de servicio del nodo i va instantáneamente a cualquier otro nodo j con probabilidad p_{ij}, o bien sinó sale del sistema (lo cual ocurre con probabilidad $p_{i0} = 1 - \sum_{j=1}^{K} p_{ij}$).

La forma de poder encontrar las distintas cantidades de interés en una red de Jackson abierta viene dada por el Teorema de Jackson que pasamos a enunciar.

Teorema 51 (Teorema de Jackson) *Consideremos una red de Jackson abierta con K nodos. Entonces, las tasas de llegada a cada nodo j de la red (Λ_j) verifican:*

$$\Lambda_j = \lambda_j + \sum_{i=1}^{K} \Lambda_i p_{ij}, \text{ para } j = 1, 2, \ldots K.$$

Además, denotando por $p(n_1, n_2, \cdots, n_K)$ la probabilidad de que haya exactamente n_1 clientes en el nodo 1, n_2 clientes en el nodo 2, ... , n_K clientes en el nodo K, y si $\Lambda_i < s_i \mu_i$, para $i = 1, 2, \ldots, K$, entonces se tiene:

$$p(n_1, n_2, \cdots, n_K) = p_1(n_1) \cdot p_2(n_2) \cdot \cdots \cdot p_K(n_K),$$

siendo $p_i(n_i)$ la probabilidad de que haya n_i clientes en el nodo i, calculada como si el nodo i se tratase de un modelo $M/M/s_i$ aislado, con tasa de llegada Λ_i y tasa de servicio μ_i. Más aún, el subsistema de cada nodo se comporta, en general, como si fuese un sistema aislado correspondiente a un modelo $M/M/s_i$ independiente de los demás.

Como consecuencia del Teorema de Jackson resulta fácil obtener las principales cantidades acerca de cada subsistema. Por analogía con la notación anterior, denotaremos esas cantidades por L_i, $L_{q,i}$, W_i y $W_{q,i}$ cuando se refieran al subsistema del nodo i. Para hacer mención a las correspondientes cantidades sobre todo el sistema (toda la red) usaremos la notación L_T, $L_{q,T}$, W_T y $W_{q,T}$.

Los pasos a seguir para resolver cualquier red abierta de Jackson serán:

1. Una vez establecidas la tasa de entrada desde el exterior a cada subsistema (λ_i), la tasa de servicio (μ_i), el número de servidores (s_i) y las probabilidades de transición (p_{ij}), plantear y resolver el sistema de ecuaciones:

$$\Lambda_j = \lambda_j + \sum_{i=1}^{K} \Lambda_i p_{ij}, \text{ para } j = 1, 2, \dots K,$$

o, equivalentemente, en forma matricial

$$\overrightarrow{\Lambda} = \overrightarrow{\lambda} + P^t \cdot \overrightarrow{\Lambda}, \text{ o simplemente, } \left(I - P^t\right) \overrightarrow{\Lambda} = \overrightarrow{\lambda},$$

siendo

$$\overrightarrow{\Lambda} = \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \\ \vdots \\ \Lambda_K \end{pmatrix} \text{ y } \overrightarrow{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_K \end{pmatrix}.$$

La solución $\overrightarrow{\Lambda} = (I - P^t)^{-1} \cdot \overrightarrow{\lambda}$ nos proporciona las tasas totales de llegada a cada subsistema (vengan de fuera o de otro nodo).

2. Utilizando los valores encontrados en el paso anterior pueden hallarse todos los valores relevantes de cada subsistema (L_i, $L_{q,i}$, W_i y $W_{q,i}$), sin más que considerarlo como un modelo $M/M/s_i$, por separado, con tasa de llegada Λ_i y tasa de servicio μ_i.

3. Debido a la naturaleza del sistema (una red de colas) y a la hipótesis de transiciones instantáneas, los parámetros de número medio de clientes en toda la red y número de clientes en cola en toda la red pueden cacularse sumando simplemente los de cada nodo:

$$L_T = \sum_{i=1}^{K} L_i, \quad \text{y} \quad L_{q,T} = \sum_{i=1}^{K} L_{q,i}.$$

Por su parte, los tiempos medios de permanencia en el sistema de toda la red y de permanencia en la cola para toda la red non son de cálculo tan directo, pues la trayectoria de un cliente a lo largo de la red de colas es aleatoria. Estamos pues ante una situación en la que el tiempo de un cliente en la red está sujeto a dos tipo de aleatoriedad. En primer lugar desconocemos qué nodos visitará (algunos quizá repetidamente) el cliente, pues su evolución en la red es aleatoria. En segundo lugar, aún conociendo cuáles han sido los nodos por los que ha pasado (y el número de veces que ha pasado por cada uno) es aleatorio el tiempo que va a estar en cada subsistema (aunque en este caso sí es conocida su media: W_i para cada vez que pasa por el nodo i). La forma más sencilla de resolver el

problema es utilizar las fórmulas de Little generalizadas al sistema de toda la red. Para ello definimos $\lambda_T = \sum_{i=1}^{K} \lambda_i$, que es el número medio de clientes que entran en la red por unidad de tiempo (por cualquier nodo) y obtenemos

$$\begin{aligned} W_T &= \frac{L_T}{\lambda_T}, \\ W_{q,T} &= \frac{L_{q,T}}{\lambda_T}. \end{aligned}$$

Veamos un ejemplo.

Ejemplo 52 *Dos servidores informáticos reciben procesos de usuarios a razón de 20 procesos por minuto el primero y 30 procesos por minuto el segundo. La llegada de procesos obedece un proceso de Poisson. Dada su mayor rapidez, el único procesador del servidor 1 puede atender una media de 100 procesos por minuto mientras que cualquiera de los dos procesadores del 2 sólo es capaz de atender hasta 25, siendo los tiempos de procesado exponenciales. Se sabe que ambos servidores atienden los procesos según una disciplina FIFO y que cuando un proceso está a punto de acabar su ejecución en el servidor 2 crea otro nuevo proceso hijo en el 1 con una probabilidad del 25%, terminando totalmente la ejecución en otro caso. Por su parte, los procesos que están terminando su ejecución en el nodo 1 crean otro nuevo en el mismo servidor en un 20% de los casos, mandan otro proceso al servidor 2 un 10% de las veces, terminando totalmente la ejecución en caso contrario. Calcular el número medio de procesos en cada servidor y el tiempo medio que tarda un proceso y toda su descendencia en terminar de ejecutarse.*

Solución: Se trata de una red de Jackson abierta con $K = 2$ nodos, para la cual hay $s_1 = 1$ servidor en el nodo 1 y $s_2 = 2$ servidores en el segundo. Las tasas de llegada y servicio (en número de procesos por minuto) desde fuera del sistema son $\lambda_1 = 20$, $\lambda_2 = 30$, $\mu_1 = 100$, $\mu_2 = 25$. Las probabilidades de transición de unos estados a otros vienen dadas, en forma matricial, por

$$P = \begin{pmatrix} 0.2 & 0.1 \\ 0.25 & 0 \end{pmatrix}.$$

Planteamos y resolvemos el sistema que permite encontrar las tasas efectivas:

$$\begin{aligned} \Lambda_1 &= 20 + 0.2 \cdot \Lambda_1 + 0.25 \cdot \Lambda_2 \\ \Lambda_2 &= 30 + 0.1 \cdot \Lambda_1 \end{aligned}$$

luego

$$\begin{aligned} \Lambda_1 &= 20 + 0.2\Lambda_1 + 0.25\,(30 + 0.1\Lambda_1) = 20 + 0.2\Lambda_1 + 7.5 + 0.025\Lambda_1 \\ &= 27.5 + 0.225\Lambda_1, \end{aligned}$$

con lo cual $0.775\Lambda_1 = 27.5$ *y, por tanto,* $\Lambda_1 = \frac{27.5}{0.775} = 35.484$. *Por otra parte* $\Lambda_2 = 30 + 0.1\Lambda_1 = 33.548$. *Primeramente debemos encontrar los valores* L_1 *y* L_2 *utilizando las fórmulas disponibles para el modelo* $M/M/1$ *con* $\Lambda_1 = 35.484$ *y* $\mu_1 = 100$ *y para el modelo* $M/M/2$ *con* $\Lambda_2 = 33.548$ *y* $\mu_2 = 25$. *Cómo* $\rho_1 = \frac{35.484}{100} = 0.35484 < 1$ *y* $\rho_2 = \frac{33.548}{2 \cdot 25} = 0.67096 < 1$, *entonces ambos nodos son estacionarios. Por tanto, se tiene que*

$$L_1 = \frac{\rho_1}{1-\rho_1} = \frac{0.35484}{1-0.35484} = 0.55.$$

Por su parte, para el nodo 2,

$$\begin{aligned} p_2(0) &= \frac{1}{1 + \sum_{n=1}^{2-1} \frac{\Lambda_2^n}{n!\mu_2^n} + \frac{\Lambda_2^2}{(2-1)!\mu_2^{2-1}(2\mu_2 - \Lambda_2)}} \\ &= \frac{1}{1 + \frac{33.548}{25} + \frac{33.548^2}{25(2 \cdot 25 - 33.548)}} = 0.19691674 \end{aligned}$$

y

$$L_2 = \frac{\Lambda_2^{2+1} p_2(0)}{(2-1)!\mu_2^{2-1}(2\mu_2 - \Lambda_2)^2} + \frac{\Lambda_2}{\mu_2} = \frac{33.548^3 \cdot 0.19691674}{25(2 \cdot 25 - 33.548)^2} + \frac{33.548}{25} = 2.4407$$

Así pues hay, por término medio, 0.55 *procesos (en ejecución o en cola) en el servidor 1 y* 2.44 *en el servidor 2. Como consecuencia tendremos* $L_T = L_1 + L_2 = 2.9907$ *procesos en media en todo el sistema y, usando la primera fórmula de Little para toda la red se obtiene*

$$W_T = \frac{L_T}{\lambda_T} = \frac{2.9907}{50} = 0.0598 \text{ minutos.}$$

Es decir, el tiempo medio que transcurre desde que un proceso se envía al sistema informático (sin prejuzgar por qué servidor entra) hasta que él y toda su descendencia está ejecutada es 0.0598 *minutos* $= 3.588$ *segundos.*

12.3 Redes de Jackson cerradas

Una red de Jackson cerrada es una red de colas cerrada con K nodos o subsistemas, en la cual cada $i = 1, 2, \ldots, K$ tiene s_i servidores en su mecanismo de servicio, siendo todos los del nodo i con tiempo de servicio de distribución exponencial de parámetro μ_i. Como en cualquier red cerrada, al no haber entradas ni salidas de clientes de la red, resulta indispensable especificar el número de clientes dentro de la red, N, que permanecerá constante siempre. Por este motivo, $L_T = N$ y cantidades como W_T o $W_{q,T}$, carecen de sentido. Lo realmente

importante aquí es determinar las probabilidades de que haya n_i clientes en el nodo i para $i = 1, 2, \ldots, K$, que se denotarán por $p_{n_1,n_2,\cdots,n_K}$. Obviamente éstas probabilidades son distintas de cero sólo si $n_1 + n_2 + \cdots + n_K = N$.

