# INDICE

# 1. CADENAS DE MARKOV

Algunas veces nos interesa saber como cambia una variable aleatoria a través del tiempo. Por ejemplo, desearíamos conocer cómo evoluciona el precio de las acciones de una empresa en el mercado a través del tiempo. El estudio de como evoluciona una variable aleatoria incluye el concepto de procesos estocásticos. En este capítulo explica esos procesos, en especial uno que se conoce como cadena de Markov. Las cadenas de Markov se han aplicado en áreas tales como educación, mercadotecnia, servicios de salud, finanzas, contabilidad y producción. Comenzaremos definiendo el concepto de un proceso estocástico. En el resto del capítulo describiremos las ideas básicas que se necesitan para comprenderlas cadenas de Markov.

## 1.1-¿QUÉ ES UN PROCESO ESTOCÁSTICO?

Supóngase que observamos alguna característica de un sistema en puntos discretos en el tiempo (que llamarnos 0,1,2. . ,). Sea $X_t$ el valor de la característica del sistema en el tiempo *t*. En la mayor parte de los casos no se conoce $X_t$ con certeza antes del tiempo *t* y se puede considerar como variable aleatoria. Un proceso estocástico de tiempo discreto es simplemente una descripción de la relación entre las variables aleatorias $X_0$, $X_1$, $X_2$, ., . A continuación daremos unos ejemplos de procesos estocásticos de tiempo discreto.

**Ejemplo La ruina del jugador**

En el tiempo 0 tengo 2 dólares. En los tiempos 1, 2. ... participo en un juego en el que apuesto 1 dólar. Gano el juego con probabilidad p, y lo pierdo con probabilidad 1 -p. Mi meta es aumentar mi capital a 4 dólares, y tan pronto como lo logre se suspende el juego. El juego también se suspende si mi capital se reduce a 0 dólares. Si definimos que $X_t$ es mi capital después del juego cuando el tiempo es *t*, si es que lo hay, entonces se puede considerar que $X_0$, $X_1$, $X_2$, ..., $X_t$ son procesos estocásticos de tiempo discreto. Nótese que $X_0 = 2$ es una constante conocida, pero que $X_1$ y las demás $X_t$, son aleatorias. Por ejemplo, $X_1 = 3$ con probabilidad p y $X_1 = 1$ con probabilidad 1 - p. Nótese que si $X_t = 4$, entonces $X_{t+1}$ y todas las demás $X_t$, también serán igual a 4. Igualmente, si $X_t = 0$, entonces $X_{t+1}$ y todas las demás $X_t$ serán cero también. Por razones obvias, a estos casos se les llama problema de la ruina del jugador.

———————————————————————————————•

**Ejemplo**

En una urna que contiene bolas hay dos sin pintar. Se selecciona una bola al azar y se lanza una moneda. Si la bola elegida no está pintada y la moneda produce cara, pintamos la bola de rojo; si la moneda produce cruz, la pintamos de negro Si la bola ya está pintada, entonces cambiamos el color de la bola de rojo a negro o de negro a rojo, independientemente de si la moneda produce cara o cruz. Para modelar este caso como proceso estocástico, definimos a *t* como el tiempo después que la moneda ha sido lanzada por t-ésima vez y se ha pintado la bola escogida En cualquier tiempo se puede representar el estado mediante el vector [u r b], donde u es el número de bolas sin pintar en la urna, r el número de bolas rojas y b el número de bolas negras. Se nos dice que $X_0$ = [2 0 0]. Después del primer lanzamiento, una bola habrá sido pintada ya sea de rojo o de negro y el estado será [1 1 0] o [1 0 1].

Por lo tanto, podemos asegurar que $X_1$ = [1 1 0] o $X_1$ = [1 0 1]. Es claro que debe haber alguna relación entre las $X_t$,. Por ejemplo, si X, = [0 2 0] podemos asegurar que $X_{t+1}$ será [0 1 1].

**Ejemplo**

Sea $X_0$ el precio de una acción de Computadoras CSL al principio de este día hábil. También, sea $X_t$, el precio de esa acción al principio del t-ésimo día hábil en el futuro. Es claro que si se conocen los valores de $X_0$, $X_1$, $X_2$, ..., $X_t$ nos dicen algo acerca de la distribución de probabilidad de $X_{t+1}$; el asunto es: ¿que nos dice el pasado (los precios de las acciones hasta el tiempo t) acerca de $X_{t+1}$? La respuesta a esta pregunta es de importancia crítica en finanzas.

Terminaremos esta sección con una explicación breve de los procesos estócasticos de tiempo continuo. Un proceso estocástico de tiempo continuo es simplemente un proceso estocástico en el que el estado del tiempo se puede considerar cualquier tiempo y no sólo en instantes discretos. Por ejemplo, se puede considerar que el número de personas en un supermercado a los t minutos después de abrir, es un proceso estocástico de tiempo continuo. Los modelos en los que intervienen estos procesos se estudian en la sección de teoría de colas. Como el precio de una acción se puede observar en cualquier tiempo, y no solo al abrir la bolsa, se puede considerar como un proceso estocástico de tiempo continuo. Al considerarlo así, se ha podido llegar a importantes resultados en la teoría de finanzas, incluyendo la famosa fórmula de Black-Scholes para opción de precio.

