Cátedra Nissan

-PROTHIUS-

Métodos Cuantitativos de Organización Industrial (Apuntes): Sistemas con esperas: Teoría de Colas y Simulación

Joaquín Bautista Valhondo, Albert Corominas Subias y Ramón Companys Pascual

D-08/2011
(Rec. MQOI1-1996-BCC)

Departamento de Organización de Empresas

Universidad Politécnica de Cataluña

Publica:

Universitat Politècnica de Catalunya
www.upc.edu

Edita:

Cátedra Nissan
www.nissanchair.com
director@nissanchair.com

Sistemas con esperas: Teoría de colas y simulación

Joaquín Bautista, Albert Corominas y Ramón Companys

21 de julio de 2011

Índice general

Agradecimientos

Los autores agradecen la recuperación, reescritura y revisión del presente texto a los profesores Rocío Alfaro Pozo y Alberto Cano Pérez.

Joaquín Bautista Valhondo

Barcelona, julio de 2011

Capítulo 1

Introducción a los sistemas con esperas

1.1. Naturaleza y características de los fenómenos de espera

Siempre que los flujos de entrada y de salida en un punto de un sistema no están perfectamente sincronizados se producen esperas. Por consiguiente, los fenómenos de espera surgen por doquier, en la vida cotidiana y en los sistemas productivos de bienes o de servicios.

Al cabo del día todos perdemos tiempo esperando, incluso muchas veces. En la gasolinera, en el peaje, en la cafetería, en el banco, para tomar el ascensor, para coger el autobús y para pagarlo, cuando llamamos por teléfono y el receptor comunica, etc.; para hacer un viaje en avión normalmente es preciso hacer varias colas. En una manufactura, las materias primas esperan en recepción y después en el almacén hasta que se incorporan al proceso productivo, los semielaborados esperan en diversos puntos de la instalación a que una máquina esté disponible, los productos acabados esperan en el almacén hasta ser distribuidos y los vehículos de reparto esperan a lo largo de su itinerario y en el momento de descargar; las máquinas averiadas esperan para ser reparadas; los operarios posiblemente harán cola ante relojes marcadores; los programas esperan para ser procesados por el ordenador, etc. En un centro hospitalario hay listas de espera de pacientes que aguardan turno para ser ingresados, pacientes en salas de espera, médicos que están pendientes de que les sea concedido el solicitado quirófano que han solicitado, colas ante los servicios administrativos para cumplimentar documentos o pagar las cuentas, colas en la cafetería, pacientes que esperan en sus habitaciones a ser atendidos por enfermeras ocupadas en la atención de otros enfermos, etc.; también hay objetos que esperan, por ejemplo, productos que, en la farmacia, permanecerán en una estantería hasta que les toque el turno de ser utilizados.

Estos ejemplos son, voluntariamente, muy dispares. La espera puede afectar a personas o a objetos. Las unidades pueden disponerse físicamente en una fila o formar una fila ideal en un archivo; pueden reunirse en una sala sin orden aparente pero con un orden real, el que será tenido en cuenta por el servicio: el de llegada a la sala, el correspondiente a unos números que les han sido asignados, a unas horas previamente convenidas o a unas fechas de fabricación o de caducidad; las unidades, finalmente, pueden permanecer fijas en su lugar, al cual deberá acudir la unidad que presta el ser-

vicio.

En todos los casos diremos que se producen *colas*, extendiendo la acepción habitual, más restrictiva, de este término.

Todas las colas son consecuencia de una armonización insuficiente entre un flujo de llegada de unidades a las que se ha de realizar una operación o prestar un servicio y el ritmo a que se realiza la actividad correspondiente. Por supuesto, se forman colas si la capacidad media de producir está por debajo de la que correspondería a la demanda media; si a un centro llega, en promedio, una unidad cada minuto y realizar el tratamiento exige dos minutos en promedio, es obvio que se formarán colas. Pero para evitarlas no basta con que el número *medio* de tratamientos por unidad de tiempo que el sistema puede realizar sea igual o mayor que el número *medio* de llegadas por unidad de tiempo. Si el sistema tarda exactamente un minuto en realizar el tratamiento y las unidades llegan exactamente a un ritmo de una por minuto, no habrá colas; pero si llegan paquetes de n unidades cada n minutos habrá colas cuya longitud será de hasta $n-1$ unidades. Más en general, hay colas (manteniendo el supuesto de tiempo de servicio constante) si el tiempo entre llegadas tiene una cierta dispersión; supóngase que las llegadas son al azar (en el sentido de que la probabilidad de que llegue una unidad es independiente del tiempo transcurrido desde la última llegada; ésta, por cierto, es una hipótesis que se cumple razonablemente en muchos sistemas reales en que las unidades que llegan no se coordinan y cada una, por consiguiente, se comporta con independencia de las demás) y que la tasa media de llegadas es igual a la tasa de prestación del servicio: evidentemente habrá intervalos de tiempo con colas (porque existe una probabilidad no nula de que lleguen unidades mientras el sistema está ocupado todavía con una unidad anterior) y habrá intervalos en que el sistema estará ocioso (porque ha dado servicio a la última unidad y no ha llegado todavía ninguna otra); por lo tanto, a pesar de que el sistema tiene una cierta capacidad de prestar servicio, sólo puede ejercerla durante una fracción del tiempo, por lo que de hecho, en un intervalo temporal dado, el flujo medio de salidas del sistema podrá ser menor que el flujo medio de llegadas y, a pesar de la igualdad entre la tasa media de llegadas y la capacidad de prestar el servicio, hay una probabilidad de formación de colas incluso muy largas. El fenómeno será más acentuado si existe también una dispersión en los tiempos de servicio.

No basta, pues, tal como se decía más arriba, con que la tasa media de llegadas sea inferior a la capacidad de servicio; para que no se formen colas es preciso además tener un control suficiente sobre la dispersión de los tiempos entre llegadas y la de los tiempos de servicio. En general es difícil controlar estas dispersiones, por motivos diversos y evidentes (las unidades que llegan a veces son externas al sistema por lo que no hay una relación jerárquica con ellas, en las llegadas intervienen causas aleatorias, el tiempo de servicio es variable tanto por factores internos al sistema como por las características diversas de las unidades y del servicio que demandan).

Ello no significa que las colas sean una fatalidad ante la que sólo cabe resignarse, pero la existencia de una cola no puede interpretarse sin más como síntoma de una mala gestión; lo será sólo si se trata de una cola de magnitud media exageradamente grande; es más, en algunos casos podría ser un síntoma de mala gestión, de derroche de

recursos, la situación opuesta.

Cuando hay esperas, hay una acumulación de unidades (de producto, de materias primas, personas, etc.) y por consiguiente aparece un stock (de materias primas, de productos acabados, de trabajos en curso, etc.) y, por tanto, un coste.

Para reducir estos efectos hay que armonizar los flujos; las capacidades medias han de ser suficientes, pero esto no basta, hay que hacer coincidir tanto como sea posible los flujos de entrada y salida a lo largo del tiempo, controlar los flujos, reducir la dispersión de los tiempos de proceso. A pesar de estos esfuerzos, subsistirán aspectos no controlables y se producirán ciertas esperas.

1.2. Costes asociados a los sistemas con esperas

Los fenómenos de espera tienen asociados unos costes, explícitos o implícitos:

- Costes de las unidades que prestan el servicio o *canales* (incluyendo el del espacio necesario para alojarlas).
- Coste del tiempo de estancia de las unidades en el sistema, que incluye el tiempo de espera más el de servicio (aunque ambos pueden tener una consideración distinta, generalmente se han de tener en cuenta los dos).
- Coste de desplazamiento de las unidades al servicio o del servicio a las unidades.
- Coste del espacio asignado para colocar las unidades que esperan.
- Coste de las perturbaciones derivadas de la falta de espacio.

Estos costes aparecen en casi todos los sistemas en que se producen esperas, pero no siempre son fáciles de cuantificar y en la elección de un sistema u otro deben tenerse en cuenta todos. Pocos canales de servicio o canales poco capaces serán poco costosos pero originarán elevados costes de espera; si comparamos dos canales de servicio iguales o uno solo con capacidad doble (y, para simplificar la comparación, de igual coste que el conjunto de los dos canales iguales), con supuestos muy poco restrictivos la segunda opción es preferible en lo que respecta a los tiempos totales de presencia de las unidades en el sistema, pero si se tiene en cuenta los costes de desplazamiento, que serán menores con dos unidades convenientemente localizadas que con una sola, tal vez la opción de menor coste sea la de dos canales. En esta comparación debería tenerse también en cuenta la distinta fiabilidad de una u otra solución: a igual probabilidad de avería la probabilidad de quedarse sin servicio es mayor en el caso de un solo canal.

Cuando las unidades que acuden al canal o a los canales de servicio son parte integrante del propio sistema productivo los costes de estancia en el sistema no son difíciles de estimar y, en cualquier caso, su existencia es indiscutible y notoria (por ejemplo, tiempo perdido por los operarios que hacen cola en un centro de distribución de herramientas). Pero si las unidades son externas al sistema, su espera no representa un coste, directamente o inmediatamente, para el sistema productivo o resulta muy difícil de evaluar; la consideración de este aspecto depende del tipo de sistema o de su posición en el entorno, que puede ser el mercado: si hay competencia, las esperas pueden

hacer perder clientes y ello tiene objetivamente un coste aunque sea difícil de estimar; si no hay competencia, el coste, a corto plazo, lo asumen las unidades que desean obtener el servicio pero lo razonable es proponerse un nivel de servicio mínimo. El nivel de servicio es un concepto muy general que se puede formalizar de diversas formas; para ello hay que elegir una o más magnitudes (por ejemplo: longitud media de la cola, tiempo medio de espera, tiempo medio de estancia en el sistema, proporción de unidades cuya estancia en el sistema es superior a un cierto valor, t_0, etc. o una cierta ponderación de estos parámetros) y, para cada una de ellas, el valor mínimo o máximo, según el caso, que se considera aceptable.

En definitiva, trátese de hacer mínimo el coste total de funcionamiento o de conseguir un cierto nivel de servicio a un coste mínimo, interesa dimensionar los sistemas productivos de bienes o de servicios con el fin de conseguir el objetivo propuesto.

¿Sobre qué características del sistema se puede actuar y de qué técnicas se dispone para resolver este tipo de problemas?

Normalmente, cabe elegir entre distintos tipos de canales (cada uno de los cuales se caracterizará por su capacidad de servicio media, por la dispersión de la misma y por sus costes) y determinar el número de los mismos, así como el sistema por el que se elige la unidad a atender (o *disciplina de cola*). Además, se puede muchas veces actuar sobre la forma en que se producen las llegadas.

Para caracterizar un sistema de colas (conjunto formado por las unidades que llegan y tal vez esperan y los canales de servicio) se debe especificar:

- la ley que rige las llegadas
- la ley que rige el tiempo de servicio (y el número de canales de cada tipo)
- la disciplina de la cola.

Algunas clasificaciones importantes son:

- En relación con las llegadas:
  - Si se pueden producir o no en grupo.
  - Si la longitud de la cola disuade (con una cierta probabilidad) o no a las unidades que pretenden incorporarse al sistema.
  - Si el número de unidades que pretenden incorporarse al sistema depende o no del número de unidades que se encuentran en el mismo.
- En relación con el servicio:
  - Si la ley de servicio es la misma siempre y para todas las unidades o si depende del tipo de unidades (de hecho, la existencia de diversos tipos de unidades puede considerarse una característica de las llegadas) o de la longitud de la cola (incluso el número de canales puede depender de la longitud de la cola).
- En relación con la disciplina de cola:

- Puede ser: primer llegado - primer servido (FIFO: *first in, first out*), último llegado - primer servido (LIFO: *last in, first out*), al azar (SIRO: *service in random order*), con prioridades según una jerarquía de unidades (con o sin interrupción del servicio). En el caso de diversos canales en paralelo la cola puede ser única o puede haber una cola para cada canal (en este caso, la distribución de las unidades entre las colas puede obedecer a una asignación previa o bien a la voluntad de las unidades que se pueden incorporar, por ejemplo, a la cola más corta en el momento que llegan al sistema); evidentemente también influye en el funcionamiento del sistema, en el caso de cola múltiple, que las unidades se tengan que mantener en la cola que han elegido o a que han sido asignadas o que puedan cambiar o ser cambiadas. También se puede presentar el caso (así ocurre en muchos sistemas reales) de que exista una cierta probabilidad de abandonar la cola por parte de las unidades en función por ejemplo del tiempo de espera ya transcurrido y del número de unidades que le preceden en la cola.