Las probabilidades de los distintos estados de la red pueden calcularse por medio de la siguiente fórmula:

$$p_{n_1,n_2,\cdots,n_K} = \frac{1}{G(N)} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{a_i(n_i)},$$

donde

$$a_i(n) = \begin{cases} n! & \text{si} \quad n \leq s_i \\ s_i! s_i^{n-s_i} & \text{si} \quad n \geq s_i \end{cases}$$

y

$$G(N) = \sum_{n_1+n_2+\cdots+n_K=N} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{a_i(n_i)},$$

siendo $\rho_i = \frac{\Lambda_i}{\mu_i}$ y los Λ_i cualquier solución no nula del sistema

$$\Lambda_j = \sum_{i=1}^{K} \Lambda_i p_{ij} \quad \text{para } j = 1, 2, \ldots, K.$$

Este último sistema es el que se obtiene en el teorema de Jackson cuando las tasas de entrada λ_i son cero. Lo que ocurre ahora es que dicho sistema no tiene solución única (ya que se trata de un sistema homogéneo). De hecho, para encontrar una solución del sistema anterior basta fijar al valor 1 cualquiera de las Λ_i (por ejemplo $\Lambda_1 = 1$) y resolver el sistema ignorando una cualquiera de las ecuaciones (combinación lineal del resto).

La implementación directa de $G(N)$ según la fórmula anterior puede resultar tremendamente ineficiente cuando N es grande ya que el número de estados distintos del sistema (que viene coincidiendo con el número de sumandos en la expresión de $G(N)$) es $\binom{N+K-1}{N}$. Para evitar dicho cálculo directo existe una fórmula recursiva que permite la obtención de

$$g_m(n) = \sum_{n_1+n_2+\cdots+n_m=n} \prod_{i=1}^{m} f_i(n_i)$$

siendo

$$f_i(n) = \frac{\rho_i^n}{a_i(n)}, \text{ para } i = 1, 2, \ldots, K \text{ y } n = 0, 1, 2, \ldots, N.$$

Obsérvese que esto resuelve el problema, pues $G(N) = g_K(N)$.

La recursividad citada se obtiene fácilmente teniendo en cuenta:

$$\begin{aligned} g_m(n) &= \sum_{n_1+n_2+\cdots+n_m=n} \prod_{i=1}^{m} f_i(n_i) \\ &= \sum_{n_m=0}^{n} \left[\sum_{n_1+n_2+\cdots+n_{m-1}=n-n_m} \prod_{i=1}^{m-1} f_i(n_i) \right] f_m(n_m) \\ &= \sum_{n_m=0}^{n} f_m(n_m)\, g_{m-1}(n-n_m) = \sum_{i=0}^{n} f_m(i)\, g_{m-1}(n-i)\,. \end{aligned}$$

Por otra parte puede verse trivialmente que $g_1(n) = f_1(n)$, lo cual permite comenzar a aplicar la ecuación recursiva. Además, resulta inmediato comprobar que $g_m(0) = 1$ para todo $m = 1, 2, \ldots, K$.

Ejemplo 53 *Un sistema informático consta de 4 estaciones de trabajo conectadas entre sí. Para control y seguridad del sistema hay tres procesos en continua ejecución en alguna de las estaciones. Una vez terminada la ejecución de un proceso en una de las estaciones éste crea una copia de él mismo que se envía a ejecutar a la propia estación o a alguna de las otras tres. En la siguiente tabla se recogen las probabilidades de que el proceso "hijo" se envíe a la estación j, sabiendo que el que termina su ejecución lo hace en la estación i:*

origen\destino	*1*	*2*	*3*	*4*
1	0.25	0.15	0.20	0.40
2	0.15	0.35	0.20	0.30
3	0.50	0.25	0.15	0.10
4	0.40	0.30	0.25	0.05

Se sabe que los servidores 1 y 2 son biprocesadores y cada uno de los procesadores de ambos tiene un tiempo de procesado con distribución exponencial y capacidad de 5 procesos por minuto. Por su parte, las estaciones 3 y 4 son ambas monoprocesadoras y pueden atender respectivamente a 10 y 15 procesos por minuto. Calcular la probabilidad de que esté alguno de los tres procesos en el servidor 4. ¿Cuál es el número medio de procesos en el servidor 4? ¿Cuál es el tiempo medio que transcurre desde que llega un proceso al servidor 4 hasta que termina su ejecución?

Solución: Teniendo en cuenta los datos del ejemplo, se puede modelizar mediante una red de Jackson cerrada con $N = 3$, $K = 4$, $\mu_1 = 5$, $\mu_2 = 5$, $\mu_3 = 10$ y $\mu_4 = 15$. Los clientes son pues cada uno de los tres procesos que están "recorriendo" el sistema informático. Dada la matriz P, en primer lugar trataremos de encontrar los Λ_i. Para ello debemos hallar una solución del

sistema:

$$\begin{aligned}
\Lambda_1 &= \Lambda_1 p_{11} + \Lambda_2 p_{21} + \Lambda_3 p_{31} + \Lambda_4 p_{41} \\
\Lambda_2 &= \Lambda_1 p_{12} + \Lambda_2 p_{22} + \Lambda_3 p_{32} + \Lambda_4 p_{42} \\
\Lambda_3 &= \Lambda_1 p_{13} + \Lambda_2 p_{23} + \Lambda_3 p_{33} + \Lambda_4 p_{43} \\
\Lambda_4 &= \Lambda_1 p_{14} + \Lambda_2 p_{24} + \Lambda_3 p_{34} + \Lambda_4 p_{44}
\end{aligned}$$

que en nuestro caso se reduce a

$$\begin{aligned}
\Lambda_1 &= 0.25\Lambda_1 + 0.15\Lambda_2 + 0.50\Lambda_3 + 0.40\Lambda_4 \\
\Lambda_2 &= 0.15\Lambda_1 + 0.35\Lambda_2 + 0.25\Lambda_3 + 0.30\Lambda_4 \\
\Lambda_3 &= 0.20\Lambda_1 + 0.20\Lambda_2 + 0.15\Lambda_3 + 0.25\Lambda_4 \\
\Lambda_4 &= 0.40\Lambda_1 + 0.30\Lambda_2 + 0.10\Lambda_3 + 0.05\Lambda_4
\end{aligned}$$

Tomando, por ejemplo, $\Lambda_3 = 1$*, se obtiene una de las infinitas soluciones del sistema homogéneo:* $\Lambda_1 = 1.536286$*,* $\Lambda_2 = 1.2716115$*,* $\Lambda_3 = 1$ *y* $\Lambda_4 = 1.153682$*. De esta forma, teniendo en cuenta las capacidades de servicio de cada procesador de las distintas estaciones, se obtienen los valores:*

$$\begin{aligned}
\rho_1 &= \frac{1.536286}{5} = 0.3072572 \\
\rho_2 &= \frac{1.2716115}{5} = 0.2543223 \\
\rho_3 &= \frac{1}{10} = 0.1 \\
\rho_4 &= \frac{1.153682}{15} = 0.0769121
\end{aligned}$$

Por otra parte,

$$a_1(n) = a_2(n) = 2^{n-1} \text{ y } a_3(n) = a_4(n) = 1 \ \forall n,$$

con lo cual

$$\begin{aligned}
f_i(n) &= 2^{1-n}\rho_i^n, \text{ para } i = 1,2 \text{ y } f_i(n) = \rho_i^n, \\
\text{para } i &= 3,4, \text{ siempre que } n \geq 1, \\
\text{mientras que } f_i(0) &= 1 \text{ para } i = 1,2,3,4.
\end{aligned}$$

En particular

$$\begin{aligned}
g_1(0) &= f_1(0) = 1,\ g_1(1) = f_1(1) = 0.3072572, \\
g_1(2) &= f_1(2) = \frac{(0.3072572)^2}{2} = 4.7203493 \times 10^{-2}, \\
g_1(3) &= f_1(3) = \frac{(0.3072572)^3}{2^2} = 7.2518066 \times 10^{-3}.
\end{aligned}$$

Ahora procedemos recursivamente, encontrando los valores $g_2(n)$*:*

$$\begin{aligned}
g_2(0) &= 1,\ g_2(1) = f_2(0)\,g_1(1) + f_2(1)\,g_1(0)\\
&= 0.3072572 + 0.2543223 = 0.5615795,\\
g_2(2) &= f_2(0)\,g_1(2) + f_2(1)\,g_1(1) + f_2(2)\,g_1(0)\\
&= 4.7203493 \times 10^{-2} + 0.2543223 \cdot 0.3072572 + \frac{(0.2543223)^2}{2}\\
&= 0.15768577,\\
g_2(3) &= f_2(0)\,g_1(3) + f_2(1)\,g_1(2) + f_2(2)\,g_1(1) + f_2(3)\,g_1(0)\\
&= 7.2518066 \times 10^{-3} + 0.2543223 \cdot 4.7203493 \times 10^{-2}\\
&\quad + \frac{(0.2543223)^2}{2} \cdot 0.3072572 + \frac{(0.2543223)^3}{2^2}\\
&= 3.3305761 \times 10^{-2},
\end{aligned}$$

luego, los valores $g_3(n)$*:*

$$\begin{aligned}
g_3(0) &= 1,\ g_3(1) = f_3(0)\,g_2(1) + f_3(1)\,g_2(0)\\
&= 0.5615795 + 0.1 = 0.6615795,\\
g_3(2) &= f_3(0)\,g_2(2) + f_3(1)\,g_2(1) + f_3(2)\,g_2(0)\\
&= 0.15768577 + 0.1 \cdot 0.5615795 + (0.1)^2\\
&= 0.22384372,\\
g_3(3) &= f_3(0)\,g_2(3) + f_3(1)\,g_2(2) + f_3(2)\,g_2(1) + f_3(3)\,g_2(0)\\
&= 3.3305761 \times 10^{-2} + 0.1 \cdot 0.15768577 + (0.1)^2 \cdot 0.5615795 + (0.1)^3\\
&= 5.5690133 \times 10^{-2}.
\end{aligned}$$