Ejemplo
Un caso muy conocido es el proceso de movimiento Browniano, el cual tiene las siguientes características

1. suponga $t_0 < t_1 < ... < t_n$; los incrementos $X_{t_1} - X_{t_0}, \cdots X_{t_n} - X_{t_{n-1}}$ son variables aleatorias independientes
2. La distribución de probabilidad es $X_{t_2} - X_{t_2}$, $t_1 < t_2$, depende solo de $t_2 - t_1$
3. $P(X_t - X_s \le x) = \frac{1}{\sqrt{2\pi B(t-s)}} \int_{-\infty}^{x} e^{\frac{-u^2}{2B(t-s)}} du$ donde B es una constante

La historia de este proceso comienza con la observación de R. Brown en 1827 de pequeñas partículas inmersas en un liquido, las cuales mostraban movimientos irregulares. En 1905 Einsten explica el movimiento mediante el postulado que las partículas bajo observación eran sujetas de perpetuas colisiones con las partículas que rodean el medio. Los resultados derivados de forma analítica por Einstein fueron posteriormente verificados por experimentación.

En este caso $X_t$ denota el desplazamiento en el tiempo t de un partícula Browniana. El desplazamiento $X_t - X_s$, es normalmente distribuido sobre un intervalo de tiempo (s,t) que puede verse como la suma de pequeños desplazamientos. El teorema del limite Central es esencialmente aplicable y razonable afirmar que $X_t - X_s$, es normalmente distribuido. También que la distribución de $X_{t+h} - X_{s+h}$, es la misma, para cualquier h>0.

## 1.2 ¿QUÉ ES UNA CADENA DE MARKOV?

Un tipo especial de procesos estocásticos de tiempo discreto se llama cadena de Markov. Para simplificar nuestra presentación supondremos que en cualquier tiempo, el proceso estocástico de tiempo discreto puede estar en uno de un número finito de estados identificados por 1.2,..., s.

DEFINICIÓN Un proceso estocástico de tiempo discreto es una cadena de Markov si, para t = 0,1,2,... y todos los estados,

(1) $P(X_{t+1} = i_{t+1} \setminus X_t = i_t, X_{t-1} = i_{t-1}, \ldots, X_1 = i_1, X_0 = i_0) = P(X_{t+1} = i_{t+1} \setminus X_t = i)$

En esencia, la ecuación (1) dice que la distribución de probabilidad del estado en el tiempo t + 1 depende del estado en el tiempo t(i) y no depende de los estados por los cuales pasó la cadena para llegar a i, en el tiempo *t*.

En el estudio de las cadenas de Markov haremos la hipótesis adicional que para lodos los estados i y j, y toda t, $P(X_{t+1} = i \setminus X_t = j)$ es independiente de t. Esta hipótesis permite escribir

(2) $P(X_{t+1} = i \setminus X_t = j) = p_{ij}$

donde $p_{ij}$ es la probabilidad de que dado que el sistema está en el estado i en el tiempo t, el sistema estará en el estado j en el tiempo t + 1. Si el sistema pasa del estado i durante un periodo al estado j durante el siguiente, se dice que ha ocurrido una transición de i a j. Con frecuencia se llaman probabilidades de transición a las $p_{ij}$ en una cadena de Markov.

La ecuación (2) indica que la ley de probabilidad que relaciona el estado del siguiente periodo con el estado actual no cambia. o que permanece estacionaria, en el tiempo. Por este motivo, a menudo se llama Hipótesis de estabilidad a la ecuación (2). Toda cadena de Markov que cumple con la ecuación (2) se llama **cadena estacionionaria de Markov**,

El estudio de las cadenas de Markov también necesita que definamos $q_i$ como la probabilidad de que la cadena se encuentre en el estado i en el tiempo 0; en otras palabras, $P(X_0 = i) \bullet = q_i$,. Al vector $q = [q_1, q_2, \ldots, q_s]$ se le llama distribución inicial de probabilidad de la cadena de Markov. En la mayoría de las aplicaciones, las probabilidades de transición se presentan como una matriz P de probabilidad de transición s x s. La matriz de probabilidad de transición P se puede escribir como

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & & \ddots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{bmatrix}$$

Dado que el estado es i en el tiempo t, el proceso debe estar en algún lugar en el tiempo t + 1. Esto significa que para cada i.

$$\sum_{j=1}^{s} P(X_{t+1} = j / X_t = i) = 1$$

$$\sum_{j=1}^{s} p_{ij} = 1$$

También sabemos que cada elemento de la matriz P debe ser no negativo. Por lo tanto, todos los elementos de la matriz de probabilidad de transición son no negativos; los elementos de cada renglón deben sumar 1.