Todo ello se refiere a sistemas con un solo canal de servicio o con diversos canales en paralelo. Pero realmente existen sistemas mucho más complejos: canales en serie, flujos que se bifurcan según diversas reglas posibles, flujos que se reúnen: en general, las unidades pasan sucesivamente por diversos sistemas de espera; en la realidad no sólo hay colas aisladas: se encuentran con frecuencia las denominadas *redes de colas*.

1.3. Gestión de las colas

Puesto que las colas no sólo existen sino que muchas veces son inevitables es absurdo ignorarlas y dejar que funcionen espontáneamente; es preciso tenerlas en cuenta como un elemento en el sistema productivo, un elemento que ha de ser objeto de una gestión específica.

Donde pueda haber esperas tiene que existir un espacio para almacenar las unidades que forman la cola. Si no, se generará desorden (unidades colocadas en lugares inapropiados, no previstos para ello: en los pasillos, a la intemperie, por ejemplo, lo que puede producir otras demoras o deterioros o repercutir en la seguridad, por ejemplo) o tal vez un bloqueo del sistema. Esto último es lo que puede ocurrir en una línea de montaje o de producción si no se ha previsto un espacio suficiente para el almacenamiento intermedio entre las estaciones de trabajo; en el caso extremo de que no haya espacio alguno para ello cualquier avería en una estación o cualquier incremento en el tiempo de una operación puede bloquear el funcionamiento de la línea, hacia arriba o hacia abajo de la estación que presenta perturbaciones.

También hay que prever dispositivos de manutención para mover las unidades y en algunos casos un sistema de prioridades (tal como ocurre muchas veces en los almacenes, con el fin de asegurar la rotación de las unidades).

En los servicios, cuando un usuario o cliente espera pierde tiempo y además suele tener una percepción subjetiva del tiempo de espera en virtud de la cual éste le parece más largo de lo que realmente es. Una mala organización de la cola puede generar en el usuario una sensación de ansiedad, por la posibilidad de que el turno no sea respe-

tado o, incluso, en grandes sistemas de servicios, porque no está seguro de ser tenido en cuenta (es decir, que admite la posibilidad de que el sistema lo haya olvidado).

Estas consideraciones (tan simples, por otra parte) se han de tener presentes en la organización de la cola.

El lugar donde físicamente esperan los clientes o usuarios ha de reunir condiciones adecuadas y contener elementos que permitan hacer la espera más amena; durante este tiempo de espera se puede proporcionar al usuario información sobre el servicio (lo que puede facilitar o mejorar la calidad de la prestación del mismo) o bien solicitar datos al usuario, a través de un formulario, por ejemplo. En general hay que preguntarse si el usuario, mientras espera, puede hacer algo útil para una mejor (más rápida o de mayor calidad) prestación del servicio o si durante este tiempo se le puede prestar algún otro servicio, quizá complementario con relación al que ha motivado su desplazamiento al sistema.

En los sistemas con múltiples canales indiferenciados en paralelo es preferible siempre organizar una sola cola en lugar de una para cada canal, puesto que así los tiempos medios de espera son menores y además no se producen desplazamientos de las unidades de una a otra cola; en la práctica ello exige una distribución en planta adecuada (con elementos de separación y, por así decir, canalización de la cola) o bien la asignación de números de orden a las unidades, en cuyo caso debe haber un procedimiento para informar a las unidades en la cola del número a que le corresponde el turno.

En general es muy importante que el usuario esté bien informado sobre cómo progresa la cola o que él mismo experimente este progreso (con un cambio de lugar, por ejemplo).

Cuando los tiempos de servicio son muy distintos para los diversos usuarios puede ser conveniente establecer canales separados por tipos de usuarios para evitar largas esperas a usuarios con tiempos de servicio muy breves. Si la naturaleza del servicio lo permite se puede actuar también a través de la disciplina de cola (por ejemplo, dando prioridad a los usuarios con un tiempo de servicio menor, lo cual contribuye a disminuir el tiempo medio de espera).

También es conveniente aumentar la capacidad de prestación del servicio cuando la cola aumenta más allá de cierto límite (número de canales variable, personal auxiliar en cada canal, personal que efectúa ciertas tareas preparatorias con los usuarios que forman cola, etc.).

Capítulo 2

Teoría de colas

2.1. Una formalización de los sistemas con esperas: La teoría de colas

Un sistema de colas simple consta de los canales de servicio y la cola. Se producen unas llegadas, procedentes de un centro emisor que se considera externo al sistema y las unidades que llegan ocupan un canal, si hay alguno disponible, o se incorporan a la cola; cuando el canal ha completado el tratamiento de una unidad, ésta sale del sistema.

Este sencillo esquema admite una gran cantidad de variantes. Muchas de ellas pueden designarse mediante un código de la forma:

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

que es una ampliación del propuesto inicialmente por Kendall y cuyos elementos tienen el siguiente significado:

a Ley de las llegadas al sistema.

b Ley de los tiempos de servicio.

c Número de canales en paralelo (supuestos iguales).

d Disciplina de cola.

e Número máximo de unidades que pueden encontrarse simultáneamente en el interior del sistema.

f Tamaño del centro emisor.

Para los elementos *a* y *b* los valores más utilizados son:

M Las llegadas corresponden a un proceso markoviano. Si la tasa media de llegadas es constante a lo largo del tiempo los tiempos entre llegadas siguen una ley exponencial y el número de llegadas por unidad de tiempo una ley de Poisson (de ahí que podamos referirnos indistintamente a tales llegadas con los términos exponenciales y poissonianas).

D Tiempo constante.

E Ley Erlang-k (una variable aleatoria que es suma de k variables aleatorias exponenciales independientes e idénticamente distribuidas sigue una ley Erlang-k).

GI/G Ley general.

Para el elemento *d*:

GD Disciplina general (cualquiera, sin prioridades).

FIFO Primer llegado, primer atendido.

LIFO Último llegado, primer atendido.

SIRO Al azar.

Obsérvese que ser primero en ser atendido no es equivalente a ser el primero en salir del sistema, puesto que cuando hay más de un canal una unidad puede salir del sistema después de otra aunque haya empezado a ser tratada antes.

Por lo que respecta a *f*, puede ser ∞ o N (un número natural). Ello debe interpretarse como sigue: que el centro emisor sea infinito significa que no se modifica por el hecho de que algunas unidades estén en el sistema, por lo que la ley de llegadas es independiente del número de unidades que contenga el sistema; en cambio, si el centro emisor es finito, con N unidades inicialmente, sus características dependen del número de elementos en el sistema, puesto que éste más el número de unidades en el centro emisor es constante, igual a N.

El código describe un tipo de sistema, en el cual caben infinitos sistemas específicos, según los valores que adopten el número de canales (c puede ser un valor concreto o un conjunto de valores definido por una propiedad, tal como $c > 1$) y los parámetros que caracterizan las leyes que rigen las llegadas y los tiempos de servicio.

Los parámetros que más se utilizan en la descripción de un sistema de colas son los siguientes:

λ Tasa de llegadas (número medio de llegadas por unidad de tiempo), que puede ser constante o no (en particular en los sistemas con centro emisor finito, la tasa de llegadas depende del número de unidades que en cada momento se encuentran en el centro emisor y, por consiguiente, del número de unidades en el sistema).

μ Tasa de servicio (número medio de unidades por unidad de tiempo que es capaz de tratar un canal).

s Número de canales en paralelo (supuestos, habitualmente, iguales).

A partir de λ y de μ se puede calcular:

$u = \frac{\lambda}{\mu}$ Intensidad de tráfico (corresponde al número de canales cuya capacidad media global sería igual a las llegadas medias); el número de canales en el sistema ha de ser entero en tanto que u, en general, no lo es, por lo que el número de canales mínimo para que el sistema no quede colapsado ha de ser igual al menor entero mayor o igual que u).

Y, haciendo intervenir s:

$\rho = \frac{u}{s} = \frac{\lambda}{s\mu}$ Factor de servicio. Es la proporción que representa la demanda total media por unidad de tiempo en relación con la capacidad total media del servicio. Si no hay pérdida de unidades y todos los canales se utilizan por un igual, el valor de ρ corresponde a la proporción de tiempo en que está ocupado cualquier canal. Salvo en ciertos sistemas (número de unidades en el sistema limitado, con pérdida de las unidades que llegan y que no caben), el factor de servicio ha de ser menor que la unidad (un factor de servicio mayor que la unidad implica, en general, colas que crecen indefinidamente o pérdida de unidades; un factor de servicio igual a la unidad corresponde al caso en que la demanda media coincide con la capacidad de servicio media -en este caso, el azar, la dispersión en los tiempos entre llegadas o en el tiempo de servicio, puede llegar a producir colas muy largas -). Aunque pueda parecer innecesario advertimos que no se ha de confundir el nivel de servicio (al que ya nos hemos referido anteriormente) con el factor de servicio (aunque, para un mismo tipo de sistema el nivel de servicio crece a medida que disminuye ρ, la relación no es lineal y, por otra parte, sistemas distintos con los mismos valores de ρ pueden tener niveles de servicio muy distintos).

En general, con los modelos de sistemas en que se producen fenómenos de espera se pretende calcular alguno o varios de los valores siguientes:

- Probabilidad de que en el interior del sistema haya un cierto número de unidades, p_n.
- Probabilidad de que un canal esté libre y probabilidad de que una unidad que llega al sistema encuentre un canal disponible para ella y, por consiguiente, no tenga que esperar. Ambas probabilidades son distintas, pero las tratamos en el mismo párrafo porque se puede pensar que coinciden en los sistemas con un solo canal y conviene aclarar que, en general, ello no es cierto: por ejemplo, si en un sistema con un canal y con un tiempo de servicio constante igual a una unidad de tiempo llegan dos unidades seguidas cada 3 unidades de tiempo, el canal está libre $1/3$ del tiempo y este valor es la probabilidad de encontrarlo libre si hacemos una observación al azar; en cambio, la probabilidad de que una unidad elegida al azar encuentre el canal disponible es igual a 0.5; ahora bien, si la ley que rige las llegadas es exponencial, la probabilidad de que una unidad no tenga que esperar es igual a la probabilidad de que el canal este vacío (ésta es una propiedad peculiar de las llegadas exponenciales a la que se ha denominado PASTA: *Poisson Arrivals See Time Averages*).
- el valor medio del número de unidades en el sistema, del número de unidades en la cola, del tiempo de estancia de una unidad en el sistema o en la cola, que se designan, respectivamente, por:

$$L, L_q, W, W_q$$

- la distribución del tiempo de estancia en el sistema o en la cola y, por tanto, la probabilidad de que una unidad permanezca en uno u otra más de un cierto tiempo.

- los costes medios por unidad de tiempo (de funcionamiento de los canales, de la espera de las unidades, de desplazamiento, totales).

Estos valores dependen del tiempo que ha transcurrido desde el instante que se considera inicial y del estado del sistema en ese instante; en general hay, pues, un *régimen transitorio*. En los sistemas bien dimensionados, la probabilidad de que el sistema adopte uno u otro estado tiende a un valor constante a medida que aumenta el tiempo transcurrido desde el instante inicial; en estos casos el sistema tiende a lo que se denomina *régimen permanente*, tanto más rápidamente cuanto menor es el factor de servicio.

En síntesis, la teoría de colas modeliza los sistemas de espera por medio de sistemas de ecuaciones diferenciales que proporcionan la probabilidad de cada estado del sistema en función del tiempo transcurrido desde el instante inicial y de la situación correspondiente al mismo.

Generalmente estos sistemas de ecuaciones son difíciles de resolver, salvo por procedimientos de cálculo numérico, especialmente cuando se pretende describir el régimen transitorio. En el caso del régimen permanente hay algunos casos fáciles, para los que hay fórmulas o algoritmos muy sencillos que proporcionan con poco esfuerzo los valores deseados; para algunos modelos hay incluso tablas y gráficos o programas de ordenador que facilitan aún más el trabajo.

Las características de funcionamiento del sistema que pueden calcularse fácilmente dependen del tipo de sistema. En los más sencillos se puede calcular incluso la distribución de los tiempos de estancia de las unidades en el sistema o en la cola o la probabilidad de que en el sistema se encuentre un número cualquiera de unidades; en otros sólo se puede calcular fácilmente L, L_q, W, W_q.