Finalmente,

$$\begin{aligned}
G(3) &= g_4(3) = f_4(0)\,g_3(3) + f_4(1)\,g_3(2) + f_4(2)\,g_3(1) + f_4(3)\,g_3(0)\\
&= 5.5690133 \times 10^{-2} + 0.0769121 \cdot 0.22384372\\
&\quad + (0.0769121)^2 \cdot 0.6615795 + (0.0769121)^3\\
&= 7.7274949 \times 10^{-2}.
\end{aligned}$$

El número medio de procesos en el servidor 4 resulta

$$\begin{aligned}
L_4 &= 0 \cdot p_{\bullet\bullet\bullet 0} + 1 \cdot p_{\bullet\bullet\bullet 1} + 2 \cdot p_{\bullet\bullet\bullet 2} + 3 \cdot p_{\bullet\bullet\bullet 3}\\
&= p_{\bullet\bullet\bullet 1} + 2 \cdot p_{\bullet\bullet\bullet 2} + 3 \cdot p_{\bullet\bullet\bullet 3},\ \textit{siendo}\\
p_{\bullet\bullet\bullet n_4} &= \sum_{n_1,n_2,n_3,n_1+n_2+n_3=N-n_4} p_{n_1,n_2,n_3,n_4},\ \textit{es decir}\\
p_{\bullet\bullet\bullet 1} &= p_{2001} + p_{1101} + p_{1011} + p_{0201} + p_{0111} + p_{0021},\\
p_{\bullet\bullet\bullet 2} &= p_{1002} + p_{0102} + p_{0012},\\
p_{\bullet\bullet\bullet 3} &= p_{0003}.
\end{aligned}$$

En lugar de calcular estos 10 términos, podemos tener en cuenta la forma de cada uno de ellos:

$$p_{n_1,n_2,\cdots,n_K} = \frac{1}{G(N)} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{a_i(n_i)}$$

para deducir

$$\begin{aligned} p_{\bullet,\bullet,\cdots,\bullet,n_K} &= \sum_{n_1,n_2,\ldots,n_{K-1},n_1+n_2+\cdots+n_{K-1}=N-n_K} p_{n_1,n_2,\cdots,n_K} \\ &= \frac{1}{G(N)} \frac{\rho_K^{n_K}}{a_K(n_K)} \sum_{n_1,n_2,\ldots,n_{K-1},n_1+n_2+\cdots+n_{K-1}=N-n_K} \prod_{i=1}^{K-1} \frac{\rho_i^{n_i}}{a_i(n_i)} \\ &= \frac{1}{G(N)} f_K(n_K)\, g_{K-1}(N-n_K). \end{aligned}$$

Salvo el término $G(N)^{-1}$ lo restante es uno de los sumandos que aparece en el cálculo de $G(N) = g_K(N)$. Es digno de mención que para obtener la eficiencia de cálculo de este procedimiento (frente al de obtener las 10 probabilidades antes planteadas) resulta de suma importancia que en los cálculos abreviados previos se haya dejado de último el nodo 4 (para así poder usar $g_3(n)$). En caso de que se deseasen calcular estas probabilidades marginales para otro nodo, resultaría más adecuado renombrar los nodos de forma que el nodo de interés fuese el último del cálculo abreviado, ya que el ahorro está basado en la fórmula recursiva de cálculo de los $g_m(n)$. En nuestro caso, resulta muy fácil calcular las probabilidades marginales relativas al nodo 4:

$$\begin{aligned} p_{\bullet\bullet\bullet 0} &= \frac{f_4(0)\, g_3(3)}{G(3)} = \frac{5.5690133 \times 10^{-2}}{7.7274949 \times 10^{-2}} = 0.72067512 \\ p_{\bullet\bullet\bullet 1} &= \frac{f_4(1)\, g_3(2)}{G(3)} = \frac{0.0769121 \cdot 0.2238472}{7.7274949 \times 10^{-2}} = 0.22279265, \\ p_{\bullet\bullet\bullet 2} &= \frac{f_4(2)\, g_3(1)}{G(3)} = \frac{(0.0769121)^2 \cdot 0.6615795}{7.7274949 \times 10^{-2}} = 5.0644542 \times 10^{-2}, \\ p_{\bullet\bullet\bullet 3} &= \frac{f_4(3)\, g_3(0)}{G(3)} = \frac{(0.0769121)^3}{7.7274949 \times 10^{-2}} = 5.8876947 \times 10^{-3}. \end{aligned}$$

Así pues, la probabilidad de que esté algún proceso en el servidor 4 es

$$1 - p_{\bullet\bullet\bullet 0} = 1 - 0.72067512 = 0.27932488.$$

Además, es inmediato el cálculo del número medio de procesos en la estación 4:

$$\begin{aligned} L_4 &= 0.22279265 + 2 \cdot 5.0644542 \times 10^{-2} + 3 \cdot 5.8876947 \times 10^{-3} \\ &= 0.34174482. \end{aligned}$$

Para calcular el tiempo medio que pasa un proceso en el servidor número 4 podemos utilizar las fórmulas de Little, si bien, debemos calcular previamente el valor Λ_4 adecuado. Anteriormente ya se había encontrado una de las infinitas soluciones no nulas del sistema de ecuaciones que relaciona las tasas de entrada. El problema es que esa soulción encontrada no tiene porqué ofrecer el valor correcto de las Λ_i (sinó más bien valores que son proporcionales a los $\overline{\Lambda}_i$ verdaderos). El problema puede resolverse fácilmente utilizando que el número medio de clientes que entran a un nodo elegido (Λ_i) ha de ser igual al número medio de clientes que salen servidos de dicho nodo. Aplicando esta condición al nodo 4 obtenemos:

$$\begin{aligned}\overline{\Lambda}_4 &= 0 \cdot p_{\bullet\bullet\bullet 0} + \mu_4 \cdot (p_{\bullet\bullet\bullet 1} + p_{\bullet\bullet\bullet 2} + p_{\bullet\bullet\bullet 4}) \\ &= 15 \cdot \left(0.22279265 + 5.0644542 \times 10^{-2} + 5.8876947 \times 10^{-3}\right) \\ &= 15 \cdot 0.27932489 = 4.1898734.\end{aligned}$$

Aunque en este caso no resulta de nuestro interés, esto permitiría encontrar las tasas de llegada al resto de los nodos de la red, pues ya comentamos que

$$\begin{aligned}\overline{\Lambda}_i &= c\Lambda_i, \text{ para } i = 1,2,3,4 \text{ y cierto } c > 0, \text{ con lo cual} \\ c &= \frac{\overline{\Lambda}_4}{\Lambda_4} = \frac{4.1898734}{1.153682} = 3.6317403 \text{ y así} \\ \overline{\Lambda}_1 &= 3.6317403 \cdot 1.536286 = 5.5793918, \\ \overline{\Lambda}_2 &= 3.6317403 \cdot 1.2716115 = 4.6181627, \\ \overline{\Lambda}_3 &= 3.6317403 \cdot 1 = 3.6317403.\end{aligned}$$

Finalmente, la fórmula de Little generalizada aplicada al nodo 4 permite calcular el tiempo medio de permanencia de un proceso en dicho nodo:

$$W_4 = \frac{L_4}{\overline{\Lambda}_4} = \frac{0.34174482}{4.1898734} = 0.081564 \text{ minutos } = 4.89 \text{ segundos.}$$

12.4 Otros modelos de colas en red

Como casos particulares de redes de colas estudiaremos las colas en serie y las redes circulares de colas.

12.4.1 Colas en serie

Se trata de una colección de K colas que se suceden unas a otras de tal manera que sólo es posible la entrada de clientes desde fuera del sistema a la primera

de ellas, produciéndose la salida de ellos tras el servicio del último nodo. Con nuestra terminología anterior, se tiene que

$$\begin{aligned}
\lambda_1 &= \lambda > 0,\ \lambda_i = 0\ \forall i = 2, 3, \ldots, K, \\
p_{ij} &= \begin{cases} 1 & \text{si} \quad j = i+1 \\ 0 & \text{si} \quad j \neq i+1 \end{cases} \\
& \quad \text{De lo que se deduce que} \\
p_{01} &= 1,\ p_{K0} = 1,\ p_{0j} = 0 \text{ para } j = 2, 3, \ldots, K, \\
p_{i0} &= 0 \text{ para } i = 1, 2, \ldots, K-1.
\end{aligned}$$

Además de todo esto, el mecanismo de servicio del nodo i está compuesto por s_i servidores con idéntica capacidad de servicio a razón de μ_i clientes por unidad de tiempo (con duración de servicio exponencial) y la capacidad de la cola es ilimitada.