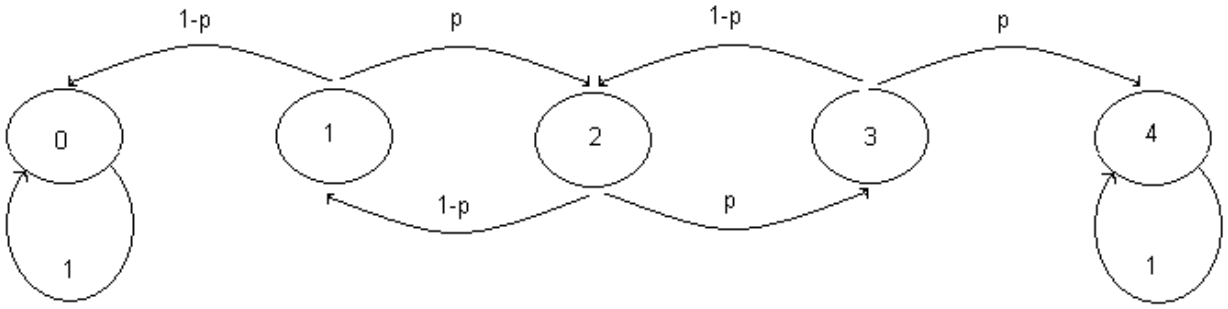
La ruina del jugador (continuación) Encuentre la matriz de transición del primer ejemplo.

Solución Como la cantidad de dinero que tengo después de t + 1 jugadas depende de los antecedentes del juego sólo hasta la cantidad de efectivo que tengo después de *t* jugadas, no hay duda que se trata de una cadena de Markov. Como las reglas del juego no varían con el tiempo, también tenemos una cadena de Markov estacionaria. La matriz de transición es la siguiente (el estado i quiere decir que tenemos i dólares):

| Estados | 0 dólares | 1 dólares | 2 dólares | 3 dólares | 4 dólares |
|---|---|---|---|---|---|
| 0 dólares | 1 | 0 | 0 | 0 | 0 |
| 1 dólares | 1-p | 0 | p | 0 | 0 |
| 2 dólares | 0 | 1-p | 0 | p | 0 |
| 3 dólares | 0 | 0 | 1-p | 0 | p |
| 4 dólares | 0 | 0 | 0 | 0 | 1 |

Si el estado es 0 dólares o 4 dólares no juego más y, por lo tanto el estado no puede cambiar; entonces $p_{\infty\,\infty} = p_{44} = 1$. Para los demás estados sabemos que, con la probabilidad p, el estado del siguiente periodo será mayor que el estado actual en 1, y con probabilidad 1 -p, el estado del siguiente periodo será menor en 1 que el estado actual.

Figura 1
Representación gráfica de la matriz, de transición para el ejemplo de la ruina del jugador



Una matriz de transición se puede representar con una gráfica en la que cada nodo represente un estado y arco (i,j) represente la probabilidad de transición $p_{ij}$. La Fig. 1 es una representación gráfica de la matriz de probabilidad de transición para este ejemplo.

EJEMPLO 2. (Continuación) Determine la matriz de transición del Ejemplo 2 en la sección anterior.

Solución Como el estado de la urna después del siguiente lanzamiento de la moneda depende sólo del pasado del proceso hasta el estado de la unía después del lanzamiento actual, se trata de una cadena de Markov. Como las reglas no varían a través del tiempo, tenemos una cadena estacionaria de Markov. La matriz de transición para el Ejemplo 2 es la siguiente:

| | [0 1 1] | [0 2 0] | [0 0 2] | [2 0 0] | [1 1 0] | [1 0 1] |
|---|---|---|---|---|---|---|
| [0 1 1] | 0 | ½ | ½ | 0 | 0 | 0 |
| [0 2 0] | 1 | 0 | 0 | 0 | 0 | 0 |
| [0 0 2] | 1 | 0 | 0 | 0 | 0 | 0 |
| [2 0 0] | 0 | 0 | 0 | 0 | ½ | ½ |
| [1 1 0] | ¼ | ¼ | 0 | 0 | 0 | ½ |
| [1 0 1] | ¼ | 0 | ¼ | 0 | ½ | 0 |

EJEMPLO 3 (Continuación) En los últimos años, los estudiantes de finanzas han dedicado mucho esfuerzo a contestar la pregunta de si el precio diario de una acción se puede describir mediante una cadena de Markov. Supongamos que el precio diario de una acción, como la de Computadoras CSL, se puede representar por una cadena de Markov. ¿Qué nos dice esto? Simplemente que la distribución de probabilidad del precio de las acciones mañana depende sólo del precio de hoy, pero no de los precios anteriores. Si el precio de una acción se puede representar con cadena de Markov, los "tablistas" que tratan de predecir los precios futuros en base a los comportamientos seguidos durante el pasado están mal. Por ejemplo, supongamos que el precio diario de una acción de CSL sigue una cadena de Markov y el preciod e hoy es 50 dólares. Entonces, para predecir el precio de mañana no importa si el precio ha aumentado o disminuido durante cada uno de los últimos 30 días. En cualquier caso, o en cualquier otro caso que pudiera haber conducido al precio actual de 50 dólares, la predicción del precio de mañana se debe basar sólo en el hecho de que hoy el precio de esas acciones es de 50 dólares. En la actualidad, el consenso es que para la mayor parte de las acciones, su cotización diaria se puede describir con una cadena de Markov. A esta idea se le llama con frecuencia hipótesis del mercado eficiente.