Al respecto conviene tener en cuenta que entre estos cuatro valores existen relaciones tales que, dado uno de ellos, se puede calcular muy fácilmente los otros tres:

- Por una parte, el tiempo medio de estancia en el sistema es igual al tiempo medio de estancia en la cola más el tiempo medio de servicio, por lo cual:

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

- Por otra, se cumple (fórmulas de Little) que:

$$L = \lambda W \text{ y } L_q = \lambda W_q$$

- expresiones en que, cuando λ no es constante debe substituirse por su valor medio, $\bar{\lambda}$:

$$L = \bar{\lambda} W \text{ y } L_q = \bar{\lambda} W_q$$

Dicho valor medio se refiere a las unidades que se incorporan efectivamente al sistema, que no deben confundirse con las que salen del centro emisor, puesto que en algunos sistemas no todas ellas llegan a entrar en el sistema (sistemas con capacidad limitada o, más en general, sistemas en que, en función del estado de los mismos, hay una

probabilidad no nula de que una unidad que pretende entrar desista de hacerlo o sea rechazada).

Intuitivamente se puede justificar la fórmula de Little considerando la evaluación, por dos vías distintas, de la media de la suma de los tiempos de estancia en el sistema (o en la cola) durante un tiempo muy largo, T; por una parte esta media es igual a LT (a lo largo de una unidad de tiempo, el número medio de unidades en el sistema es L); por otra, es igual al número medio de unidades llegadas al sistema, $\bar{\lambda}T$, por el tiempo medio de estancia en el sistema de las unidades consideradas individualmente, W, por lo que el tiempo medio de estancia del conjunto es $\bar{\lambda}WT$; de la igualdad entre LT y $\bar{\lambda}WT$ se deduce $L = \bar{\lambda}W$. El mismo razonamiento se puede aplicar a una parte del sistema; si esta parte es la cola se deduce análogamente que $L_q = \bar{\lambda}W_q$.

Los modelos más sencillos y difundidos son los correspondientes a procesos de llegada poissonianos y a tiempos de servicio con distribución exponencial, con uno o varios canales idénticos en paralelo y con cola finita o infinita y con centro emisor finito o infinito. También es muy conocido el modelo con ley de llegadas exponencial y un solo canal con un tiempo de servicio con distribución de probabilidad cualquiera (con la condición de que el tiempo de servicio de una unidad no dependa del de las demás). Las fórmulas y en su caso tablas y gráficos correspondientes a estos modelos, para el *régimen permanente* se encuentran en las introducciones a la teoría de colas que se encuentran en cualquier texto de introducción a la Investigación Operativa. Se han publicado también gráficos para el modelo $M/D/s : GD/\infty/\infty$ y tablas correspondientes a modelos con leyes de llegadas y de servicio distintas de la exponencial (concretamente, leyes que son combinaciones lineales de leyes exponenciales) lo que amplía las posibilidades de aplicación de la teoría de colas; algunos paquetes de software especializados incluyen otros modelos. El apéndice A recoge fórmulas correspondientes al modelo más sencillo (llegadas y servicio exponenciales, un solo canal, con tasa de llegadas constante y sin limitación en la capacidad del sistema para albergar unidades).

La utilización de estos modelos es rápida y barata. Pero son una aproximación a veces muy grosera de la realidad, por dos motivos, principalmente: por una parte las leyes que siguen las llegadas y los tiempos de servicio en el sistema real pueden no ajustarse a las de los modelos asequibles; por otra, el comportamiento de muchos sistemas puede que no llegue a aproximarse siquiera al que correspondería al régimen permanente, bien sea porque no funcione durante un tiempo suficientemente largo sin interrupción, bien porque las tasas de llegadas y de prestación del servicio sean variables a lo largo del tiempo.

Si se tiene conciencia de estas limitaciones, la teoría de colas es un instrumento útil y fácil de aplicar para hacerse una primera idea de las características del funcionamiento de un sistema o para comparar diversas soluciones.

A continuación se estudian con mayor detalle el modelo denominado proceso de nacimiento y muerte.

2.2. Procesos de nacimiento y muerte

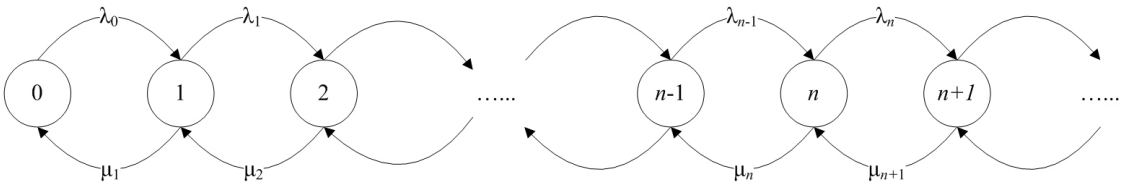
Los modelos a que se ha hecho referencia en el punto anterior, con llegadas poissonianas y tiempos de servicio exponenciales son casos particulares de un modelo más general que es el proceso de nacimiento y muerte.

Se puede describir este modelo, de un modo relativamente informal, a partir de las siguientes hipótesis:

1. La probabilidad de que se produzca una llegada (o nacimiento) en un intervalo dt es igual a $\lambda_n dt$, donde λ_n es constante, dado n, y no depende, por consiguiente, del tiempo transcurrido desde la última llegada.
2. La probabilidad de que se produzca una salida o muerte en un intervalo dt es igual a $\mu_n dt$, donde μ_n es constante, dado n, y no depende, por consiguiente, del tiempo transcurrido desde el comienzo del servicio.
3. La probabilidad de que en un intervalo dt se produzca más de un acontecimiento es un infinitésimo de orden superior.

Se puede decir, pues, que el estado del sistema queda definido, por el número de unidades que contiene el mismo, n, y que los cambios de estado se producen cuando hay un nacimiento o llegada (el sistema pasa del estado n al estado $n+1$) o una muerte o salida (el sistema pasa del estado $n+1$ al n).

El esquema de transiciones entre estados se puede representar mediante un grafo como el siguiente:



Obsérvese que en régimen permanente la probabilidad de entrar en un estado en un intervalo dt ha de ser igual a la probabilidad de salir del mismo (puesto que el régimen permanente se define precisamente como aquella situación en que la probabilidad del sistema se encuentre en un cierto estado es independiente del tiempo, lo que no sucedería si la probabilidad de entrada fuera distinta de la de salida). Por consiguiente, si consideramos un estado cualquiera, n, se ha de cumplir que:

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

y, para $n = 0$:

$$p_1\mu_1 = p_0\lambda_0$$

de donde:

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

por lo que $\forall n$ se puede expresar p_n en función de p_0 y teniendo en cuenta que la suma de las probabilidades de todos los estados ha de ser igual a 1, calcular todas las probabilidades, a partir de las cuales se puede determinar fácilmente otros parámetros, como L, por ejemplo.

Las hipótesis que definen el proceso de nacimiento y muerte corresponden a un proceso sin memoria (la probabilidad de que ocurra algo sólo depende del estado actual del sistema y no de su historia, de la trayectoria a través de la cual se ha alcanzado dicho estado). En muchos sistemas, las llegadas se producen de forma independiente, por lo que la hipótesis puede ajustarse bastante a la realidad; en lo que respecta a las salidas es más difícil que se produzca esta concordancia.

Se demuestra que en un proceso de nacimiento puro (sólo llegadas), la distribución de los tiempos entre llegadas es exponencial. Lo mismo sucede en un proceso de muerte pura en relación con los tiempos de servicio.

Particularizando el diagrama y las ecuaciones de los procesos de nacimiento y muerte se obtienen las expresiones correspondientes a los modelos $M/M/s$ con capacidad finita o infinita y con centro emisor finito e infinito.

2.3. Introducción a las redes de colas

En los sistemas productivos es frecuente que un sistema de colas se relacione con otros sistemas de colas, constituyendo así, en conjunto, lo que se denomina una red de colas. El flujo de salida de uno o más sistemas de colas puede constituir el flujo de entrada a otro sistema; el flujo de salida de un sistema puede salir de la red, dirigirse a otro sistema, o dividirse y dirigirse a destinos diversos.

Es obvio que el tratamiento de las redes de colas (que se ha desarrollado fuertemente en los últimos años, inicialmente en tomo a las redes de informática y de comunicaciones) es complejo, en general.

Algunas redes de colas, no obstante, admiten un tratamiento sencillo.

En primer lugar, si a un sistema de colas llega un flujo exponencial, sin pérdidas, y el tiempo de servicio es asimismo exponencial, el flujo de salida también lo es, por lo cual los sistemas de colas en serie con tales características pueden tratarse sin dificultades con las fórmulas correspondientes a los sistemas simples.

Por otra parte la mezcla de flujos exponenciales es también exponencial con un parámetro igual a la suma de los que corresponden a los flujos componentes.

Algo parecido ocurre cuando un flujo exponencial se divide, pero sólo cuando se divide al azar; en este caso los flujos resultantes son exponenciales con un parámetro igual al del flujo inicial multiplicado por la proporción que corresponde al flujo de salida. Pero si la división no es al azar, la distribución de los flujos de salida ya no es exponencial; considérese el caso de un flujo que se divide al 50 % en otros dos: si las unidades

se asignan alternativamente al uno o al otro, el tiempo entre unidades en cada uno de ellos sigue una ley Erlang-2.

Las observaciones precedentes permiten tratar algunas redes de colas por medio de las fórmulas correspondientes a sistemas simples.

Otra posibilidad, ya más compleja, para el tratamiento de redes de colas es la generalización del modelo de proceso de nacimiento y muerte.

En este texto se ha descrito sucintamente dicho modelo. Una de sus características es que el estado del sistema queda definido por el número de elementos en el mismo, n. Si en cada sistema de colas simple de los que constituyen una red se cumplen las hipótesis del proceso de nacimiento y muerte, el estado del sistema queda definido por el número de elementos en cada sistema simple y, en algunos casos, por el hecho de que un sistema simple está bloqueado o no (una estación de trabajo de una línea de montaje puede haber terminado las tareas que tiene asignadas y quedar bloqueada por el hecho de que la estación siguiente no ha terminado y no hay espacio de almacenamiento intermedio). Una vez establecidos los estados que el sistema puede adoptar se puede dibujar el grafo de transiciones y, para cada estado, la ecuación que expresa la igualdad de las probabilidades de entrar y salir de un estado, en régimen permanente; estas ecuaciones, junto con la condición de que la suma de probabilidades de los estados es igual a 1, forman, cuando el conjunto de estados es finito, un sistema de ecuaciones lineales (con una ecuación redundante), cuya resolución permite determinar los valores de las probabilidades de los estados.

2.4. Aplicación de los modelos de colas al diseño de sistemas

Los modelos de colas se pueden utilizar para determinar las características de funcionamiento de un sistema dado o para diseñar un sistema de acuerdo con determinados objetivos.

En este último caso se tratará de establecer una relación de soluciones alternativas, algunas de las cuales pueden tener un parámetro indeterminado, tal como el número de canales.

En algunos casos el objetivo será minimizar el coste, teniendo en cuenta los costes de funcionamiento, los de desplazamiento y los de espera. Para evaluar estos últimos deberemos conocer cuanto cuesta la espera de una unidad durante una unidad de tiempo; si este coste es C para todas las unidades los costes de espera en el sistema durante una unidad de tiempo son iguales a LC (o a $\bar{\lambda}WC$, por la fórmula de Little).

En otros casos, el objetivo será alcanzar un determinado nivel de servicio con un coste mínimo (muchas veces ello corresponderá a la determinación del número mínimo de canales con que se puede alcanzar el nivel de servicio prefijado). Ello supone que se ha de definir qué parámetro o parámetros se consideran representativos del nivel de servicio y cuáles son los valores mínimos o máximos que se desea que adopten dichos

parámetros.

2.5. Estimación de los parámetros de un sistema de colas

Esta es una cuestión que ni siquiera se menciona en muchos textos pero de considerable importancia práctica.

Para aplicar un modelo de colas se necesita información sobre las leyes de llegadas y de servicio y, al menos, el valor de los parámetros λ y μ. Evidentemente, estas informaciones proceden del sistema real y éste, en general, se comporta aleatoriamente y sólo cabe estimar los datos a partir de una muestra. Con ella se puede verificar si algo se opone a las hipótesis sobre el tipo de ley y estimar los parámetros, pero no se ha de olvidar que dispondremos de estimaciones, con un cierto intervalo de confianza y que, por consiguiente, los valores de ρ, L, etc. tendrán asociado asimismo un intervalo de confianza; si la muestra no es muy grande, los intervalos de confianza de los parámetros calculados con el modelo de colas pueden ser muy amplios, por lo que, en este caso, resulta muy aventurado sacar conclusiones.