Se trata, por tanto, de una red de Jackson abierta muy particular, a la cuál le es aplicable el Teorema de Jackson. Las ecuaciones que permiten encontrar las tasas efectivas de llegada se reducen a:

$$\begin{aligned}
\Lambda_1 &= \lambda \\
\Lambda_2 &= \Lambda_1 \\
&\vdots \\
\Lambda_K &= \Lambda_{K-1}
\end{aligned}$$

con lo cual

$$\Lambda_1 = \Lambda_2 = \cdots = \Lambda_K = \lambda,$$

de manera que se pueden resolver los modelos de cada nodo como si se tratase de subsistemas de colas $M/M/s_i$ independientes con tasa de entrada λ y tasa de servicio μ_i. Dado que en este caso el flujo de un cliente a través de la red es secuencial, desde el nodo 1 hasta el nodo K, sí será cierto que los tiempos medios de un cliente en el sistema o en la cola son suma de los correspondientes a cada subsistema. Matemáticamente esto se sustenta en que

$$W_T = \frac{L_T}{\lambda_T} = \frac{\sum_{i=1}^{K} L_i}{\lambda} = \sum_{i=1}^{K} \frac{L_i}{\lambda} = \sum_{i=1}^{K} \frac{L_i}{\Lambda_i} = \sum_{i=1}^{K} W_i.$$

12.4.2 Red de colas cíclica

En este caso se tienen K subsistemas de colas, tipo $M/M/s_i$, como en el caso anterior, pero conectadas de forma circular. Se trata, por tanto, de una red de Jackson cerrada en la que cada cliente que sale servido del nodo i (con $i = 1, 2, \ldots, K-1$) se dirige al nodo $i+1$ con probabilidad 1. Si el cliente sale

servido del nodo K entonces va al nodo 1 con toda probabilidad. Es necesario especificar el número total de clientes en la red, N, y a partir de las tasas de servicio, μ_i, pueden encontrarse las distintas probabilidades de interés. Así, el sistema de ecuaciones lineales homogéneo resulta:

$$\begin{aligned} \Lambda_1 &= \Lambda_K \\ \Lambda_2 &= \Lambda_1 \\ &\vdots \\ \Lambda_K &= \Lambda_{K-1} \end{aligned}$$

lo cual implica que

$$\Lambda_1 = \Lambda_2 = \cdots = \Lambda_{K-1} = \Lambda_K = c,$$

siendo c una constante positiva arbitrariamente elegida. De esto se deduce que $\rho_i = \frac{c}{\mu_i}$ y puede procederse tal y como ya se detalló para una red de Jackson cerrada genérica.

En el caso particular en el que haya un sólo servidor en cada nodo ($s_i = 1$ para $i = 1, 2, \ldots, K$) y las tasas de servicio de todos los nodos coincidan ($\mu_1 = \mu_2 = \cdots = \mu_K = \mu$), eligiendo la constante $c = \mu$ se tendría que todos los ρ_i valen 1. Como $a_i(n) = 1$ para todo $i = 1, 2, \ldots, K$ y todo $n \geq 0$, resulta muy fácil obtener las probabilidades de los distintos estados del sistema, ya que todas son iguales:

$$p_{n_1, n_2, \cdots, n_K} = \frac{1}{G(N)}$$

donde

$$G(N) = \sum_{n_1+n_2+\cdots+n_K=N} 1 = \binom{N+K-1}{N}.$$

En un caso como ese, también es muy fácil obtener las probabilidades marginales, pongamos por caso del nodo 1:

$$\begin{aligned} p_{n_1, \bullet, \cdots, \bullet} &= \sum_{n_2+n_3+\cdots+n_K=N-n_1} p_{n_1, n_2, \cdots, n_K} = \frac{1}{G(N)} \sum_{n_2+n_3+\cdots+n_K=N-n_1} 1 \\ &= \frac{\binom{N-n_1+K-2}{N-n_1}}{\binom{N+K-1}{N}}, \text{ para } n_1 = 0, 1, \ldots, N. \end{aligned}$$

12.4.3 Colas en serie con bloqueo

En algunas situaciones las redes de colas que se examinan tiene capacidad limitada para algunas o todas las colas. En tal caso resulta indispensable especificar claramente qué consecuencias tendrá para el futuro flujo de clientes

que una de esas colas se encuentre saturada. Típicamente este tipo de situaciones suelen modelizarse mediante los llamados "bloqueos" que consisten en que cualquier cliente que fuese a salir servido de otro subsistema que compone la red pero que se encuentra su cola de destino saturada, mantiene bloqueado el mecanismo de servicio sin que otro cliente pueda acceder a él. En este tipo de modelos y en muchísimos otros resulta perfectamente posible plantear desde el principio las ecuaciones de balance de la red y resolverlas, obteniendo así las probabilidades de los diferentes estados del sistema. Veamos un ejemplo.

Considérese un sistema formado por dos colas de tipo $M/M/1/0$ en serie. Denótese por λ la tasa de llegada de clientes por unidad de tiempo desde fuera del sistema al primer nodo (no hay entradas desde fuera al segundo) y por μ_1 y μ_2 las tasas de servicio del único servidor del nodo 1 y 2, respectivamente. Dado que no está permitido hacer cola en ninguno de los dos subsistemas, cuando a un cliente se le termine el servicio en el primero de ellos y encuentre que el servidor del segundo subsistema está ocupado, permanecerá bloqueando el servidor del primero. De esta forma denotando por (i, j) un estado genérico del sistema, con el que se indica que hay i clientes en el nodo 1 y j clientes en el nodo 2, tendríamos los estados:

$$(0,0)\,,(0,1)\,,(1,0)\,,(1,1)\,.$$

Como además es posible que haya un cliente en el primer nodo, ya servido, pero bloqueando el único servidor de dicho nodo (lo cual ocurre cuando, a su vez, haya un cliente sirviéndose en el segundo nodo) entonces tendremos un estado más que denotaremos por $(b, 1)$. Por lo comentado anteriormente, se ve que $(b, 1)$ es un estado diferente de $(1, 1)$.

Aunque podríamos proceder paso a paso para la obtención de las ecuaciones de balance (tal y como se hizo con un proceso de nacimiento y muerte estacionario arbitrario) resulta más sencillo e intuitivo deducirlas directamente de la idea de igualar, para cada posible estado, la probabilidad de dicho estado multiplicada por la suma de las tasas correspondientes a formas de salir de este estado hacia otro distinto con la suma de la probabilidades de aquellos estados desde los cuales se puede llegar al estado en cuestión (en un única transición), multiplicadas por las tasas correspondientes a dicha transición. Así se obtienen

$$\begin{aligned}
\lambda p_{00} &= \mu_2 p_{01} \\
(\lambda + \mu_2)\, p_{01} &= \mu_1 p_{10} + \mu_2 p_{b1} \\
\mu_1 p_{10} &= \lambda p_{00} + \mu_2 p_{11} \\
(\mu_1 + \mu_2)\, p_{11} &= \lambda p_{01} \\
\mu_2 p_{b1} &= \mu_1 p_{11}
\end{aligned}$$

Definiendo $\rho_1 = \frac{\lambda}{\mu_1}$ y $\rho_2 = \frac{\lambda}{\mu_2}$ e imponiendo la condición de que todas las

probabilidades sumen 1 se llega al siguiente sistema de ecuaciones:

$$\begin{aligned}
\rho_2 p_{00} &= p_{01} \\
(1+\rho_2)\,\rho_1 p_{01} &= \rho_2 p_{10} + \rho_1 p_{b1} \\
\rho_2 p_{10} &= \rho_1\rho_2 p_{00} + \rho_1 p_{11} \\
(\rho_2+\rho_1)\,p_{11} &= \rho_1\rho_2 p_{01} \\
\rho_1 p_{b1} &= \rho_2 p_{11} \\
p_{00}+p_{01}+p_{10}+p_{11}+p_{b1} &= 1
\end{aligned}$$

Así se obtiene:

$$\begin{aligned}
p_{01} &= \rho_2 p_{00} \\
p_{11} &= \frac{\rho_1\rho_2}{\rho_2+\rho_1}p_{01} = \frac{\rho_1\rho_2^2}{\rho_2+\rho_1}p_{00} \\
p_{b1} &= \frac{\rho_2}{\rho_1}p_{11} = \frac{\rho_2^3}{\rho_2+\rho_1}p_{00} \\
p_{10} &= \rho_1 p_{00} + \frac{\rho_1}{\rho_2}p_{11} = \rho_1 p_{00} + \frac{\rho_1^2\rho_2}{\rho_2+\rho_1}p_{00} \\
&= \frac{\rho_1\rho_2+\rho_1^2+\rho_1^2\rho_2}{\rho_2+\rho_1}p_{00} \\
1 &= p_{00}+p_{01}+p_{10}+p_{11}+p_{b1} \\
&= \left(1+\rho_2+\frac{\rho_1\rho_2+\rho_1^2+\rho_1^2\rho_2}{\rho_2+\rho_1}+\frac{\rho_1\rho_2^2}{\rho_2+\rho_1}+\frac{\rho_2^3}{\rho_2+\rho_1}\right)p_{00} \\
\Leftrightarrow\ 1 &= \frac{\rho_2+\rho_1+\rho_2^2+2\rho_1\rho_2+\rho_1^2+\rho_1^2\rho_2+\rho_1\rho_2^2+\rho_2^3}{\rho_2+\rho_1}p_{00}
\end{aligned}$$

De esta forma, denotando por $A = \rho_2+\rho_1+\rho_2^2+2\rho_1\rho_2+\rho_1^2+\rho_1^2\rho_2+\rho_1\rho_2^2+\rho_2^3$, se tiene:

$$\begin{aligned}
p_{00} &= \frac{\rho_1+\rho_2}{A} \\
p_{01} &= \frac{\rho_1\rho_2+\rho_2^2}{A} \\
p_{10} &= \frac{\rho_1\rho_2+\rho_1^2+\rho_1^2\rho_2}{A} \\
p_{11} &= \frac{\rho_1\rho_2^2}{A} \\
p_{b1} &= \frac{\rho_2^3}{A}
\end{aligned}$$

En el caso particular en que los servidores de ambos nodos tengan la misma

capacidad de servicio ($\mu_1 = \mu_2 = \mu$) entonces, denotando $\rho = \frac{\lambda}{\mu}$ se tiene:

$$\begin{aligned} p_{00} &= \frac{2}{2+4\rho+3\rho^2} \\ p_{01} &= \frac{2\rho}{2+4\rho+3\rho^2} \\ p_{10} &= \frac{2\rho+\rho^2}{2+4\rho+3\rho^2} \\ p_{11} &= \frac{\rho^2}{2+4\rho+3\rho^2} \\ p_{b1} &= \frac{\rho^2}{2+4\rho+3\rho^2} \end{aligned}$$