**PROBLEMAS**

1. En Smalltown, al 90% de los días soleados siguen días soleados, y al 80% de los días nublados siguen días nublados. Con esta información modelar el clima de Smalltown como cadena de Markov.

2. Se tiene un sistema de inventario en el que la secuencia de eventos durante cada periodo es como sigue: (1) Se observa el nivel de inventario (llamémosle í) al principio del periodo. (2) Si i ≤1, se

piden 4 –i unidades. Si i ≥ 2, no se hace ningún pedido. (3) Los Clientes no piden unidades durante el periodo, con probabilidad 1/3; se pide una unidad durante el periodo, con probabilidad 1/3. y se piden 2 unidades durante el periodo, con probabilidad 1/3 (4) Se observa el nivel de inventario al principio del siguiente periodo.
Defina un estado de periodo como el nivel de inventario al principio del periodo. Determine la matriz de transición que pudiera usarse para modelar este sistema de inventario como una cadena de Markov.

3. Una fábrica tiene dos máquinas. Durante cualquier día, cada máquina que trabaja al principio del día tiene probabilidad x de descomponerse. Si se descompone una máquina durante el día, se manda a un taller de reparación y estará trabajando dos días después que se descompuso. Por ejemplo, si una máquina se descompone durante el día 3, estará trabajando al principio del día 5. Si se hace que el estado del sistema sea el número de máquinas que trabajan al principio del día, formule, una matriz de probabilidad de transición para este caso.

4. En relación con el problema 1, suponga que el tiempo de mañana en Smalltown depende del tiempo que haya prevalecido los últimos dos días, como sigue; (1) Si los últimos dos días han sido soleados, entonces el 95% de las veces mañana será soleado. (2) Si ayer estuvo nublado y hoy soleado, entonces el 70% de las veces mañana estará soleado. (3) Si ayer estuvo soleado y hoy está nublado, entonces el 60%, de las veces mañana estará nublado. (4) Si los últimos dos días fueron nublados, entonces el 80% de las veces mañana será nublado.

Con esta información modele el clima de Smalltown como cadena de Markov. Si el tiempo de mañana dependiera del de los últimos tres días, ¿cuántos estados se necesitarían para modelar el clima como cadena de Markov? Nota: El método que se usa en este problema se puede aplicar para modelar un proceso estocástico de tiempo discreto como cadena de Markov. aun si $X_{t+1}$ depende de los estados anteriores a $X_t$, tal como $X_{t-1}$ en este ejemplo.

5. Sea X, la ubicación de su ficha en el tablero deMonopoly después de t tiradas de dados. ¿Se puede modelar X, como cadena de Markov? Si no esa así ¿cómo podemos modificar la definición del estado en el tiempo t para que $X_0$, $X_1$,. . ., $X_i$, • • • sea una cadena de Markov? Sugerencia: ¿Cómo va un jugador a la cárcel? En este problema, suponga que los jugadores que van a prisión permanecen allí hasta que en la tirada de dados -sacan doble número o hasta que hayan estado tres turnos en prisión, lo que se represente primero.

6. En el Prob. 3, suponga que una máquina que se descompone regresa al servicio tres días después.
Por ejemplo, la máquina que se descompone el día 3 estará trabajando al principio del día 6. Determine una matriz de transición de probabilidad para este caso.

# 1.3 PROBABILIDADES DE TRANSICIÓN DE n ETAPAS

Suponga que estudiamos una cadena de Markov con matriz P de probabilidad de transición conocida. Como todas las cadenas con las que trataremos son estacionarias, no nos importará identificar nuestras cadenas de Markov como estacionarias. Una pregunta de interés es: si una cadena de Markov está en el estado i en el tiempo m, ¿cuál es la probabilidad que n periodos después la cadena de Markov este en el estado *j*? Como se trata de una cadena de Markov estacionaria, esta probabilidad será independiente de m y, por lo tanto, podemos escribir

$$P(X_{m+n} = j \setminus X_m = i) = P(X_n = j \setminus X_0 = i) = P_{ij}(n)$$

donde $P_{ij}(n)$ se llama probabilidad en la etapa n de una transición del estado i al estado j.

Es claro que $P_{ij}(1) = P_{ij}$. Para determinar $P_{ij}(2)$ nótese que si el sistema se encuentra hoy en el estado i, entonces para que el sistema termine en el estado j dentro de 2 periodos, debemos pasar del estado i al estado k y después pasar del estado k al estado j(Fig. 2). Este modo de razonar nos permite escribir

$P_{ij}(2) = \sum_{k=1}^{s}$ (probabilidad de transición de i a k}x (probabilidad de transición de k a j)