Así pues, en las aplicaciones es necesario calcular los intervalos de confianza para la muestra disponible o determinar qué tamaño de muestra es necesario para obtener las estimaciones con una precisión dada.

Capítulo 3

Simulación

3.1. Concepto y clasificaciones

La Investigación Operativa se puede definir, sucintamente (y, por tanto, esquemáticamente) como la disciplina que aborda los problemas organizativos mediante el uso de modelos simbólicos o matemáticos.

Dichos modelos responden a la siguiente formalización:

$$X \in E(Y)$$
$$V = g(X, Y)$$
$$U = f(X, Y, V)$$

donde:

Y conjunto de parámetros o variables exógenas; corresponden a magnitudes sobre cuyo valor no se puede actuar pero que influyen en el comportamiento del sistema.

X conjunto de variables de acción, es decir, aquellas cuyo valor puede gobernarse dentro de ciertos límites, que dependen del conjunto de parámetros, tal como indica la expresión $X \in E(Y)$

V conjunto de variables cuyo valor resulta de los adoptados por Y y X, a través de la relación g

U conjunto de variables de evaluación del comportamiento del sistema, que depende de Y, X y V, a través de f.

En los modelos de optimización el objetivo es hacer máxima o mínima una función de U, $z(U)$, es decir, dado Y, determinar $X*$ que optimice dicha función.

A veces, las características de U o las del modelo no permiten plantear una $z(U)$ o calcular X^*; en ocasiones, simplemente no se desea optimizar sino conocer y evaluar el comportamiento del sistema para unos valores determinados de Y y de X. En este caso, se introducen en el modelo los valores supuestos con el fin de calcular V y U; se estudia y evalúa el comportamiento del sistema a través del modelo: se puede decir que se *simula* dicho comportamiento. Esta es una acepción muy amplia del término

simulación, que comprende, por ejemplo, la utilización de un modelo de colas para evaluar el tiempo medio de espera de las unidades, dados unos valores de los parámetros. No obstante, el uso del término *simulación* se suele reservar sólo para aquellos casos en que se pretende estudiar el comportamiento del sistema *a lo largo del tiempo*.

El sistema real que se desea estudiar puede ser determinista o aleatorio; obviando la discusión sobre si el carácter aleatorio es inherente al sistema o expresa simplemente un conocimiento insuficiente del mismo por parte el observador, diremos que un sistema es aleatorio cuando los valores de alguna o algunas variables no pueden determinarse con certeza y únicamente cabe asignar probabilidades a una cierta gama de valores.

Por su parte, el modelo utilizado para simular puede ser también determinista o aleatorio; se dará este último caso cuando al menos una de las variables del modelo tiene carácter estocástico.

Potencialmente hay, pues, cuatro tipos de simulación según el carácter determinista o aleatorio del sistema o de su entorno y del modelo. De estos cuatro tipos, aquí nos centraremos en los tres siguientes:

1. Sistema determinista / modelo determinista.

   Una vez establecido el modelo, basta organizar los cálculos para obtener los valores de las variables a lo largo del tiempo. La dificultad puede residir en la naturaleza de estos cálculos (integración de ecuaciones diferenciales en derivadas parciales, por ejemplo) o en el procedimiento para determinar el avance del reloj o variable que expresa el transcurso del tiempo.

2. Sistema aleatorio / modelo determinista.

   Se utiliza muchas veces este tipo de simulación en problemas en que el entorno (variables Y) se comporta aleatoriamente. El procedimiento consiste en aplicar el modelo para varios supuestos del valor de Y.

3. Sistema aleatorio / modelo aleatorio.

   Es la simulación más típica, hasta el punto que muchos textos sobre simulación sólo se refieren a ella.

La combinación sistema determinista / modelo aleatorio se suele utilizar en otros ambientes. Se puede hacer uso del azar para estimar un valor (el de π mediante la aguja de Buffon, por ejemplo, o el de una integral, que puede corresponder a un área o volumen), pero ello no corresponde a ninguna de las definiciones de simulación discutidas más arriba.

Desde luego, no es ésta la única clasificación posible. Es interesante también la distinción entre los sistemas cuyo estado se modifica o puede modificarse continuamente y aquellos en que sólo se modifica en instantes singulares (o en que nos basta, para describir su comportamiento con indicar su estado en ciertos instantes singulares); po-

demos distinguir entonces la simulación *continua* y la simulación *discreta*. Esta última es la que casi siempre se aplica en los problemas de organización y de gestión.

3.2. Aplicaciones de la Simulación

En cualquier caso, la simulación puede utilizarse para:

- Predecir el comportamiento de un sistema.
- Evaluar el rendimiento de un sistema funcionando en condiciones especificadas.
- Analizar la sensibilidad del funcionamiento el sistema a variaciones de las variables en un entorno.
- Establecer relaciones entre diversas magnitudes asociadas al sistema.
- Comparar políticas de gestión y determinar la política óptima (en el conjunto de las previamente especificadas).

3.3. Gestión del reloj en la simulación discreta

A costa de una cierta simplificación, se puede decir que hay dos procedimientos básicos para la gestión del reloj, a saber, intervalos fijos e intervalos variables (simulación *síncrona* y simulación *asíncrona*).

En el primero de estos procedimientos se ha de determinar la magnitud del intervalo. Ésta en algunos casos se impone de forma natural como consecuencia de las características del problema. En otros problemas existe una cierta gama en la que elegir; un intervalo demasiado pequeño obliga a hacer cálculos o anotaciones innecesarios, uno demasiado grande, no permite reflejar con la precisión deseada el comportamiento del sistema.

En la simulación asíncrona, el reloj sólo marca los instantes en que se producen cambios de estado del sistema. Ello supone definir, teniendo en cuenta los objetivos de la simulación, qué se entiende por estado del sistema y establecer qué tipos de acontecimientos producen cambios de estado; en todo momento, a lo largo de la simulación, se ha de prever en qué instante se producirá el primer acontecimiento de cada tipo (lo que puede hacerse según el caso, a partir de los datos del problema -por ejemplo, porque se sabe que las unidades llegan a intervalos de cuatro minutos - o a partir de informaciones generadas en la propia simulación -por ejemplo, una vez se ha determinado en qué instante comienza un tratamiento se puede determinar, si no hay posibilidad de interrupción, en qué instante terminará-): el menor de estos instantes es aquel en que se producirá el próximo cambio de estado del sistema (una forma de realizar la simulación consiste en establecer una cola de acontecimientos). Determinando los instantes de cambio de estado y el estado del sistema inmediatamente después del cambio se puede reconstruir, para los objetivos perseguidos, toda la evolución del sistema a lo largo del tiempo. Este procedimiento habitualmente es algo más laborioso y complejo que el de intervalo constante, pero la simulación, en general, resulta mucho más compacta; además permite seguir con todo detalle sin errores, la evolución del sistema,

cosa que no sucede en la simulación síncrona, salvo que se discreticen los tiempos y se utilice un intervalo suficientemente pequeño o suceda que todos los tiempos son múltiplos de un cierto valor, como en los ejemplos anteriores.

3.4. Problemas específicos que presenta la simulación aleatoria

Las consideraciones anteriores sobre la gestión del reloj son comunes a la simulación con modelos deterministas o con modelos aleatorios.

Lo específico de la simulación aleatoria es el carácter estocástico de las entradas y de las salidas. Ello plantea dos problemas:

¿Cómo obtener valores de las variables aleatorias de entrada?

¿Cómo evaluar las salidas?

En un modelo determinista, cada variable, para cada instante, tiene asociado un valor; en un modelo aleatorio, las variables estocásticas no tienen asociado un valor sino una distribución de probabilidad. Al realizar la simulación se ha de tomar un valor numérico para la variable; este valor no se puede elegir de cualquier manera, no basta con que sea verosímil: si repetimos la simulación o si en el curso de la simulación se utiliza repetidamente una variable, la sucesión de valores utilizados ha de ser una muestra de observaciones de la variable; es decir, la probabilidad de utilizar un valor de una variable ha de ser la que corresponda según la distribución que se supone para la misma. El problema reside, pues, en cómo obtener muestras de las variables aleatorias.

Por otra parte, de lo dicho se deduce que los resultados de la simulación son una muestra, más o menos amplia, de las variables aleatorias que expresan el comportamiento del sistema. La muestra permite hacer estimaciones, pero éstas llevan asociado un intervalo de confianza que, si la muestra es pequeña, puede ser muy grande. Surgen entonces dos cuestiones; primera: ¿cómo efectuar las estimaciones y establecer los correspondientes intervalos de confianza?; segunda: ¿cómo conseguir que los intervalos de confianza sean menores, para un tamaño de muestra dado?.

3.5. Bases para la obtención de muestras de variables aleatorias

Supóngase que disponemos de una fuente de dígitos aleatorios (del 0 al 9, con probabilidad 1/10 para cada valor y con extracciones independientes), como por ejemplo un bombo de los que se utilizan en la lotería.

Es fácil entonces obtener muestras de una variable aleatoria discreta cualquiera, con toda la precisión que se desee. Supóngase que la variable x puede tomar los valores x_1 y x_2 con probabilidades 0.6 y 0.4, respectivamente; para obtener muestras de esta variable basta tomar dígitos de uno en uno y asignar el valor x_1 o el x_2 según que el

dígito esté comprendido en el intervalo 0-5 o en el 6-9, respectivamente; la generalización de este procedimiento es obvia para variables que puedan tomar un número cualquiera, finito, de valores, cuyas probabilidades se expresen mediante un mayor número de decimales (esto último exige utilizar para cada muestra de la variable un grupo de tantos dígitos como decimales; ejemplo si x puede tomar los valores x_1 y x_2 con probabilidades 0.874 y 0.126, tomaremos grupos de tres dígitos y los compararemos con los intervalos 000-873 y 874-999).

Este procedimiento se puede aplicar gráficamente, representando la probabilidad acumulada de la variable y determinando qué valor de la misma corresponde al grupo de dígitos, interpretado como la parte decimal de un número comprendido entre 0 y 1 (ver figura 3.5 (izquierda)). Se comprende, a partir de esta representación gráfica, la generalización del procedimiento a variables continuas (ver figura 3.5 (derecha)), para las cuales se utiliza la función de distribución, $F(x)$ y, dada una muestra, y, de una variable distribuida uniformemente entre 0 y 1 se determina qué valor de x cumple:

$$y = F(x)$$

es decir:

$$x = F^{-1}(y)$$

(para algunas funciones F se puede despejar la variable x; en otras, esto no es posible y el valor de x que corresponde a un valor específico de y se determinará numéricamente).

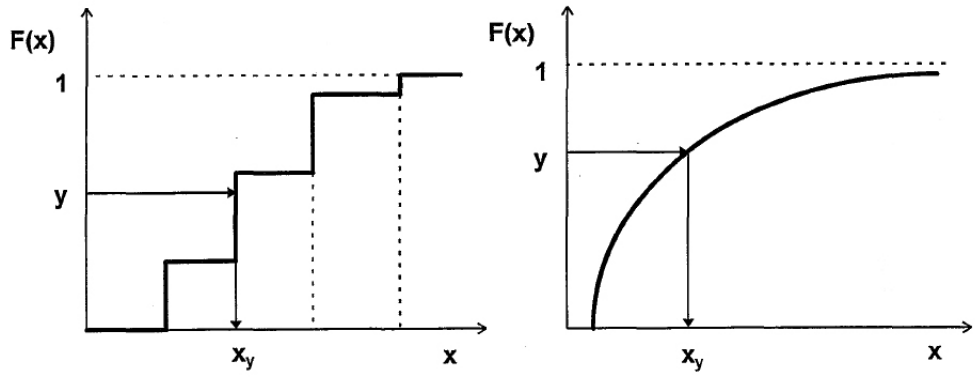


Figura 3.1: Parte decimal (izquierda) y variable continua (derecha)

Se dispone, pues, de un procedimiento de generación de muestras de cualquier variable aleatoria, a condición de tener una fuente de dígitos aleatorios o de valores distribuidos uniformemente entre 0 y 1. El problema de obtención de muestras se ha simplificado, pues, muchísimo, pero dista de ser trivial.