A partir de estas cantidades puede obtenerse una fórmula para el número medio de clientes en el sistema:

$$L = 0p_{00} + 1\,(p_{01} + p_{10}) + 2\,(p_{11} + p_{b1}) = \frac{4\rho+5\rho^2}{2+4\rho+3\rho^2}.$$

Como puede verse, si ρ tiende a cero, dicho número también tiende a cero (como era de esperar). Por el contrario, si $\rho \to \infty$ entonces $L \to \frac{5}{3} < 2$. La explicación de este hecho es que, aunque la entrada de clientes sea desmesurada respecto al servicio, el hecho de que las tasas de servicio sean iguales en ambos servidores provoca que, cuando $\rho \to \infty$, $p_{10} \to \frac{1}{3}$ y por tanto no tiende a estar siempre el sistema lleno. El tiempo medio de un cliente en todo el sistema puede hallarse fácilmente mediante la fórmula de Little generalizada. Así,

$$\overline{\lambda} = \lambda\,(p_{00} + p_{01}) = \lambda\frac{2+2\rho}{2+4\rho+3\rho^2}$$

y, por tanto,

$$W = \frac{L}{\overline{\lambda}} = \frac{4\rho+5\rho^2}{\lambda\,(2+2\rho)} = \frac{5\lambda+4\mu}{2\lambda\mu+2\mu^2}.$$

12.5 Ejercicios propuestos

1. Al servicio técnico de la casa de ordenadores marca ACME llegan aparatos averiados según un proceso de Poisson a razón de 16 cada día (jornada laboral de 8 horas). En el 25% de los casos, debido al tipo de avería descrita por el usuario, los ordenadores son remitidos directamente (perdiendo un tiempo despreciable) al servicio central, donde serán reparados por uno de los cuatro empleados asignados al efecto. En tal caso, el tiempo de reparación es exponencial con media de 4 horas. En el 75% restante

de los casos los ordenadores son examinados (y posiblemente reparados) por uno de los dos técnicos del propio servicio técnico, quienes emplean para ello un tiempo exponencial, siendo capaces de atender cada uno a una media de 8 cada jornada laboral. Aún así, un 25% de los aparatos que son revisados en el servicio técnico no pueden arreglarse allí y han de ir definitivamente al servicio central. Calcular:

(a) El número medio de días laborables que un aparato elegido al azar tardará en ser arreglado.

(b) La misma cantidad que antes para aquellos ordenadores que son enviados directamente al servicio central.

(c) El número medio de ordenadores averiados que se encuentran en el servicio central en espera de ser atendidos.

(d) El porcentaje de tiempo que cada reparador estará desocupado en el servicio técnico y en el servicio central respectivamente.

2. Un sistema informático envía mensajes que llegan, según un proceso de Poisson de parámetro λ, mediante un único servidor. El tiempo que tarda en transmitirse el mensaje y recibir, el servidor, acuse de recepción, se ajusta a una exponencial de parámetro μ. Se supone que con cierta probabilidad, c, el mensaje ha sido correctamente transmitido y sigue su camino, mientras que en caso contrario, el mensaje vuelve a hacer cola en el servidor para ser enviado nuevamente. Sabiendo que la disciplina en el servidor es FIFO, ¿cuál debe ser la condición para que este sistema sea estacionario? Dar una expresión, lo más sencilla posible, en función de λ, μ y c, para el número medio de mensajes que esperan a completar totalmente su transmisión.

3. Un sistema informático está compuesto por dos estaciones de trabajo monoprocesadoras idénticas, conectadas entre sí. A la estación 1 llegan procesos originados fuera del sistema, a razón de λ por unidad de tiempo, así como también otros procedentes de la estación 2. Cada vez que se termina la ejecución de un proceso de la estación 1, con probabilidad de un 90%, éste genera un nuevo proceso a ejecutar en la estación 2. En el 10% de casos restantes, el proceso termina sin crear más procesos "hijos". Cuando un proceso (hijo de otro) acaba su ejecución en la estación 2, puede crear otro hijo en la estación 1 (con probabilidad de un 10%) o extinguirse completamente (en el 90% de casos). Sabiendo que el tiempo de ejecución de los procesos (en ambas máquinas) se ajusta a una distribución exponencial de parámetro μ y que el tiempo entre llegadas de dos procesos a la estación 1 es también exponencial de parámetro λ, se pide:

(a) ¿Para qué valores de ρ es el sistema estacionario?

(b) Tomando $\rho = 0.9$, ¿cuál es el número medio de procesos totales en cada estación de trabajo?

(c) En las hipótesis del apartado anterior y tomando $\lambda = 5$ procesos por minuto, ¿cuál es el tiempo medio que transcurre desde que llega un proceso a la estación 1 hasta que él y todos sus descendientes se extinguen?

4. Una estación de trabajo recibe peticiones según un proceso de Poisson, a razón de 20 cada hora. Los tiempos de CPU de los trabajos siguen una distribución exponencial con media de dos minutos. Se sabe que, al finalizar cada proceso, en un 20% de los casos éste crea otro proceso hijo, con idénticas características que los trabajos que recibe originalmente la estación, y que ha de procesarse nuevamente por la misma. Calcular

   (a) el tiempo medio que tarda un trabajo en ser procesado.

   (b) el tiempo medio que tarda un trabajo y toda su descendencia (procesos hijo, nieto, etc.) en ser procesados.

   (c) el número medio de trabajos que esperan en trámite de procesarse (incluyendo el que está siendo procesado).

   (d) la probabilidad de que un proceso pueda ejecutarse directamente sin demora.

5. A una cola, A, de tipo $M/M/1$ llegan clientes según una tasa λ_A. Dichos clientes son atendidos por un servidor en un tiempo aleatorio con distribución exponencial de parámetro μ_A. Cada cliente que sale servido del sistema A se adentra en una nueva cola, B, la cual dispone de un mecanismo de servicio con un sólo servidor que tarda en dar servicio un tiempo exponencial y es capaz de atender μ_B clientes por unidad de tiempo. ¿Cuáles son las condiciones para que exista estacionariedad en todo el sistema formado por ambas colas? Suponiendo que se alcanza una condición de estado estacionario, calcular (en función de λ_A, μ_A, λ_B y μ_B) el tiempo medio que un cliente pierde en el sistema conjunto.

6. Dos estaciones de trabajo conectadas en red ejecutan procesos que llegan desde fuera del sistema (según un proceso de Poisson) a razón de 6 procesos por minuto a la estación 1 y 8 procesos por minuto a la estación 2. El tiempo de ejecución de los procesos en cada estación se ajusta a una distribución exponencial con media de 3 segundos para la estación 1 y de 2.5 segundos para la estación 2. Cuando un proceso termina su ejecución en la estación 1, en un 90% de los casos éste no genera ningún nuevo proceso hijo, el 6% de las veces crea un nuevo proceso que se envía (instantáneamente) a la estación 2 y el 4% restante se crea un nuevo proceso que se ejecutará en la estación 1. Algo semejante ocurre con los procesos ejecutados en la estación 2, siendo un 95% la probabilidad de

que no se genere un nuevo proceso, un 3% la de que se envíe un nuevo proceso a la estación 1 y un 2% la de que se cree un nuevo proceso a ejecutar en la estación 2. Se pide:

(a) El tiempo medio que transcurre desde que un proceso se envía a la estación 2 hasta que termina su ejecución.

(b) El número medio de procesos en cada estación.

(c) La probabilidad de que, entre las dos estaciones, haya más de cuatro procesos no totalmente ejecutados.

(d) Dado un proceso elegido al azar, calcular el tiempo medio que transcurre desde su entrada en el sistema hasta el fin de la ejecución de todos los procesos que, directamente o indirectamente, fueron creados a partir de él.

7. Considérese una red de colas cíclica formada por tres colas, es decir, un total de 3 colas de tal forma que cuando un cliente sale servido del nodo i ($i = 1, 2$) pasa a la cola del nodo $i + 1$ y que cuando el cliente sale del nodo 3, se dirigirá al nodo 1. En el caso de que los mecanismos de servicio en todos los nodos consten de un único servidor y de que el tiempo de servicio del mismo sea exponencial, encontrar las ecuaciones de balance en el caso de que existan 2 clientes en toda la red. Calcular las probabilidades de los distintos estados en el caso de que los parámetros de las distribuciones exponenciales de los respectivos tiempos de servicio fuesen $\mu_1 = 1, \mu_2 = 2, \mu_3 = 4$.

8. Por un pequeño sistema de comunicaciones con tres ordenadores en red circulan constantemente dos paquetes de bits que tardan en ser procesados en cada ordenador (con un único procesador) un tiempo con distribución exponencial. Los tiempos medios de procesado (en segundos) para cada ordenador vienen dados por la siguiente tabla:

Ordenador	1	2	3
Tiempo medio de procesado (seg.)	10	20	30

El 90% de las veces que un paquete termina de ser procesado por el ordenador 1, se dirige luego al ordenador 3, mientras que en caso contrario vuelve a procesarse en el ordenador 1. Una de cada cuatro veces que un paquete es procesado en el ordenador 2, vuelve a ser procesado nuevamente en el mismo ordenador mientras que las otras tres de cada cuatro veces se dirije al ordenador 3. Finalmente, el 40% de las veces que un paquete se procesa en el ordenador 3, vuelve a dicho ordenador, repartiéndose el 60% restante de forma equiprobable entre los otros dos ordenadores. Se pide:

(a) ¿Cuál es la probabilidad de que haya algún paquete procesándose en el ordenador 1?

(b) Calcular el número medio de paquetes procesándose en el ordenador 3.

(c) ¿Cuál es el tiempo medio que transcurre desde que llega un paquete al ordenador 3 hasta que termina de procesarse (sin incluir realimentación)?