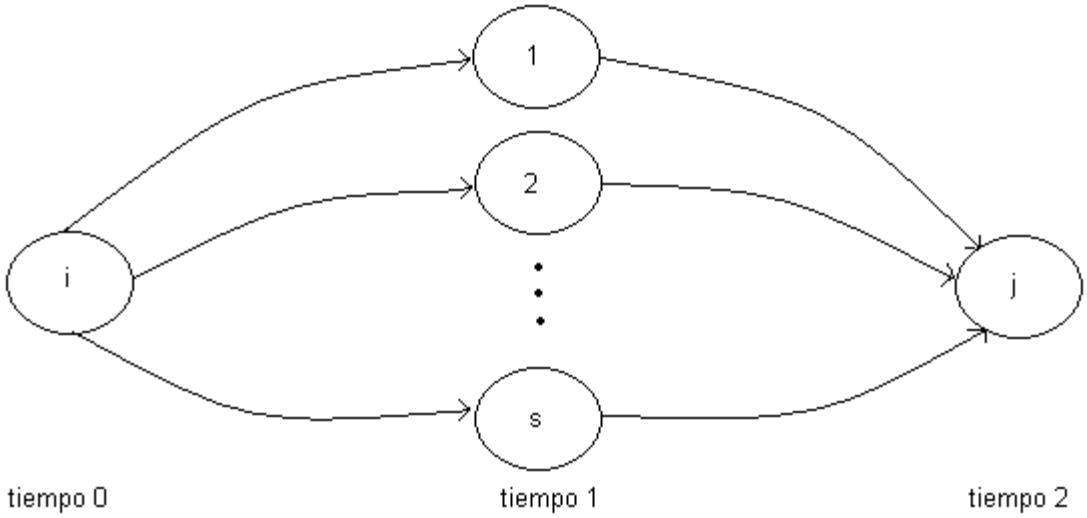
De acuerdo con la definición de P, la matriz de probabilidad de transición, replanteamos la última ecuación en la siguiente forma:

(3) $$P_{ij}(2) = \sum_{k=1}^{s} p_{ik}\, p_{kj}$$

El segundo miembro de la ecuación (3) es tan sólo el producto escalar del renglón i de la matriz P por la columna j de esa matriz. Por lo tanto, $P_{ij}(2)$ es el ij-ésimo elemento de la matriz $P^2$. Generalizando este modo de razonar, se puede demostrar que para n >1,

(4) $P_{ij}(n)$ = elemento ij-ésimo de $P^n$

Figura 2
Representación de la transición de i a j en dos pasos




Naturalmente, para n = 0, $P_{ij}(0) = P(X_0 = j \mid X_0 = i)$ y, por lo tanto, debemos escribir

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

En el Ejemplo. 4 mostraremos el uso de la ecuación (4).

EJEMPLO 4 Ejemplo de Cola Suponga que toda la industria de refrescos produce dos colas. Cuando una persona ha comprado la cola 1, hay una probabilidad de 90% de que su siguiente compra sea de cola 1. Si una persona compró cola 2, hay 80% de probabilidades que su próxima compra sea de cola 2.

1. Si actualmente una persona es comprador de cola 2, ¿cuál es la probabilidad que compre cola 1 pasadas dos compras a partir de hoy?

2. Si en la actualidad una persona es comprador de cola 1, ¿cuál es la probabilidad que compre cola 1 pasadas tres compras a partir de ahora?

Solución Consideraremos que las compras de cada una de las personas son una cadena de Markov, y que el estado en cualquier momento es el tipo de cola que compró la persona por última vez. Por lo tanto, las compras de cola por parte de cada una de las personas se pueden representar con una cadena de Markov de dos estados, donde

Estado 1 = la persona acaba de comprar cola 1
Estado 2 = la persona acaba de comprar cola 2

Sí definimos $X_n$ como el tipo de cola que compra una persona en la n-ésima compra futura (la compra actual = $X_0$), entonces $X_0$, $X_1$, ... se puede describir como la cadena de Markov con la siguiente matriz de transición:

$$P = \begin{array}{c} \\ \textit{cola}\ 1 \\ \textit{cola}\ 2 \end{array} \begin{array}{cc} \textit{cola}\ 1 & \textit{cola}\ 2 \\ .90 & .10 \\ .20 & .80 \end{array}$$

Podemos contestar ahora las preguntas 1 y 2.