Una fuente de dígitos aleatorios puede ser un dispositivo fisico dotado de cierta simetría que garantice la equiprobabilidad de los valores (por ejemplo, un bombo o una ruleta); por supuesto, en general estos dispositivos son lentos y difícilmente conectables con un ordenador. A partir de un dispositivo como los mencionados se puede establecer tablas, pero entonces los dígitos contenidos en ellas y la secuencia en que

aparecen en las mismas serán invariables y, por consiguiente, no se podrá hacer un uso repetido de los mismos.

John von Neumann propuso el método que ha sido adoptado de forma general: generar a través de un procedimiento determinista dígitos cuya frecuencia y cuya secuencia tengan propiedades parecidas a las de los números aleatorios, es decir, lo que se denomina números pseudoaleatorios.

La forma en que el propio von Neumann concretó el método resultó muy poco afortunada, pero ilustra la idea básica; se trata del procedimiento denominado de los cuadrados medios: tomamos un número de cuatro cifras, por ejemplo, lo elevamos al cuadrado y retenemos las cuatro cifras centrales del resultado y repetimos este proceso cuantas veces sea necesario. La tabla 3.1 incluye los grupos de 4 dígitos obtenidos a partir de 3456; en la propia tabla puede observarse fácilmente que los dígitos así generados no tienen propiedades parecidas a los aleatorios: a partir del grupo 6100 se produce un ciclo muy corto en el que además el 0 aparece con una frecuencia del 50 % (por supuesto, la aparición de ciclos más o menos largos, es inevitable ya que el número de valores distintos que se pueden formar con un grupo de n dígitos, en base 10, es 10 elevado a n).

Tabla 3.1: Generación de números psudoaleatorios con s_0 = 3456; s_0^2 =11943936; s_1= 9439

3456	166	0609
9439	275	3708
947	756	7492
8968	5715	1300
4250	6612	6900
625	7185	6100
3906	6242	2100
2568	9265	4100
5946	6406	8100
3549	368	6100
5954	1354	2100
4501	8333	4100
2590	4388	8100
7081	2545	6100
1405	4770	2100
9740	7529	4100
8676	6858	8100
2729	321	6100
4474	1030	2100

Pese a ello, el método general subyacente a la propuesta de Von Neuman era interesante y práctico. El procedimiento que se suele utilizar en los generadores de números aleatorios de los lenguajes de programación, especializados o no en simulación, se denomina congruencial. Se toma un valor inicial o semilla s_0 y a partir del mismo se calcula una sucesión de valores según la expresión:

$$s_{i+1} = (a + bs_i) \bmod L$$

es decir, s_{i+1} es el resto en la división por L de $a+bs_i$; a cada s_i le corresponde un valor, r_i, comprendido entre 0 y $(L-1)/L$:

$$r_i = \frac{s_i}{L}$$

Habitualmente, L es el máximo valor que se puede representar en un cierto tipo de variable del ordenador (por ejemplo, para un ordenador de 16 bits, L=32767). No todos los valores de los parámetros a y b en la expresión correspondiente al cálculo de las s_i son adecuados, pero se pueden elegirlos de modo que el ciclo sea el más largo posible (L valores). Pero el procedimiento es determinista, el ciclo se repite inexorablemente y, dada la semilla s_0 la sucesión de valores está determinada.

Como se ha dicho, se desea que los números pseudoaleatorios tengan propiedades parecidas a los números aleatorios: las probabilidades de los valores de cada dígito son iguales e independientes de las de los otros dígitos. Ello se puede estudiar (en el sentido habitual en estadística: hacemos pruebas para ver si algo se opone a la hipótesis que hemos formulado) mediante pruebas muy variadas. Desde luego, lo primero es comprobar (mediante una prueba de χ^2 o de Kolmogorov-Smirnov) que nada se opone a la hipótesis de que los valores son equiprobables (pero secuencias como 0123456789 ... repetidas indefinidamente superan perfectamente ambas pruebas). Más complejo resulta comprobar que nada se opone a la independencia entre unos y otros dígitos; una prueba adecuada para esta propiedad se describe a continuación:

Dada la sucesión de valores r_i ($i = 1, ...N$) se determina su mediana y se cuentan las ráfagas contenidas en la sucesión. Una ráfaga es una subsecuencia tal que todos sus elementos tienen un valor superior (o inferior) al de la mediana (los valores iguales al de la mediana se pueden agrupar con los superiores o con los inferiores, según se convenga, pero teniendo en cuenta que el test se basa en el supuesto de que el número de valores inferiores y el de valores superiores es el mismo: esto puede obligar a suprimir alguna de las observaciones cuyo valor coincida con el de la mediana). Sea R el número de ráfagas y n el número de valores por debajo de la posición de la mediana; si el valor de R es muy bajo ello significa que la probabilidad de que a un valor bajo (o alto) le siga otro valor bajo (o alto) es superior a la que corresponde a la independencia entre valores consecutivos; si el valor de R es muy alto, significa que los valores de r_i fluctúan (a un valor alto le sigue otro bajo y viceversa). El valor de R sigue aproximadamente, para valores de N superiores a 10, al menos, una ley normal de parámetros:

$$E(R) = n + 1; \sigma^2(R) = \frac{n(n-1)}{2n-1}$$

Otra prueba parecida se basa en la distribución del número de ráfagas de otro tipo (subsecuencias de valores crecientes o decrecientes). Para valores de N grandes, la distribución de este número tiende a una ley normal de media $(2N-1)/3$ y variancia $(16N-29)/90$.

3.6. Obtención de muestras de variables aleatorias

El método general descrito en el punto anterior es aplicable a cualquier distribución; para algunas, no obstante, puede ser preferible utilizar otros procedimientos basados en propiedades específicas.

3.6.1. Método de la transformada inversa: Aplicación a las leyes uniforme y exponencial

En algunos casos, la aplicación del método general descrito en el punto anterior, permite obtener fórmulas sencillas. Es lo que ocurre con las leyes exponencial y uniforme:

$$\text{Ley exponencial: } x = -\frac{1}{\lambda}\ln(1-y) = -\frac{1}{\lambda}\ln y'$$

$$\text{Ley uniforme en } [a,b]\text{: } x = a + (b-a)y$$

3.6.2. Método de composición: Aplicación a las leyes binomial, de Poisson, Erlang-k y $\chi 2$

Algunas variables aleatorias son suma de otras variables aleatorias; la variable suma se puede simular a partir de muestras de las variables sumandos.

- Ley binomial (n, ω): La variable correspondiente al número de sucesos de cierto tipo en n experimentos, en cada uno de los cuales el suceso tiene probabilidad ω, se puede considerar como la suma de n variables que pueden tomar únicamente los valores 0 y 1, con probabilidades $1-\omega$ y ω, respectivamente. La simulación de dichas variables es muy sencilla, dado el número aleatorio, y, se compara con ω: si $y \leq \omega$ se atribuye a la variable el valor 1 (y el valor 0 en caso contrario).

- Ley de Poisson: La suma de dos leyes de Poisson es una ley de Poisson cuyo parámetro es la suma de los parámetros de las variables sumandos. Esta propiedad puede permitir la simulación de una ley de Poisson de parámetro cualquiera a partir de tablas correspondientes a leyes de Poisson de parámetros determinados.

- Ley Erlang-k: Puesto que una variable que sigue una ley Erlang-k es la suma de k variables independientes e idénticamente distribuidas según una ley exponencial, si μ es la media de la variable se puede simular por medio de la expresión siguiente:

$$-\frac{\mu}{k}\sum_{i=1}^{k}\ln y_i = -\frac{\mu}{k}\ln\prod_{i=1}^{k} y_i$$

- Ley χ^2 con n grados de libertad: La suma de los cuadrados de n variables independientes distribuidas según una ley normal (0,1) sigue una ley de χ^2 con n grados de libertad, de lo que se desprende inmediatamente un procedimiento para obtener muestras de dicha ley a partir de muestras de una ley normal centrada y reducida.

3.6.3. Método de las transformaciones equivalentes: Aplicación a la ley de Poisson

A veces existe una relación entre dos distribuciones de probabilidad que permite obtener muestras de una variable que sigue una de ellas a partir de una simulación de la otra. Es lo que ocurre con las leyes exponencial y de Poisson: si los intervalos entre realizaciones de cierto suceso obedecen a una distribución exponencial, el número de realizaciones del suceso en un intervalo de tiempo dado sigue una ley de Poisson (si el tiempo medio entre sucesos es $1/\lambda$, el número medio de sucesos por unidad de tiempo es λ); por consiguiente, se puede obtener una muestra de una ley de Poisson de parámetro λ simulando una ley exponencial del mismo parámetro, tantas veces como sea necesario para que la suma de los valores obtenidos supere la unidad: dicho número de veces menos 1 es la muestra de la ley de Poisson.

3.6.4. Obtención de muestras de la ley normal

Las muestras de una variable z que sigue una distribución $N(\mu, \sigma)$ se obtienen a partir de muestras de una variable, t, que sigue una ley normal centrada y reducida, $N(0, 1)$:

$$z = \mu + t\sigma$$

Para obtener muestras de t se puede hacer uso de tablas de la función de distribución de la ley normal centrada y reducida, interpolando cuando sea necesario. También se puede hacer uso de aproximaciones a la función inversa de la de distribución.

Existe asimismo la posibilidad de aplicar el teorema central del límite, según el cual la suma de variables aleatorias independientes tiende a seguir una ley normal (cuyas media y variancia son la suma de las medias y variancias de las variables sumandos) cuando aumenta el número de sumandos; la aproximación es mejor si las variables sumandos tienen distribuciones simétricas. La distribución más fácil de simular es la uniforme en el intervalo [0,1], de media 0,5 y variancia 1/12; la suma de k variables con una distribución de este tipo, si k es suficientemente grande, se aproxima a una ley normal de media $k/2$ y variancia $k/12$: con $k = 12$ la aproximación es buena, la media vale 6 y la variancia (y la desviación tipo), 1. Así pues, se puede aproximar una ley normal centrada y reducida restando 6 de la suma de 12 muestras independientes de una ley uniforme en [0,1]; evidentemente, todos los valores así obtenidos pertenecerán al intervalo [-6,6]; la probabilidad de obtener valores fuera de este intervalo es estrictamente nula, cosa que no sucede con la ley normal, que toma valores en $(-\infty, \infty)$.

Otro procedimiento para obtener muestras de una ley normal centrada y reducida son las fórmulas de Box- Müller, que se obtienen a partir de una ley normal de dos variables expresadas en coordenadas polares. Las fórmulas son:

$$x' = \sqrt{-2\ \ln\ y'}\ \sin\ 2\pi y''; x'' = \sqrt{-2\ \ln\ y'}\ \cos\ 2\pi y''$$

donde y' e y'' son dos números aleatorios y x' y x'' dos muestras, independientes, de una variable normal centrada y reducida. Para aplicar las fórmulas de Box-Müller con números pseudoaleatorios, y' e y'' deben proceder de dos generadores distintos, ya que si son dos extracciones sucesivas de un mismo generador, la relación que existe entre ellos induce propiedades indeseables en los pares x', x''.

3.7. Análisis de los resultados de una simulación

Lo más delicado de una simulación aleatoria es el análisis de los resultados. Por las complicaciones técnicas que presenta, alguno textos pasan por esta cuestión como sobre ascuas, y la consecuencia es que muchos estudios de simulación llegan a conclusiones incorrectas o insuficientemente fundamentadas.

A este respecto hay que distinguir si deseamos estudiar el comportamiento del sistema en régimen transitorio o en régimen permanente. Según veremos a continuación, en la simulación ocurre lo contrario que en la teoría de colas; ésta nos proporciona modelos con los que resulta bastante sencillo el estudio del régimen permanente y resulta muy poco útil para el régimen transitorio; en cambio, el estudio del régimen permanente mediante la simulación presenta dificultades difíciles de salvar.

En el estudio del régimen transitorio tiene una importancia decisiva la caracterización de las condiciones iniciales. Éstas pueden ser fijas (el sistema empieza a funcionar siempre en las mismas condiciones; por ejemplo, en un taller puede suceder que al comienzo del turno todas las máquinas estén siempre dispuestas para funcionar porque existe un servicio nocturno de reparaciones, etc.) o aleatorias (el número de clientes que esperan a que se abra un establecimiento comercial).