9. Un sistema informático consta de dos estaciones de trabajo monoprocesadoras. Los tiempos de procesado de los trabajos son de distribución exponencial y media de 6 y 15 segundos para las estaciones 1 y 2 respectivamente. Se sabe que el 60% de los trabajos procesados en la primera estación son enviados a la segunda, realimentándose la primera estación en un 20% de los casos y saliendo totalmente ejecutados los procesos en el 20% restante. Por su parte, ocho de cada diez procesos ejecutados en la estación 2 son enviados a la estación 1, siendo terminada la ejecución en su totalidad en el 20% restante de los casos. Sabiendo que las llegadas de procesos al sistema obedecen un proceso de Poisson con media de 120 procesos por hora, ¿de qué manera debe distribuirse dicha entrada de procesos entre las dos estaciones para que el número medio total de procesos en el sistema sea mínimo? ¿Cuánto vale dicho número medio mínimo? ¿Hace esto mínimo el tiempo medio total en el sistema? ¿Por qué?

10. En un sistema informático, que consta de tres estaciones de trabajo, se dispone de un mecanismo de seguridad consistente en la ejecución sucesiva de dos programas idénticos de "detección de intrusos" en las distintas máquinas. Así, para un programa fijo, primero se ejecuta completamente en la estación 1, luego se ejecuta en la 2, una vez terminada su ejecución se ejecuta en la estación 3, luego pasa a ejecutarse de nuevo en la 1 y así sucesivamente. Debido a las características del sistema se sabe que los dos programas no pueden ejecutarse simultáneamente en una misma estación, esperando, el último en llegar, a que termine la ejecución el primero. Obviamente, cuando los programas se encuentran en distintas máquinas, la ejecución es simultánea. El tiempo de ejecución de cada programa es aleatorio y con distribución exponencial de media 1 segundo, para las estaciones 1 y 2, y 2 segundos para la estación 3 (con procesador más lento). Calcular:

(a) La probabilidad de que no haya ningún "detector de intrusos" en la estación 1.

(b) El número medio de "detectores de intrusos" en la estación 3.

(c) El tiempo medio que transcurre desde que un "detector de intrusos" llega a la estación 3 hasta que termina totalmente su ejecución.

(d) El tiempo medio que tarda un "detector de intrusos" en dar una vuelta completa de comprobación en el sistema.

11. Dos impresoras de un centro de cálculo se averían con mucha frecuencia. Para cada impresora, el tiempo hasta la próxima avería tiene distribución exponencial con media de 15 días. Cada vez que una de ellas está fuera de servicio, en un 75% de los casos la intenta reparar un técnico del propio centro de cálculo, siendo el tiempo que emplea, exponencial con media de 6 días. En el 25% restante de los casos la impresora es enviada directamente a un servicio técnico, en el que hay un operario que tarda en subsanar la avería un tiempo exponencial con media de 3 días. En realidad, después de que una impresora averiada haya sido revisada en el centro de cálculo, existe la posibilidad (en la tercera parte de casos) de necesitar todavía el ser reparada por el servicio técnico. ¿Cuál es la probabilidad de que las dos impresoras estén fuera de servicio?

13

Colas con distribuciones arbitrarias de llegada y servicio

En este capítulo se estudiarán algunas situaciones en las que la distribución del tiempo entre llegadas o la del tiempo de servicio deja de ser exponencial. En la práctica habitual raramente estas distribuciones se ajustan a una exponencial pero, como se verá el hecho de deshacernos de esta restricción acarreará notables problemas de cálculo para deducir los valores de los parámetros de comportamiento del sistema. A lo largo de este tema nos centraremos tan sólo en el estudio de uno de los modelos más simples que podemos imaginar, el $M/G/1$, es decir el caso de una cola con un sólo servidor, tiempos entre llegadas exponenciales y tiempos de servicio con distribución general, G. Como colofón incluiremos algunas ideas muy generales sobre cómo utilizar la simulación a modo de herramienta para poder aproximar las soluciones a problemas de teoría de colas cuando nos desviamos de las hipótesis de exponencialidad.

13.1 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 seguirá 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 porqué ser exponencial, el parámetro μ pasará ahora a significar el inverso del tiempo medio de servicio, es decir

$$\mu = \frac{1}{E(G)}.$$

Intuitivamente guarda una relación directa con el número medio de clientes que es capaz de servir el servidor por unidad de tiempo (aunque pueden no coincidir exactamente).

En lo que sigue trataremos de deducir las fórmulas que nos permitan obtener los valores de los parámetros de interés en un sistema como este. Para ello, denotaremos por X_n la variable aleatoria que contabiliza el "número de clientes en el sistema cuando el n-ésimo cliente lo abandona". Denótese por A_n el número de clientes que llegan durante el tiempo de servicio del n-ésimo cliente. Resulta fácil relacionar ambas cantidades de la siguiente forma:

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

donde

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

Tomando esperanzas en la expresión anterior y suponiendo estacionariedad (i.e., $L = E(X_n) = E(X_{n+1})$) se tiene:

$$L = L - E(U(X_n)) + E(A_{n+1}),$$

con lo cual:

$$E(U(X_n)) = E(A_{n+1}) = P(X_n > 0).$$

Esta última cantidad puede hallarse usando la distribución de la variable aleatoria S_{n+1}, que denota el tiempo de servicio del $(n+1)$-ésimo cliente, que en este caso tiene distribución general, que supondremos continua y con densidad g:

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

Los cálculos serían semejantes si la distribución de S_{n+1} fuese discreta o inlcuso con parte continua y parte discreta. En cualquier caso el resultado final al que se llega es el mismo.

Por otra parte, elevando al cuadrado los términos de la expresión que relaciona X_{n+1} con X_n y A_{n+1}, se obtiene:

$$\begin{aligned} X_{n+1}^2 &= X_n^2 + U(X_n)^2 + A_{n+1}^2 - 2X_n U(X_n) + 2X_n A_{n+1} - 2U(X_n) A_{n+1} \\ &= X_n^2 + U(X_n) + A_{n+1}^2 - 2X_n + 2X_n A_{n+1} - 2U(X_n) A_{n+1}. \end{aligned}$$

Tomando de nuevo esperanzas y teniendo en cuenta la independencia entre X_n y A_{n+1} se sigue:

$$\begin{aligned} E\left(X_{n+1}^2\right) &= E\left(X_n^2\right) + E(U(X_n)) + E\left(A_{n+1}^2\right) - 2E(X_n) \\ &\quad +2E(X_n) E(A_{n+1}) - 2E(U(X_n)) E(A_{n+1}) \\ &= E\left(X_n^2\right) + \rho + E\left(A_{n+1}^2\right) - 2L + 2L\rho - 2\rho^2. \end{aligned}$$

Utilizando otra vez la estacionariedad llegamos a

$$\begin{aligned} 2L - 2L\rho &= \rho - 2\rho^2 + E\left(A_{n+1}^2\right), \text{ con lo cual} \\ 2L(1-\rho) &= \rho - 2\rho^2 + E\left(A_{n+1}^2\right) \text{ y así} \\ L &= \frac{\rho - 2\rho^2 + E\left(A_{n+1}^2\right)}{2(1-\rho)}. \end{aligned}$$

Para encontrar una expresión cerrada de L debemos obtener otra que permita expresar $E\left(A_{n+1}^2\right)$ en términos de cantidades conocidas:

$$E\left(A_{n+1}^2\right) = Var(A_{n+1}) + E(A_{n+1})^2 = Var(A_{n+1}) + \rho^2,$$

pero, a su vez,

$$\begin{aligned} Var\left(A_{n+1}\right) &= E\left[Var\left(A_{n+1}|_{S_{n+1}}\right)\right] + Var\left[E\left(A_{n+1}|_{S_{n+1}}\right)\right] \\ &= E\left(\lambda S_{n+1}\right) + Var\left(\lambda S_{n+1}\right) = \lambda\frac{1}{\mu} + \lambda^2\sigma_S^2 = \rho + \lambda^2\sigma_S^2, \end{aligned}$$

siendo σ_S^2 la varianza del tiempo de servicio (es decir $\sigma_S^2 = Var\left(G\right)$). Así, podemos llegar a la expresión

$$E\left(A_{n+1}^2\right) = \rho + \lambda^2\sigma_S^2 + \rho^2,$$

que nos permite finalmente obtener la del número medio de clientes en el sistema:

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

Esta última expresión:

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

es la llamada fórmula de Pollaczek-Khintchine y a partir de ella pueden obtenerse W, W_q y L_q ya que siguen verificándose las fórmulas de Little y la relación entre tiempos medios en el sistema y en la cola.

Es digno de mención que, como caso particular, se obtiene la ya conocida fórmula para el modelo $M/M/1$, pues, en el caso $G = M$ se tiene que $\sigma_S^2 = \frac{1}{\mu^2}$, con lo cual

$$L_{M/M/1} = \rho + \frac{\rho^2 + \rho^2}{2\left(1-\rho\right)} = \rho + \frac{\rho^2}{1-\rho} = \frac{\rho - \rho^2 + \rho^2}{1-\rho} = \frac{\rho}{1-\rho}.$$

Obviamente, la distribución G que provoca un valor de L menor (fijado μ) es la distribución degenerada en el valor $\frac{1}{\mu}$ (que denotaremos por D), que corresponde con un servicio de duración constante e igual a dicho valor. Esto es inmediato pues $\sigma_S^2 = 0$ y, por tanto

$$L_{M/D/1} = \rho + \frac{\rho^2}{2\left(1-\rho\right)} = \frac{2\rho - 2\rho^2 + \rho^2}{2\left(1-\rho\right)} = \frac{\rho}{1-\rho}\left(1 - \frac{\rho}{2}\right),$$

que en el caso límite ($\rho \to 1^-$) tendería a ser la mitad del análogo para un $M/M/1$.