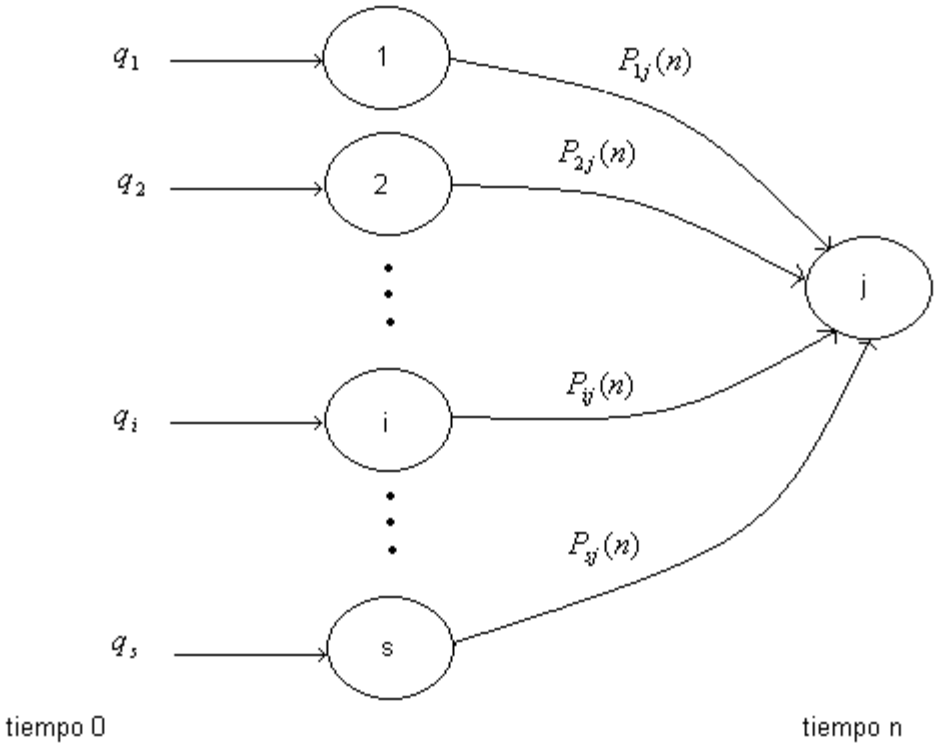
1. Se busca $P(X_2 = 1 / X_1 = 2) = P_{21}(2)$ = elemento 21 de $P^2$:

$$P^2 = \begin{bmatrix} .90 & .10 \\ .20 & .80 \end{bmatrix} \begin{bmatrix} .90 & .10 \\ .20 & .80 \end{bmatrix} = \begin{bmatrix} .83 & .17 \\ .34 & .66 \end{bmatrix}$$

Por lo tanto, $P_{21}(2) = .34$. Esto significa que hay probabilidad .34 de que la persona 2 compre cola 1, después de dos compras a partir de ahora. Con la teoría básica de probabilidad, podemos obtener esta respuesta siguiendo un camino distinto a través del teorema total de probabilidades.

En muchos casos no conocemos el estado de la cadena de Markov en el tiempo 0. Como se definió en la Sec. 1.2, sea $q_i$ la probabilidad que la cadena esté en el estado i en el tiempo 0. Entonces podemos determinar la probabilidad de que el sistema este en el estado j en e! Tiempo n mediante el siguiente razonamiento (Fig. 3):

Figura 3
Determinación dela probabilidad de estar en el estado j en el tiempo n cuando se desconoce el estado inicial.




Probabilidad de estar en el estado j en el tiempo n =

$\sum_{k=1}^{s}$ (Probabilidad que el estado original sea i ) X (Probabilidad de ir de i a j en n transiciones) =

(5) $\sum_{k=1}^{s} q_k P_{kj}(n)$ = q X Columna j de $P^n$

donde $q = [q_1, q_2, ..., q_s]$

Para mostrar el uso de la ecuación (5) contestaremos la siguiente pregunta:
supongamos que el 60% de toda la gente toma hoy cola 1 y el 40% cola 2. A tres compras a partir de ahora, ¿que fracción de los compradores estará tomando cola 1? Como q = [.60 .40] y

q(columna 1 de $P^3$) = probabilidad de que a tres compras a partir de este momento una persona tome cola 1

la probabilidad que se busca es

$$[.60 \quad .40]\begin{bmatrix} .781 \\ .438 \end{bmatrix} = .6438$$

Por lo tanto, a tres compras de este momento el 64% de las personas estará comprando cola 1.

Para mostrar el comportamiento de las probabilidades de transición en n etapas para grandes valores de n hemos calculado algunas de las probabilidades de transición de n etapas para el ejemplo de la cola y las mostramos en la siguiente tabla (Tabla 1).

Cuando n es grande, $P_{11}(n)$ y $P_{21}(n)$ son casi constantes y tienden a .67. Esto quiere decir que para n grande, independientemente del estado inicial, hay una probabilidad de .67 de que una persona compre cola 1. Igualmente. vemos que para n grande, tanto $P_{12}(n)$ como $P_{22}(n)$ son casi constantes y tienden a .33. Esto significa que para n grande, haciendo caso omiso del estado inicial, hay una probabilidad .33 de que una persona sea comprador de cola 2. En la Sección 1.5 estudiaremos con detenimiento estas tendencias de probabilidad de transición en la etapa n.

Tabla 1 Probabilidades de transición en n etapas para el ejemplo de las colas

| n | $P_{11}(n)$ | $P_{12}(n)$ | $P_{21}(n)$ | $P_{22}(n)$ |
|---|---|---|---|---|
| 1 | .90 | .10 | .20 | .80 |
| 2 | .83 | .17 | .34 | .66 |
| 3 | .78 | .22 | .44 | .56 |
| 4 | .75 | .25 | .51 | .49 |
| 5 | .72 | .28 | .56 | .44 |
| 10 | .68 | .32 | .65 | .35 |
| 20 | .67 | .33 | .67 | .33 |
| 30 | .67 | .33 | .67 | .33 |
| 40 | .67 | .33 | ,67 | .33 |

### 1.3.1 Tiempos de Parada o de primer pasaje

Un tiempo de parada se define como $T_j = \min(n>0$, tal que $X_n=j)$. Es el primer instante tal que la variable aleatoria Xn toma el valor el j, a este tiempo también será referenciado como tiempo de primer pasaje. Si definimos $P_i(T_j = m)$ como la probabilidad del que primer tiempo de parada en j, dado que estamos en el estado i, podemos redefinir la probabilidad de ir de i a jo en n pasos como

$$\text{Pij(n)} = \sum_{m=1}^{n} P_i(T_j = m) Pjj(n-m)$$

Esta definición es útil para la prueba de algunos resultados futuros.