Para estimar la media de una variable que caracterice el funcionamiento de un sistema en régimen transitorio deberemos ejecutar un cierto número de simulaciones, n, independientes. La independencia de los resultados obtenidos en las diversas simulaciones permite determinar fácilmente el intervalo de confianza de la estimación, en el supuesto de que las observaciones proceden de una población distribuida normalmente:

$$\bar{x_i} \pm t_{|O(i)|-1,\alpha} \frac{\sqrt{|O\,(i)| \sum_{k_i \in O(i)} x_{i_k}^2 - \left(\sum_{k_i \in O(i)} x_{i_k}\right)^2}}{\sqrt{N_i\,|O\,(i)|\,(|O\,(i)|-1)}} \forall i \in I$$

donde $\bar{x_i}$ es la media de la muestra i, $|O\,(i)|$ el número de observaciones de la muestra i, N_i el número de elementos de la población y t la correspondiente a la ley de Student-Fisher, con $n-1$ grados de libertad. Desde luego, el supuesto de distribución normal no siempre se cumple y se ha de comprobar; si no es aceptable, se han de transformar los resultados para obtener valores distribuidos normalmente. Para ello se puede sacar partido del teorema central del límite (la suma de variables independientes tiende a distribuirse normalmente cuando el número de sumandos crece; en particular, ello sucede con las medias de muestras), formando grupos de m observaciones y calculando después el intervalo de confianza considerando la media de cada uno de estos grupos como una observación.

En el caso de proporciones, si en una muestra de n observaciones la proporción del acontecimiento es ω'_j, el intervalo de confianza es:

$$\hat{\omega_j} \pm\ t_\alpha \sqrt{\frac{\omega'_j(1-\omega'_j)}{N}} \forall j \in J$$

donde J es el conjunto de propiedades, N es el tamaño teórico de la población y la t es la de la ley normal (ejemplo: la proporción puede corresponder a la de clientes que tienen que hacer cola durante un tiempo superior a media hora; por cierto, obsérvese que no es lo mismo la proporción de clientes de este tipo a lo largo del tiempo que la media de las proporciones diarias).

En cuanto al régimen permanente se ha de tener en cuenta que es un límite ideal, al que el comportamiento del sistema sólo se aproxima aceptablemente después de un tiempo más o menos largo, según las características del propio sistema y de las condiciones iniciales. De hecho, el concepto mismo de régimen permanente no es aplicable en muchos casos, puesto que las características del entorno varían a lo largo del tiempo; en otros casos el tiempo que ha de transcurrir para que el comportamiento se aproxime suficientemente al correspondiente al régimen permanente puede ser más largo que la vida técnica del propio sistema. En definitiva, antes de abordar el estudio del funcionamiento de un sistema en régimen permanente hemos de preguntarnos si ello es o no pertinente (y aseguramos de que no nos dejamos arrastrar por el ejemplo de la teoría de colas).

De todos modos está claro que en ocasiones tendremos que estudiar mediante la simulación las características del funcionamiento en régimen permanente de un sistema.

Para ello, el procedimiento más satisfactorio, desde un punto de vista teórico, es el denominado método regenerativo. Si el funcionamiento del sistema presenta *ciclos de regeneración* (intervalos separados por instantes, *puntos de regeneración*, tales que las probabilidades de los acontecimientos futuros dependen del estado del sistema en dichos instantes pero no de la historia), la observación correspondiente a cada ciclo es independiente de las demás. Dado que la duración, el número de transacciones, etc. son distintos para los diversos ciclos, es necesario tenerlo en cuenta al calcular los estimadores, pero esto no ofrece grandes dificultades. Éstas consisten en que la definición de los puntos de regeneración, si existen, no siempre es fácil, y en que los ciclos de regeneración pueden ser muy largos, especialmente en sistemas muy saturados, y, por consiguiente, el tiempo simulado necesario para obtener un número de ciclos suficiente puede ser muy elevado. Por ejemplo, en un sistema de colas con un solo canal con llegadas individuales separadas por un tiempo que sigue una cierta ley estadística, un instante en que llega una unidad al sistema vacío es un punto de regeneración.

Si no se utiliza el método regenerativo, para estimar el comportamiento del sistema en régimen permanente hay que simular un cierto tiempo, en primer lugar, para amortiguar la influencia de las condiciones iniciales (es conveniente no utilizar en la estimación las primeras observaciones, aunque no existe ninguna regla sencilla con fundamento teórico claro para determinar cuántas hay que despreciar) y, en segundo lugar, para tener un número de observaciones suficiente, que permita obtener una buena estimación. El problema es que las observaciones correspondientes a periodos sucesivos no son independientes: la media de las observaciones es una buena estimación y asimismo la variancia puede estimarse de la forma habitual, pero deja de ser válida la fórmula habitual para la estimación de la variancia de la media.

Fundamentalmente, las observaciones independientes pueden obtenerse de dos for-

mas distintas: por repetición y por formación de lotes.

La primera consiste simplemente en repetir la simulación un cierto número de veces, con un número aleatorio inicial distinto. Cada simulación es entonces independiente de las demás y se pueden aplicar las técnicas que hemos recordado a propósito del régimen transitorio. Esta forma de proceder puede exigir mucho tiempo de ordenador (cada simulación incluye un tiempo de régimen transitorio, a eliminar, que puede ser largo, y si la variable no tiene una distribución normal hay que formar grupos y el número de simulaciones puede ser relativamente elevado).

Es menos costoso, aunque presenta mayores dificultades teóricas, utilizar el procedimiento de formación de lotes. Si consideramos una variable tal como los costes diarios de funcionamiento de un sistema, los costes de dos días sucesivos no son, en general independientes, pero sí lo son si los dos días están muy separados. De esta observación surge el procedimiento de las medias de los lotes: se divide el conjunto de observaciones en lotes de la misma magnitud (cada lote corresponde a un cierto número de, digamos, días consecutivos) y el intervalo de confianza se calcula a partir de la media de los lotes, con las expresiones indicadas más arriba para el caso de ciclo fijo. Recuérdese, no obstante, que dichas fórmulas correspondían a dos supuestos: normalidad de la distribución e independencia de las observaciones; la hipótesis de normalidad se cumplirá aceptablemente en la mayoría de los casos a partir de tamaños de lote no muy grandes (tal vez no superiores a 30 observaciones, por ejemplo) pero no hay una regla sencilla que permita prever el tamaño de los lotes necesario para que sus medias sean independientes (por supuesto, se puede calcular el coeficiente de correlación y, si resulta muy elevado, aumentar el tamaño de los lotes, pero esto presenta dificultades porque el estimador del coeficiente de correlación tiene mucha dispersión y las conclusiones pueden ser aventuradas; además, al aumentar el tamaño de los lotes, para un tiempo de simulación dado, se reduce el número de valores utilizables para el cálculo del intervalo de confianza). Han sido propuestos diversos procedimientos prácticos (en el sentido de que no están completamente justificados desde un punto de vista teórico), alguno de los cuales es bastante complicado y algún otro tan sencillo que resulta francamente difícil de justificar (por ejemplo: dividir los datos disponibles en 5 ó 10 lotes).

3.8. Técnicas de reducción de la variancia

Cuanto mayor es la variancia de los resultados, mayor es el tamaño de la muestra requerido para obtener en la estimación una precisión dada. De ahí el interés que ofrecen las técnicas para la reducción de la variancia.

La simulación se utiliza con frecuencia para comparar políticas; por ejemplo, políticas de gestión de stocks. Se puede hacer n simulaciones para cada una de m políticas, de forma completamente independiente, pero también se puede simular n veces la demanda y utilizar los mismos valores de la demanda para cada una de las políticas. Con el primer procedimiento, hay dos causas de variación de los resultados (demanda y política), con el segundo, las diferencias en los resultados son atribuibles sólo a las diferencias en las políticas y la variabilidad es menor.

Cada vez que se realiza una simulación se obtiene una muestra de los resultados. Si la secuencia de números pseudoaleatorios está sesgada también lo estarán los resultados. Por ejemplo, si los números pseudoaleatorios tienen valores bajos una demanda simulada a partir de ellos resultará sesgada y lo mismo sucederá con el stock medio y el número de rupturas, pongamos por caso. Si se hace una segunda simulación con números pseudoaleatorios que sean complementarios de los utilizados en la primera, los sesgos serán de sentido opuesto y generalmente la estimación obtenida a partir de estas dos simulaciones tendrá menos variancia que la que se obtendría con números pseudoaleatorios independientes.

Es bien sabido que, cuando se practica un sondeo de opinión o de mercado se puede hacer un muestreo al azar en el conjunto de la población o un muestreo estratificado, en el que se asigna a cada estrato o grupo de los que resultan al efectuar una partición en dicha población una cierta proporción de la muestra total. Con ello, por una parte, se evita que a causa del azar determinados grupos queden sin representación en la muestra y, por otra, la proporción de cada grupo o estrato puede tener en cuenta su mayor dispersión y así conseguir un intervalo de confianza mínimo para una muestra de tamaño dado. Esta misma idea se puede aplicar a la simulación, aunque en la práctica no siempre resulta fácil; pero conviene retenerla sobre todo en sistemas en que una situación que se presenta muy raramente tiene consecuencias muy variables (en estos casos puede ser conveniente una simulación en que la situación referida esté sobrerrepresentada).

Otra técnica para reducir la variancia consiste en utilizar estimadores indirectos. Por ejemplo, considérese la simulación de una cola para estimar el tiempo medio de espera de las unidades, conocidas las leyes de llegada y de servicio. Sea τ el tiempo medio entre llegadas (al hacer una simulación, el tiempo medio entre las llegadas simuladas generalmente no coincidirá con este valor τ); sea x_k el tiempo de espera de la unidad k e y_k el tiempo transcurrido entre las llegadas de las unidades $k-1$ y k; se puede calcular:

$$z_k = x_k + y_k - \tau$$

de donde:

$$\bar{z} = \bar{x} + \bar{y} - \tau$$

pero $\bar{y} = \tau$, por lo cual $\bar{z} = \bar{x}$, es decir, en lugar de estimar el tiempo medio de espera directamente a través de las x_k se puede hacer indirectamente a través de las z_k; puesto que existe una correlación negativa entre las x_k y las y_k (a tiempos medios entre llegadas altos, corresponden tiempos de espera bajos), la variancia de las z_k puede ser menor que la de las x_k.

3.9. El uso de los ordenadores en la simulación. Introducción a los lenguajes de simulación

Desde luego, sin la ayuda de ordenadores no se puede simular y estimar los valores requeridos con una mínima precisión, salvo en sistemas muy sencillos, que la mayoría de las veces serán ejemplos y ejercicios.

Las aplicaciones requieren el uso de ordenadores. Para ello, se puede utilizar un lenguaje de programación cualquiera, de uso más o menos general, o un lenguaje especializado. Los del primer tipo dan la máxima flexibilidad, pero obligan a un trabajo de programación específico, incluso para muchas funciones que son comunes a la mayor parte de las simulaciones (histogramas, cálculo de medias y desviaciones tipo, etc.). En cambio, los lenguajes de simulación más o menos especializados pueden ser rígidos en algún aspecto (determinadas situaciones o reglas de funcionamiento pueden ser difíciles de simular) pero ahorran mucho trabajo de programación.

Normalmente, alguien que se disponga a realizar una simulación conocerá y dispondrá de uno o más lenguajes de uso general; puede en cambio que no conozca ningún lenguaje de simulación o no disponga de ellos. Tales circunstancias pueden ser decisivas para inclinarse por un lenguaje de uso general, cuando la simulación no es muy compleja y previsiblemente no se van a realizar otras; en caso contrario, vale la pena realizar el esfuerzo que requiera utilizar un lenguaje de tipo especializado. Se ha de tener en cuenta que el enfoque de los lenguajes de simulación es en general muy distinto del de los lenguajes de programación de uso general, por lo que el conocimiento de estos últimos constituye al principio, más que una ayuda, un cierto obstáculo.

En general los lenguajes de simulación se caracterizan por la forma de describir el sistema y el flujo de las unidades en el mismo. A partir de dicha descripción, el programa se encarga de la gestión de los acontecimientos, del cálculo de estadísticos y de la presentación de resultados. Todos los lenguajes incluyen generadores de números pseudoaleatorios, sentencias para generar muestras de las distribuciones más habituales, para el cálculo de estadísticos adicionales a los que obtiene el programa de forma estándard, etc.