De la fórmula de Pollaczek-Khintchine puede deducirse que el efecto de la distribución del tiempo de servicio en el número medio de clientes en el sistema por unidad de tiempo puede ser infinitamente grande. Así, por ejemplo,

considerando para el tiempo de servicio, S, una distribución discreta con masa de probabilidad sobre el cero y otro punto para que su media sea $\frac{1}{\mu}$, tenemos:

$$\begin{aligned} P(S=0) &= 1-p \\ P\left(S=\frac{1}{\mu p}\right) &= p \end{aligned}$$

con $p \in (0,1]$. Es evidente que $E(S)=\frac{1}{\mu}$. Además,

$$E\left(S^2\right)=\frac{1}{\mu^2 p^2}p=\frac{1}{\mu^2 p},$$

con lo cual

$$\sigma_S^2=Var(S)=\frac{1}{\mu^2 p}-\left(\frac{1}{\mu}\right)^2=\frac{1-p}{\mu^2 p}.$$

De esta manera, la fórmula de Pollaczek-Khintchine nos lleva a

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

que es una cantidad que tiende a ∞ cuando $p \to 0^+$. Dicho en otros términos, aún considerando valores fijos de λ y μ, el número medio de clientes en un sistema $M/G/1$ puede ser tan elevado como se desee sin más que elegir una distribución del tiempo de servicio suficientemente desfavorable (es decir con varianza suficientemente grande). Esto incide, una vez más, en las graves consecuencias que puede tener el especificar incorrectamente la distribución del tiempo de servicio.

Como se ha visto, la obtención de una fórmula para L en un modelo $M/G/1$ general es algo bastante laborioso. Todavía más lo es el cálculo de las probabilidades de los distintos estados del sistema y aún peor para las funciones de distribución $W(t)$ y $W_q(t)$. Daremos sólo una idea general de cómo calcular las p_n.

Puede demostrarse que las probabilidades de los distintos estados (p_n) son solución del siguiente sistema de infinitas ecuaciones:

$$p_n=p_0k_n+\sum_{m=1}^{n+1}p_m k_{n-m+1}, \text{ para } n=0,1,2,\ldots$$

siendo

$$k_n=P(A_m=n)=\begin{cases} \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!}g(t)\,dt & \text{en el caso continuo} \\ \sum_i \frac{e^{-\lambda s_i}(\lambda s_i)^n}{n!}q_i & \text{si } S \text{ es discreta con} \\ & q_i\text{: masa de probabilidad} \end{cases}$$

Dado que en muchas ocasiones resulta difícil resolver el sistema suele acudirse a las funciones características para encontrar la solución. La idea consiste en definir

$$P(z) = \sum_{n=0}^{\infty} p_n z^n \text{ y } K(z) = \sum_{n=0}^{\infty} k_n z^n,$$

que están relacionadas a través de la igualdad

$$P(z) = \frac{(1-\rho)(1-z)K(z)}{K(z)-z}.$$

Dado que las constantes k_n son calculables (con mayor o menor dificultad) a partir de la distribución del tiempo de servicio y habida cuenta de las relaciones entre las constantes del desarrollo en serie de las funciones características y sus derivadas sucesivas en el punto cero

$$\begin{aligned} P(0) &= p_0,\ K(0) = k_0 \\ P'(0) &= p_1,\ K'(0) = k_1 \\ P''(0) &= 2p_2,\ K''(0) = 2k_2 \\ &\vdots \\ P^{(n)}(0) &= n!p_n,\ K^{(n)}(0) = n!k_n \\ &\vdots \end{aligned}$$

podemos encontrar fórmulas que permitan el cálculo de las p_n. Veamos un ejemplo.

Ejemplo 54 *Dado un modelo $M/D/1$, calcular las fórmulas para las probabilidades de n clientes en el sistema.*

Solución: Al ser degenerada la distribución del servicio se tiene

$$P\left(S = \frac{1}{\mu}\right) = 1$$

y, por tanto,

$$k_n = \frac{e^{-\lambda\frac{1}{\mu}}\left(\lambda\frac{1}{\mu}\right)^n}{n!} = \frac{e^{-\rho}\rho^n}{n!}, \text{ para todo } n = 0, 1, 2, \ldots$$

De esto se sigue que

$$\begin{aligned} K(z) &= \sum_{n=0}^{\infty} k_n z^n = \sum_{n=0}^{\infty} \frac{e^{-\rho}\rho^n}{n!} z^n = e^{-\rho} \sum_{n=0}^{\infty} \frac{(\rho z)^n}{n!} \\ &= e^{-\rho} e^{\rho z} = e^{\rho(z-1)}. \end{aligned}$$

Utilizando esta expresión resulta

$$P(z) = \frac{(1-\rho)(1-z)e^{\rho(z-1)}}{e^{\rho(z-1)}-z} = (1-\rho)\frac{1-z}{1-ze^{-\rho(z-1)}}.$$

En particular

$$\begin{aligned}
p_0 &= P(0) = 1-\rho,\\
P(z) &= (1-\rho)\frac{1-z}{1-ze^{-\rho(z-1)}},\\
P'(z) &= (1-\rho)\frac{-1+e^{-\rho(z-1)}z^2\rho+e^{-\rho(z-1)}-z\rho e^{-\rho(z-1)}}{\left(1-ze^{-\rho(z-1)}\right)^2},\\
p_1 &= P'(0) = (1-\rho)(e^{\rho}-1),\\
P''(z) &= (1-\rho)e^{-\rho(z-1)}\\
&\quad\times\frac{2-4z\rho+z^2\rho^2+2\rho-z\rho^2+e^{-\rho(z-1)}(z^3\rho^2-2+2z\rho-z^2\rho^2)}{\left(-1+ze^{-\rho(z-1)}\right)^3},\\
p_2 &= \frac{P''(0)}{2} = \frac{(1-\rho)e^{\rho}(-2+2e^{\rho}-2\rho)}{2}\\
&= (1-\rho)e^{\rho}(-1+e^{\rho}-\rho).
\end{aligned}$$

Con cálculos semejantes pero laboriosos se puede obtener

$$\begin{aligned}
p_3 &= \frac{P'''(0)}{3!} = \frac{(1-\rho)e^{\rho}(6\rho+3\rho^2+6e^{2\rho}-6e^{\rho}-12\rho e^{\rho})}{6}\\
&= (1-\rho)e^{\rho}\left(\rho+\frac{1}{2}\rho^2+e^{2\rho}-e^{\rho}(2\rho+1)\right),\\
p_4 &= \frac{P^{(4)}(0)}{4!}\\
&= \frac{(1-\rho)e^{\rho}(48\rho e^{\rho}-24e^{2\rho}-12\rho^2-4\rho^3+48\rho^2e^{\rho}-72e^{2\rho}\rho+24e^{3\rho})}{24}\\
&= (1-\rho)e^{\rho}\left(2\rho e^{\rho}-e^{2\rho}-\frac{1}{2}\rho^2-\frac{1}{6}\rho^3+2\rho^2e^{\rho}-3e^{2\rho}\rho+e^{3\rho}\right).
\end{aligned}$$

De hecho, puede demostrarse por inducción la siguiente fórmula general

$$\begin{aligned}
p_n &= (1-\rho)\left\{e^{n\rho}+\sum_{i=1}^{n-1}(-1)^{n-i}e^{i\rho}\left[\frac{(i\rho)^{n-i}}{(n-i)!}+\frac{(i\rho)^{n-i-1}}{(n-i-1)!}\right]\right\},\\
\textit{para } n &= 0,1,2,\ldots
\end{aligned}$$

Aunque en el ejemplo anterior se pudo obtener una fórmula general para todos los k_n y estó permitió encontrar una expresión cerrada para la función $K(z)$ esto no es siempre posible. De todos modos es fácil darse cuenta de que basta conocer un número finito de k_n (los primeros) para poder ir obteniendo

unas cuantas p_n. La forma de conseguirlo es derivar formalmente sucesivamente ambos miembros de la expresión

$$P(z) = \frac{(1-\rho)(1-z)K(z)}{K(z)-z}$$

e ir evaluando dichas derivadas sucesivas en $z = 0$. Con esto se irán encontrando expresiones de las sucesivas p_n en términos de $k_0, k_1, \ldots, k_n$. Así, por ejemplo:

$$P(0) = \frac{(1-\rho)(1-0)K(0)}{K(0)-0} = \frac{(1-\rho)k_0}{k_0} = 1-\rho,$$

con lo cual se tiene que

$$p_0 = P(0) = 1-\rho$$

para cualquier modelo $M/G/1$. Además

$$P'(z) = (1-\rho)\frac{[-K(z)+(1-z)K'(z)][K(z)-z]-(1-z)K(z)[K'(z)-1]}{[K(z)-z]^2}$$

y, por tanto,

$$\begin{aligned} P'(0) &= (1-\rho)\{[-K(0)+(1-0)K'(0)][K(0)-0] \\ &\quad -(1-0)K(0)[K'(0)-1]\}[K(0)-0]^{-2} \\ &= (1-\rho)\frac{(-k_0+k_1)k_0-k_0(k_1-1)}{k_0^2} \\ &= (1-\rho)\frac{-k_0^2+k_0}{k_0^2} = (1-\rho)\left(\frac{1}{k_0}-1\right). \end{aligned}$$

Pudiendo seguir (con el inconveniente de la complejidad de cálculo) tanto como se desee.

13.2 Aproximación mediante simulación

Como hemos visto en el caso de un modelo $M/G/1$ la resolución analítica de un modelo de colas puede complicarse mucho cuando la distribución del tiempo entre llegadas o la del tiempo de servicio no son exponenciales. La simulación proporciona una herramienta que permite encontrar, de forma aproximada, los valores de las cantidades que habitualmente nos interesan en tales modelos.

Como sabemos, una fase previa a la resolución de un problema mediante simulación es la modelización de la realidad. Esto consiste en especificar las

variables realmente importantes que parecen regir el sistema real y las relaciones que existen entre ellas. En ocasiones la fase de modelización es muy laboriosa con frecuentes realimentaciones en las que se va depurando el modelo construído paso a paso. En este sentido, la teoría de colas es precisamente una disciplina que trata de modelizar situaciones reales en las que existe espera para dar un servicio. Así, por tanto, supondremos que el problema ya se ha expresado como un modelo de teoría de colas y que se han especificado ya las distribuciones del tiempo entre llegadas y del tiempo de servicio. Obviamente esto último ha de requerir una experimentación real previa por la cual se hayan ido observando los tiempos para una serie de clientes de ambas variables. Ni qué decir tiene que, siempre que el modelo pueda resolverse de forma analítica (por ejemplo un $M/M/s/\infty/H$) ésta será preferible a la aproximación por simulación, pues da una solución exacta. De todas formas, en muchos casos (sobre todo ante la falta de exponencialidad) dicha resolución analítica no será factible o será muy costosa y es entonces cuando la aproximación por simulación resulta muy útil.