**PROBLEMAS**

1. Cada familia norteamericana se puede clasificar como habitante de zona urbana, rural u suburbana. Durante un año determinado, el 15% de todas las familias urbanas se cambian a una zona suburbana y el 5% se cambian a una zona rural. También, el 6% de las familias suburbanas pasan a zona urbana y

el 4% se mudan a zona rural. Por último, el 4% de las familias rurales pasan a una zona suburbana y el 6% se mudan a una zona urbana.

a) Si una familia actualmente vive en una zona urbana, ¿cuál es la probabilidad que después de 2 años viva en una zona urbana? ¿En una zona suburbana? ¿En una zona rural?

(b) Supongamos que en la actualidad el 40% de las familias viven en una zona urbana, el 25% en zona suburbana y el 15% en zona rural. Después de dos años, ¿qué porcentaje de las familias norteamericanas vivirá en zona urbana?

(c) ¿Que problemas se pueden presentar si este modelo se usara para predecir la distribución futura de la población en los Estados Unidos?

2. Se pregunta lo siguiente acerca del Ejemplo 1.

(a) Después de jugar dos veces, ¿cuál es la probabilidad que tenga 3 dólares? ¿Cuál la de que tenga 2 dólares?

(b) Después de jugar tres veces, ¿cuál es la probabilidad que tenga 2 dólares?

3. En el Ejemplo a 2, determine las siguientes probabilidades de transición en n etapas:

(a) Después de haber pintado 2 bolas, ¿cuál es la probabilidad que el estado sea [0 2 0]?

(b) Después de haber pintado tres bolas, ¿cuál es la probabilidad que el estado sea [0 1 1]?

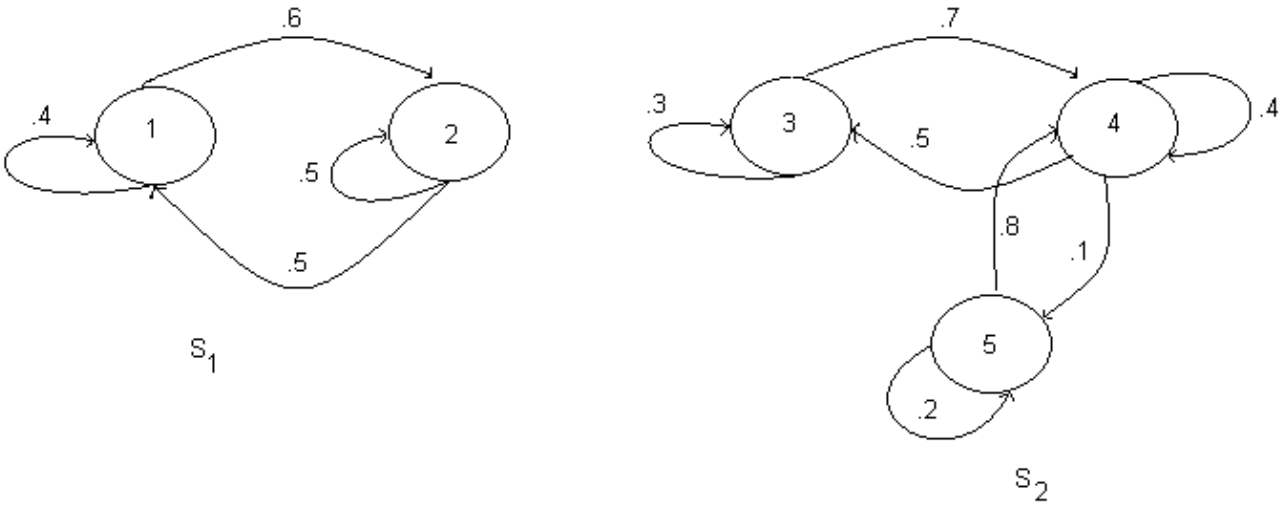
# 1.4 CLASIFICACIÓN DE ESTADOS DE UNA CADENA DE MARKOV

En la sesión 1.3 se menciono que después de muchas transiciones, las probabilidades de transición de n etapas tienden a estabilizarse . Antes de poder describir esto con más detalle, necesitamos estudiar cómo los matemáticos clasifican los estados de una cadena de Markov. La siguiente matriz de transición se usará para mostrar la mayoría de las definiciones siguientes (Fig. 4):

$$P = \begin{bmatrix} .4 & .6 & 0 & 0 & 0 \\ .5 & .5 & 0 & 0 & 0 \\ 0 & 0 & .3 & .7 & 0 \\ 0 & 0 & .5 & .4 & .1 \\ 0 & 0 & 0 & .8 & .2 \end{bmatrix}$$

DEFINICIÓN: Dados dos estados i y j, la trayectoria de i a j es la sucesión de transiciones que comienza en i y termina en j, de modo que cada transición de la secuencia tenga probabilidad positiva de presentarse.