Entre los numerosos lenguajes de simulación son muy conocidos el ARENA y el WITNESS. Para aprenderlos hay que recurrir a textos específicos, y mejor aún, a los manuales del paquete informático correspondiente.

Apéndices

Apéndices A

Comentarios sobre el modelo $M/M/1 : GD/\infty/\infty$

El modelo más sencillo de la teoría de colas corresponde al sistema de referencia: tiempo entre llegadas y tiempo de servicio distribuidos exponencialmente, un solo canal, disciplina de cola general, capacidad de la cola infinita y centro emisor infinito.

Las fórmulas para dicho modelo se encuentran en el apéndice B; aquí se incluyen algunos comentarios, para ilustrar algunas consecuencias que pueden deducirse de la teoría de colas.

Sean:

λ	número medio de llegadas por unidad de tiempo
μ	capacidad de servicio (número medio de servicios por unidad de tiempo)
$\rho = \frac{\lambda}{\mu}$	factor de servicio
p_n	probabilidad de que en el sistema se encuentren n unidades
L	número medio de unidades en el sistema
L_q	número medio de unidades en la cola
W	tiempo medio de estancia de las unidades en el sistema
W_q	tiempo medio de estancia de las unidades en la cola

Se cumple:

$$p_0 = \rho; p_n = (1-\rho)\rho^n$$

$$L = \frac{\rho}{1-\rho}; L_q = \frac{\rho^2}{1-\rho}; W = \frac{1}{\mu-\lambda}; W_q = \frac{\rho}{\mu-\lambda}$$

Estas expresiones nos permiten representar, por ejemplo, el valor de L en función de ρ. Dicho valor crece primero lentamente y después progresivamente más deprisa, para tender a infinito cuando el factor de servicio tiende a 1. Obsérvese que la gráfica presenta un codo, que empieza, aproximadamente, para $\rho = 0{,}7$, lo cual significa que para

valores de la demanda superiores al 70 % de la capacidad del canal, el sistema puede presentar una congestión apreciable; en los supuestos del modelo, el número medio de unidades en el sistema crece ilimitadamente si el valor de la demanda se aproxima al de la capacidad (aunque entonces ha de transcurrir un tiempo muy largo para que el comportamiento del sistema se aproxime al correspondiente al régimen permanente; de hecho, el número de unidades en el sistema es entonces muy inestable). Para $\rho = 1$ no hay régimen permanente ni, claro está, para $\rho > 1$ (en este último supuesto el número de unidades en el sistema tiende a crecer ilimitadamente, aunque en ciertos intervalos puede decrecer).

Es interesante comparar este modelo $M/M/1 : GD/\infty/\infty$ con el $M/D/1 : GD/\infty/\infty$. Para este último:

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

es decir, la longitud media de la cola es exactamente la mitad de la correspondiente al modelo $M/M/1 : GD/\infty/\infty$; así pues, la reducción en la dispersión del tiempo de servicio tiene un efecto muy considerable, positivo, en el funcionamiento del sistema.

Apéndices B

Fórmulas, Tablas y Gráficos

B.1. Fórmulas para procesos de nacimiento y muerte

B.1.1. Régimen transitorio

$$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t)$$
$$\frac{dP_n(t)}{dt} = \lambda_{n-1} P_{n-1}(t) - (\lambda_n + \mu_n) P_n(t) + \mu_{n+1} P_{n+1}(t)$$

B.1.2. Régimen permanente

$$P_0 = \left[1 + \sum_{n=1}^{\infty} \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} \right]^{-1}$$

$$P_n = \frac{\lambda_{n-1}\lambda_{n-2}...\lambda_1\lambda_0}{\mu_n\mu_{n-1}...\mu_2\mu_1} P_0 = \frac{\lambda_{n-1}}{\mu_n} P_{n-1}$$

$$\text{Erlang-}k\text{: } f(t) = \frac{1^k}{k-1!} t^{k-1} e^{-1t}$$

B.2. Modelos de colas de los que se incluyen fórmulas

B.2.1. $M/M/1 : GD/\infty/\infty$

$$\rho = \frac{\lambda}{\mu}; P_0 = 1 - \rho; P_n = \rho^n(1-\rho); \sum_{k=n}^{\infty} P_k = \rho^n$$

$$L = \frac{\rho}{1-\rho}; L_q = \frac{\rho^2}{1-\rho}; W = \frac{1}{\mu - \lambda}; W_q = \frac{\rho}{\mu - \lambda}$$

B.2.2. $M/M/1 : FIFO/\infty/\infty$

$$\text{Prob}(T_q > t) = \rho e^{-(\mu-\lambda)t}$$

$$\text{Prob}(T > t) = e^{-(\mu-\lambda)t}$$

B.2.3. $M/M/s : GD/\infty/\infty$

$$\rho = \frac{\lambda}{\mu s}; P_0 = \left[\sum_{n=0}^{s-1} \frac{s^n \rho^n}{n!} + \frac{\rho^s s^s}{s!}\frac{1}{1-\rho}\right]^{-1}$$

- $P_n = \frac{s^n \rho^n}{n!} P_0$, si $n \leq s$
- $P_n = \frac{s^s \rho^n}{s!} P_0$, si $n \geq s$

$$L_q = \frac{s^s \rho^{s+1}}{(1-\rho)^2 s!} P_0; L = L_q + \rho s; W = \frac{L}{\lambda}; W_q = \frac{L_q}{\lambda}$$

B.2.4. $M/M/s : FIFO/\infty/\infty$

$$\text{Prob}(T_q > t) = e^{-s\mu t(1-\rho)} \frac{s^s \rho^s}{s!} \frac{1}{1-\rho} P_0$$

- Prob($T > t$)=$e^{-\mu t}\left\{1 + \frac{P_s}{1-\rho}\left[\frac{1-e^{-\mu t\left(s-1-\frac{\lambda}{\mu}\right)}}{s-1-\frac{\lambda}{\mu}}\right]\right\}$, si $s-1-(\lambda/\mu) \neq 0$
- Prob($T > t$)=$e^{-\mu t}\left(1 + \frac{P_s}{1-\rho}\mu t\right)$, si $s-1-(\lambda/\mu) = 0$

B.2.5. $M/M/\infty : GD/\infty/\infty$

$$P_0 = e^{-(\lambda/\mu)}; P_n = \frac{(\lambda/\mu)^n}{n!} e^{-(\lambda/\mu)}$$

$$L = \frac{\lambda}{\mu}; W = \frac{1}{\mu}; L_q = W_q = 0$$

B.2.6. $M/M/1 : GD/M/\infty$

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

- si $\rho = 1$:

  $P_0 = \frac{1}{M+1}$; $P_n = \frac{1}{M+1}$

$L = \frac{M}{2}$; $L_q = \frac{M(M-1)}{2(M+1)}$

- si $\rho \neq 1$:

  $P_0 = \frac{1-\rho}{1-\rho^{M+1}}$; $P_n = \frac{1-\rho}{1-\rho^{M+1}}\rho^n$

  $L = \frac{\rho}{1-\rho} - \frac{(M+1)\rho^{M+1}}{1-\rho^{M+1}}$; $L_q = L - (1 - P_0)$

- $\forall \rho$

  $W = \frac{L}{\lambda(1-P_M)}$; $W_q = \frac{L_q}{\lambda(1-P_M)}$

B.2.7. $M/M/s : GD/M/\infty$

$M \leq s$

$\rho = \frac{\lambda}{\mu s}$; $P_0 = \left[\sum_{n=0}^{M} \frac{\rho^n s^n}{n!}\right]^{-1}$; $P_n = \frac{\rho^n s^n}{n!} P_0$

$L = \rho s \left(1 - \frac{\rho^M s^M}{M!} P_0\right)$; $L_q = 0$

$W = \frac{L}{\lambda}$; $W_q = 0$

$M > s$

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

- Si $\rho = 1$:

  $P_0 = \left[1 + \sum_{n=1}^{s} \frac{s^n}{n!} + \frac{s^s}{s!}(M - s)\right]^{-1}$

  - $P_n = \frac{s^n \rho^n}{n!} P_0$; si $n \leq s$
  - $P_n = \frac{s^s \rho^n}{s!} P_0$; si $s \leq n \leq M$

  $L = P_0 \left[\sum_{n=0}^{s} \frac{n s^n}{n!} + \frac{s^s}{2s!}(M - s)(M + s + 1)\right]$

  $L_q = \frac{s^s}{s!} P_0 \frac{M-s}{2}(M - s + 1)$

- Si $\rho \neq 1$:

  $P_0 = \left[\sum_{n=0}^{s} \frac{\rho^n s^n}{n!} + \frac{s^s}{s!}\frac{\rho}{1-\rho}\left(\rho^s - \rho^M\right)\right]^{-1}$

  - $P_n = \frac{s^n \rho^n}{n!} P_0$, si $n \leq s$
  - $P_n = \frac{s^s \rho^n}{s!} P_0$, si $s \leq n \leq M$

  $L_q = \frac{s^s \rho^{s+1}}{(1-\rho)^2 s!} P_0 \left\{1 - \rho^{M-s}\left[1 + (1 - \rho)(M - s)\right]\right\}$

  $L = L_q + \rho s$

- $\forall \rho$

  $W = \frac{L}{\lambda(1-P_M)}$; $W_q = \frac{L_q}{\lambda(1-P_M)}$

B.2.8. $M/M/1 : GD/\infty/N$

$$P_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}; P_n = \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n P_0$$

$$L = N - \frac{\mu}{\lambda}(1 - P_0); L_q = N - \frac{\mu + \lambda}{\lambda}(1 - P_0)$$

$$W = \frac{L}{\lambda(N - L)}; W_q = \frac{L_q}{\lambda(N - L)}$$

B.2.9. $M/M/s : GD/\infty/N$ ($N \geq s$)

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{N!}{(N-n)!n!} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s}^{N} \frac{N!}{(N-n)!s!s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1}$$

- $P_n = \frac{N!}{(N-n)!n!} \left(\frac{\lambda}{\mu}\right)^n P_0$; si $n \leq s$
- $P_n = \frac{N!}{(N-n)!s!s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n P_0$; si $s \leq n \leq N$

$$L = \sum_{n=0}^{N} nP_n; L_q = \sum_{n=s}^{N} (n - s)P_n$$

$$W = \frac{L}{\lambda(N - L)}; W_q = \frac{L_q}{\lambda(N - L)}$$

B.2.10. $M/G/1 : GD/\infty/\infty$

σ = desviación típica de la ley de distribución de los tiempos de servicio.

$$\rho = \frac{\lambda}{\mu}; P_0 = 1 - \rho$$

$$L = \rho + \frac{\rho^2 + \lambda^2\sigma^2}{2(1 - \rho)} \text{(fórmula de Pollaczek-Khintchine)}$$

$$L_q = \frac{\rho^2 + \lambda^2\sigma^2}{2(1 - \rho)}; W = \frac{L}{\lambda}; W_q = \frac{L_q}{\lambda}$$

B.2.11. $M/D/1 : GD/\infty/\infty$

Ver modelo $M/G/1 : GD/\infty/\infty$ con σ=0.