Para fijar las ideas, consideremos un modelo de colas muy sencillo pero para el que no disponemos (con los conocimientos vistos hasta ahora) de una herramienta analítica para resolverlo. Pensemos en una cola con un único servidor, capacidad de hasta tres clientes en la cola y fuente de entrada ilimitada. Supongamos que, en base a experimentación previa, hemos encontrado que la forma razonable de modelizar los tiempos entre llegadas es la distribución de Weibull, $W(\beta, \alpha)$, que es una generalización de la distribución $\exp(\alpha)$. Su función de densidad de probabilidad es

$$f(x) = \alpha\beta^{\alpha}x^{\alpha-1}e^{-(\beta x)^{\alpha}}, \text{ para todo } x \geq 0.$$

Como se ve, $W(\beta, 1) \stackrel{\mathrm{d}}{=} \exp(\beta)$, pero nosotros supondremos que en nuestro ejemplo los tiempos entre llegadas (en minutos) parecen ajustarse a una $W(0.1, 0.25)$. De forma parecida, tras los estudios oportunos hemos llegado a la conclusión de que la duración en minutos del tiempo de servicio sigue una $U(2, 8)$. Se trata, por tanto, de un modelo $W/U/1/3$.

Para encontrar mediante simulación aproximaciones de las cantidades de interés podemos proceder del siguiente modo:

1. Suponer inicialmente que no hay ningún cliente en el sistema (i.e. $N = 0$).

2. Simular artificialmente el tiempo que tardará en llegar el primer cliente (de acuerdo a una distribución $W(0.1, 0.25)$).

3. Una vez dejado transcurrir dicho tiempo en el cronómetro artificial del sistema, actualizar el número de clientes del sistema ($N = 1$), obtener, mediante el algoritmo de generación de números aleatorios según una $U(2, 8)$, el valor del tiempo que tardaría en dársele servicio a dicho cliente y también el tiempo que transcurriría hasta la llegada al sistema del segundo cliente (según una $W(0.1, 0.25)$).

4. Comparar ambos tiempos para saber cuál de ambas cosas sucederá antes y dejar transcurrir el tiempo de ese suceso (el que antes sucede).

5. Si lo que ocurre antes es que llega el nuevo cliente entonces se haría $N = 2$, contabilizando el tiempo durante el cuál ha habido un cliente en el sistema. Además deberíamos simular el tiempo hasta la llegada del tercer cliente y seguir el proceso.

6. En el caso de que el menor de los tiempos del paso 4 fuese el de servicio, se actualizaría N dándole el valor 0 y teniendo en cuenta el tiempo durante el cuál ha habido un cliente en el sistema. Luego habría que dejar transcurrir lo que resta de tiempo hasta la llegada del segundo cliente y continuar el proceso.

7. Según el esquema anterior se simularían las llegadas de nuevos clientes al sistema y los tiempos de servicio de clientes, teniendo en cuenta la restricción adicional de que si el instante de llegada (teórico) de un cliente al sistema se produce cuando ya hay tres en cola (además del que está siendo servido) entonces la limitación de la cola provocaría que ese cliente no puede entrar, simulándose el tiempo hasta la siguiente llegada.

Como se puede ver en los pasos anteriores, una simulación de esta naturaleza (la llamada simulación por eventos) requeriría la contabilización de una serie de variables que nos vayan dando el estado del sistema en cada instante: el número de clientes en el sistema, el tiempo que falta para que se produzca la siguiente llegada y el tiempo que resta para terminar el servicio del cliente que está siendo servido (si lo está siendo alguno). Lógicamente, para implementar esta simulación será necesario utilizar los algoritmos vistos en los primeros capitulos, que permitirán obtener valores articiales que sigan las distribuciones especificadas.

Así, en este caso que estamos viendo a título de ejemplo, puede usarse el método de inversión para simular fácilmente las distribuciones $W(0.1, 0.25)$ y $U(2, 8)$. Para ello basta obtener primeramente sus funciones de distribución:

$$\begin{aligned} F(x) &= \int_0^x f(t)\,dt = \int_0^x \alpha\beta^\alpha t^{\alpha-1} e^{-(\beta t)^\alpha} dt = \left[-e^{-(\beta t)^\alpha}\right]_{t=0}^{t=x} \\ &= 1 - e^{-(\beta x)^\alpha}, \text{ si } x \geq 0, \text{ para la Weibull,} \\ \text{con lo cual, } F(x) &= 1 - e^{-(0.1\cdot x)^{0.25}}, \text{ si } x \geq 0, \text{ para la } W(0.1, 0.25) \text{ y} \\ G(x) &= \int_2^x g(t)\,dt = \int_2^x \frac{1}{6}\,dt \\ &= \frac{x-2}{6}, \text{ si } x \in [2, 8], \text{ para la } U(2, 8) \end{aligned}$$

y luego encontrar sus respectivas inversas:

$$\begin{aligned} F(x) &= y \Leftrightarrow 1 - e^{-(0.1\cdot x)^{0.25}} = y \Leftrightarrow e^{-(0.1\cdot x)^{0.25}} = 1 - y \\ &\Leftrightarrow -(0.1\cdot x)^{0.25} = \ln(1-y) \Leftrightarrow 0.1\cdot x = [-\ln(1-y)]^4 \\ &\Leftrightarrow x = 10\cdot[-\ln(1-y)]^4 \Leftrightarrow F^{-1}(y) = 10\cdot[\ln(1-y)]^4, \\ G(x) &= y \Leftrightarrow \frac{x-2}{6} = y \Leftrightarrow x = 6y+2 \Leftrightarrow G^{-1}(y) = 6y+2. \end{aligned}$$

Finalmente, los algoritmos para simular ambas distribuciones serían:

Algoritmo de inversión para simular la distribución $W(0.1, 0.25)$
1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $X = 10\cdot[\ln(1-U)]^4$.

Algoritmo de inversión para simular la distribución $U(2,8)$
1. `Generar` $U \sim U(0,1)$.
2. `Devolver` $Y = 6U + 2$.

Mediante una programa de simulación como el esbozado más arriba sería posible obtener aproximaciones para las cantidades de interés. Así, por ejemplo, la probabilidad de que hubiese 2 clientes en el sistema se aproximaría mediante la fracción del tiempo simulado en el que el sistema artificial tuvo ese número de clientes. El tiempo medio que un cliente pasa en la cola puede aproximarse mediante simulación utilizando el promedio de los tiempos en la cola de los clientes que han pasado por la simulación.

13.3 Ejercicios propuestos

1. Tomando como fijas las tasas de llegada y servicio en un modelo $M/G/1$, ¿cuál es la distribución del tiempo de servicio que proporciona un menor número medio de clientes en el sistema? ¿Qué relación guarda dicho valor mínimo con el correspondiente a un modelo $M/M/1$?

2. A una centralita telefónica llegan llamadas, según un proceso de Poisson, a razón de 4 por minuto. El tiempo que emplea la única telefonista en direccionar cada llamada sigue una distribución discreta con los valores de 5 segundos, con probabilidad $\frac{1}{2}$ y 15 segundos con probabilidad $\frac{1}{2}$. Suponiendo que las llamadas recibidas mientras la operadora está ocupada se mantienen en espera para ser atendidas según su orden de llegada, calcular:

   (a) El número medio de llamadas en espera (incluyendo la que está atendiendo la telefonista).

(b) La probabilidad de que haya más de una llamada esperando a ser atendida por la telefonista.

3. Dado un modelo $M/\Gamma\left(q\cdot\mu,q\right)/1$, obtener el valor del número medio de clientes en el sistema como función de la tasa de llegadas, λ, y los dos parámetros involucrados en la distribución del servicio, μ y q. Dado que la media de la distribución de servicio es $\frac{q}{\mu\cdot q}=\frac{1}{\mu}$, discutir los valores que se obtienen para la cantidad anteriormente hallada, con respecto a la de un modelo $M/M/1$ (con tasa de llegadas λ y tasa de servicios μ), según los valores de $q>0$. ¿A que valores tiende dicha cantidad en los casos extremos ($q\to 0^{+}$ y $q\to\infty$)?

14

Bibliografía

Allen, A.O. (1990). *Probability, statistics and queueing theory with computer science applications.* Academic Press.

Azarang, M. R. y García Dunna, E. (1996). *Simulación y análisis de modelos estocásticos.* McGraw-Hill.

Bratley, P., Fox, B. L. y Schrage L. E. (1990). *A guide to simulation.* Springer-Verlag.

Bunday, B.D. (1996). *An introduction to queueing theory.* Arnold.

Devroye, L. (1986). *Non-uniform random variate generation.* Springer-Verlag.

Gelenbe, E. y Pujolle, G. (1987). *Introduction to queueing networks.* Wiley.

Gentle, J.E. (1998). *Random number generation and Monte Carlo methods.* Springer-Verlag.

Gross, D. y Harris, C.M. (1985). *Fundamentals of queueing theory.* Wiley.

Karian, Z.A. y Dudewicz E.J. (1991). *Modern statistical systems and GPSS simulation.* Computer Science Press.

Kleinrock, L. (1975). *Queueing systems. Volume I: Theory.* Wiley.

Kleinrock, L. (1975). *Queueing systems. Volume II: Computer applications.* Wiley.

Law, A.M. y Kelton, W.D. (1991). *Simulation, modeling and analysis.* McGraw-Hill.

Medhi, J. (1991). *Stochastic models in queueing theory.* Academic Press.

Moeschlin, O., Grycko, E., Pohl, C. y Steinert, F. (1998). *Experimental stochastics.* Springer-Verlag.

Nelson, R. (1995). *Probability, stochastic processes, and queueing theory : the mathematics of computer perfomance modelling.* Springer-Verlag.

Ross, S. M. (1999). *Simulación.* Prentice Hall.

Saaty, T.L. (1983). *Elements of queueing theory with applications.* Dover.

Trivedi, K.S. (1982). *Probability and statistics with reliability, queueing theory and computer science applications.* Prentice Hall.

Van Dijk, N.M. (1993). *Queueing networks and product forms. A systems approach.* Wiley.