B.2.12. $M/E_k/1 : GD/\infty/\infty$

$$\rho = \frac{\lambda}{\mu}; P_0 = 1 - \rho$$

$$L = \frac{\rho}{1-\rho}\left[1 - \frac{\rho}{2}\left(1 - \frac{1}{k}\right)\right]; L_q = \frac{\rho^2}{2(1-\rho)}\left(1 + \frac{1}{k}\right)$$

$$W = \frac{L}{\lambda}; W_q = \frac{L_q}{\lambda}$$

B.2.13. $M/M/s : GD/N/N$

$$P_0 = \left\{\sum_{n=0}^{s} \left(\begin{array}{c} N \\ n \end{array} \right) \rho^n + \sum_{n=s+1}^{N} \left(\begin{array}{c} N \\ n \end{array} \right) \frac{n!\rho^n}{s!s^{n-s}}\right\}^{-1}$$

- $P_n = \left(\begin{array}{c} N \\ n \end{array} \right) \rho^n P_0$, si $0 \leq n \leq s$
- $P_n = \left(\begin{array}{c} N \\ n \end{array} \right) \frac{n!\rho^n}{s!s^{n-s}} P_0$, si $s+1 \leq n \leq N$

$$L_q = \sum_{n=s+1}^{N} (n-s)P_n$$

$$L = L_q + (s - \bar{s}) = L_q + \frac{\lambda_f}{\mu}, \text{con } \lambda_f = \mu(s - \bar{s}) = \lambda(N - L); \bar{s} = \sum_{n=0}^{s} (s-n)P_n$$

B.2.14. $M/M/1 : GD/N/N$

$$L_q = N - \left(1 + \frac{1}{\rho}\right)(1 - P_0); L = N - \frac{1 - P_0}{\rho}$$

Tabla B.1: $M/M/s : GD/\infty/\infty$ (Valores de P_0)

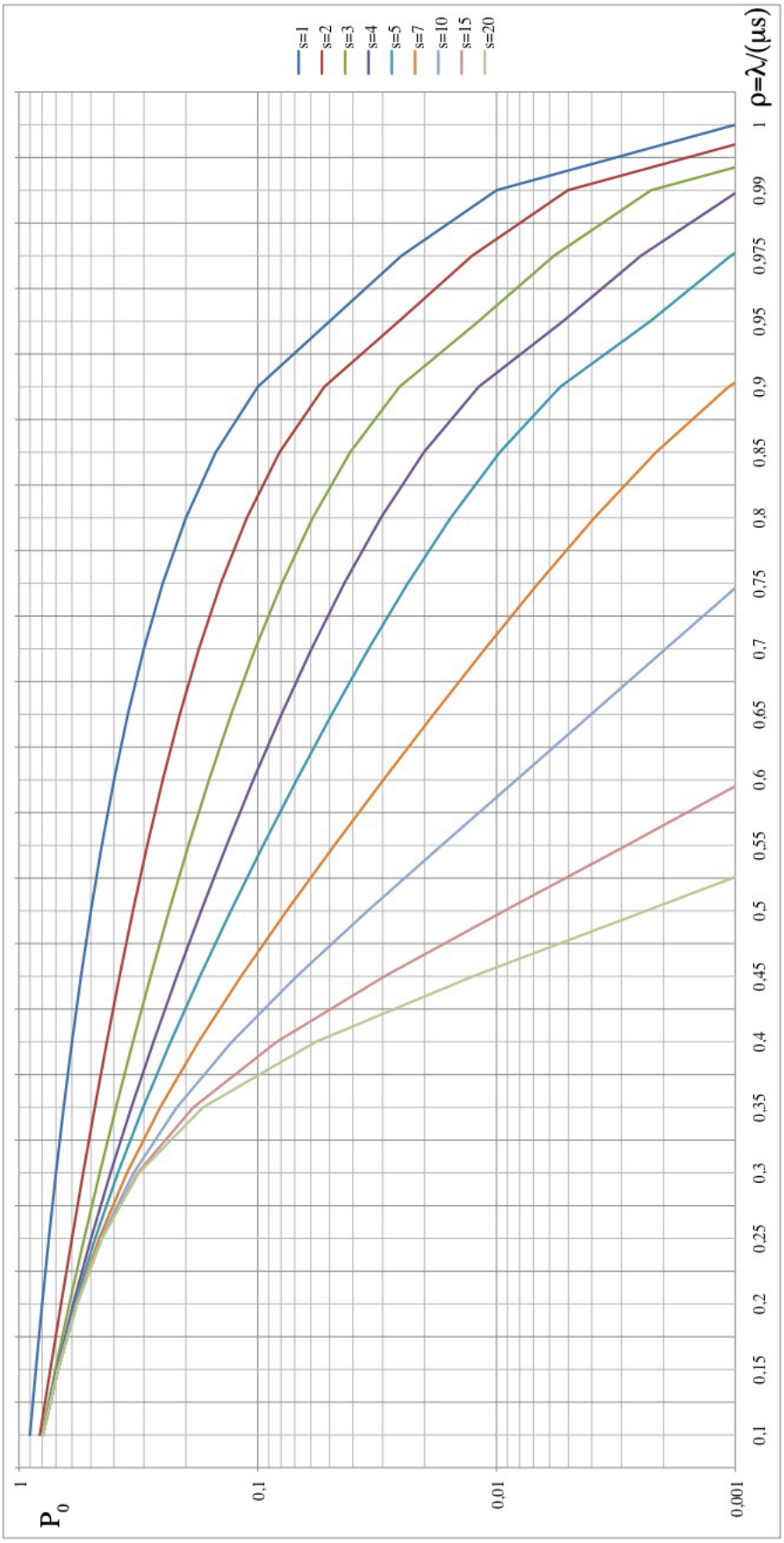


Figura B.1: $M/M/s : GD/\infty/\infty$ (Valores de P_0)

Tabla B.2: $M/M/s : GD/\infty/\infty$ (Valores de L_q)

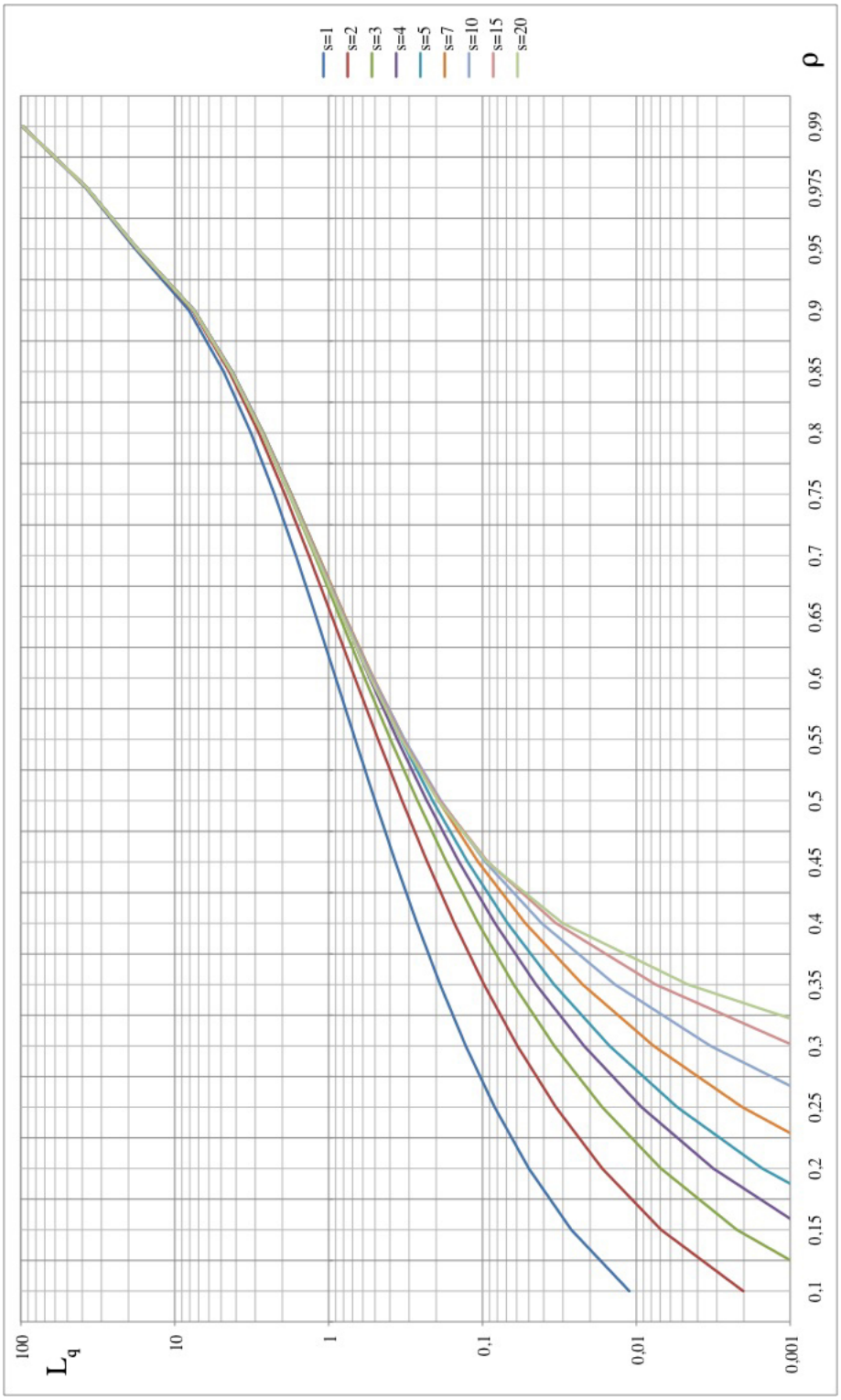


Figura B.2: $M/M/s : GD/\infty/\infty$ (Valores de L_q)

Tabla B.3: $M/M/s : GD/\infty/\infty$ (Valores de L)

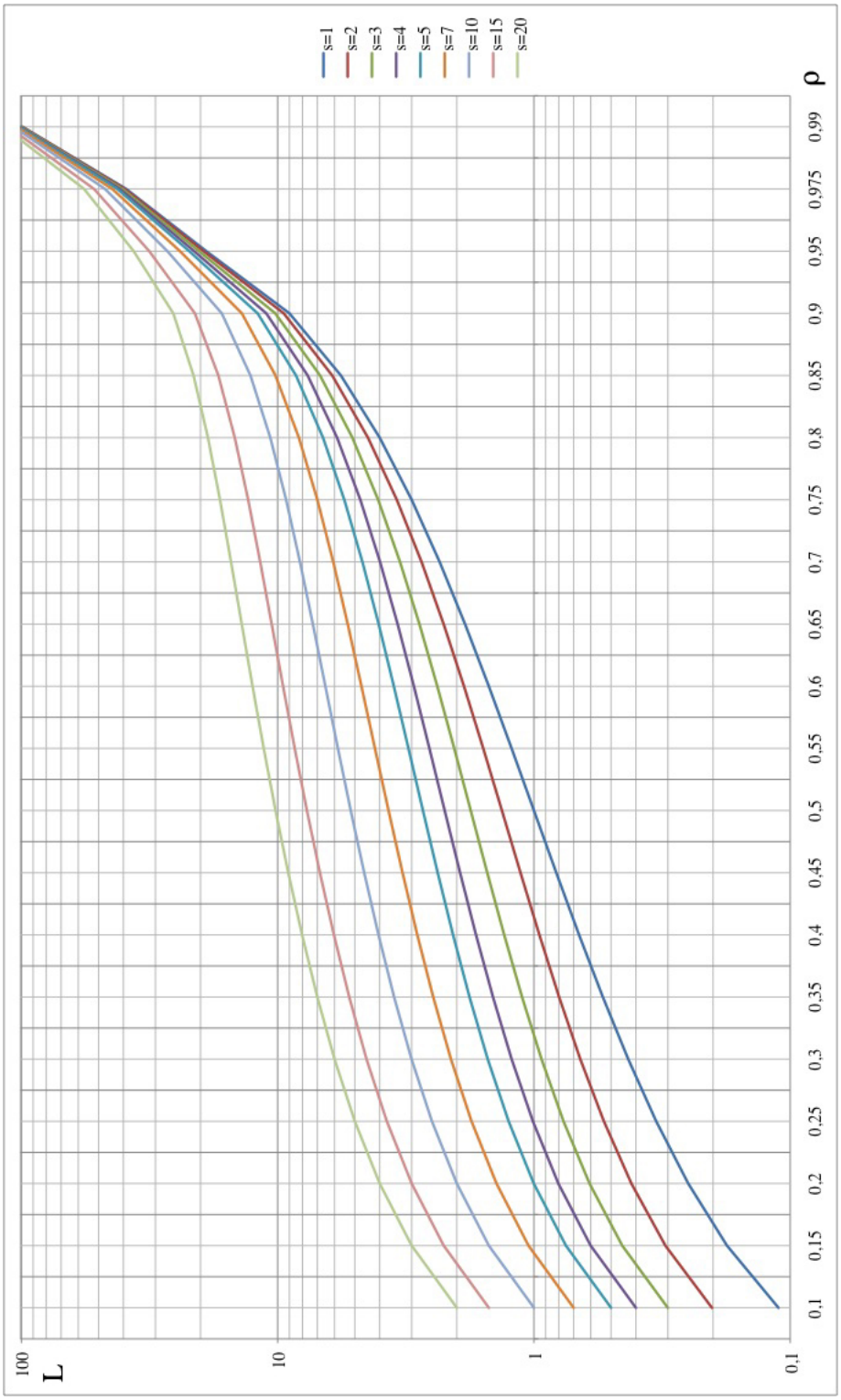


Figura B.3: $M/M/s : GD/\infty/\infty$ (Valores de L)