Figura 4
Representación gráfica de la matriz de transición




Otra manera de ver esta definición es que el tiempo de primer pasaje es un número finito o sea que la probabilidad de llegar a *j*, dado que se está en *i*, en una cantidad de finita de pasos sea positiva: $P_i(T_j < \infty) > 0$

DEFINICIÓN un estado es **alcanzable** desde un i si hay una trayectoria que vaya de i a j.

DEFINICIÓN Se dice que dos estados i y j se **comunican** si j es alcanzaba desde i, e i es alcanzable desde j.

Para la matriz P de probabilidad de transición representada en la Fig. 4, el estado 5 es alcanzable desde el estado 3 (a través de la trayectoria 3-4-5), pero el estado 5 no es alcanzable desde el estado 1 (no hay trayectoria que vaya de 1 a 5 en la Fig. 6). También, los estados 1 y 2 se comunican: podemos pasar de 1 a 2 y de 2 a 1.

DEFINICIÓN Un conjunto de estados S en una cadena de Markov es **conjunto cerrado** si ningún estado fuera de S es alcanzable desde un estado en S.

De la cadena de Markov con la matriz P de la FÍg. 4, tanto S1 = {1,2} como S2 = {3, 4, 5} son conjuntos cerrados. Observe que una vez que entramos a un conjunto cerrado no podemos dejarlo nunca. En la Fig. 4 ningún arco comienza en $S_1$ y termina en $S_2$ o principia en $S_2$ y termina en $S_1$.

DEFINICIÓN Un estado i es un estado **absorbente** si $p_{ij} = 1$. O sea $P_j(T_j = 1) = 1$

Siempre que entramos a un estado de absorción, nunca lo podremos dejar. En el Ejemplo 1, la ruina del jugador, los estados 0 y 4 son absorbentes. Es natural que un estado absorbente sea un conjunto cerrado que sólo contenga un estado.

DEFINICIÓN Un estado i es un estado **transitorio** si hay un estado j alcanzable desde i, pero el estado i no es alcanzable desde el estado j. Esto es lo mismo que afirmar que $P_i(T_j < \infty) < 1$, no siempre existe una cantidad finita de pasos para alcanzar a *j* desde *i*.

En otras palabras, un estado i es transitorio si hay manera de dejar el estado i de tal modo que nunca se regrese a él. En el ejemplo de la ruina del jugador, los estados 1, 2 y 3 son estados transitorios. Por ejemplo (FÍg. 1), desde el estado 2 es posible pasar por la trayectoria 2-3-4. pero no hay modo de regresar al estado 2 desde el estado 4. Igualmente, en el Ejemplo 2, [2 0 0], [1 1 0] y [l 0 1] son estados transitorios. Hay una trayectoria desde [1 0 1] a [0 0 2], pero una vez que se hayan pintado ambas bolas, no hay manera de regresara [1 0 1].

Después de un gran número de periodos, la probabilidad de encontrarse en cualquier estado de transición i es cero. Cada vez que entramos a un estado i de transición, hay una probabilidad positiva de dejar i para siempre y terminar en el estado j descrito en la definición de estado transitorio. Así, al final, tenemos la seguridad de entrar al estado j (y en ese caso nunca regresaremos al estado i). Así, suponga que en el Ejemplo 2 nos encontramos en el estado transitorio [1 0 1]. Con probabilidad 1, la bola no pintada la pintaremos finalmente y nunca regresaremos a ese estado [1 0 1].

DEF1NÍCIÓN Si un estado no es transitorio, se llama estado **recurrente**. Esto es lo mismo que afirmar que $P_i(T_j < \infty) = 1$, o sea existe una cantidad finita de pasos para llegar a j desde i.

En el Ejemplo 1, los estados 0 y 4 son estados recurrentes (y también estados absorbentes). En el Ejemplo 2, [0 2 0], [0 0 2[ y [0 1 1] son estados recurrentes. Para la matriz de transición de la Fig. 4, todos los estados son recurrentes.

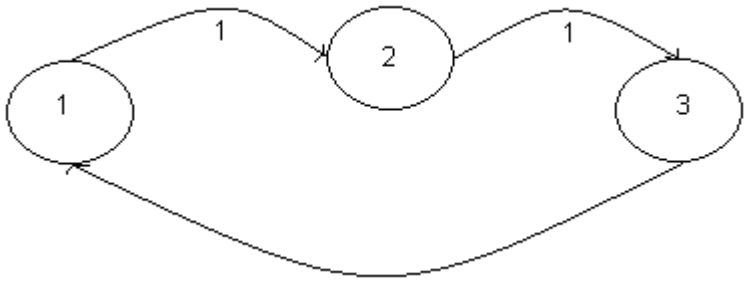
DEFINICIÓN Un estado i es **periódico** con periodo k >1 si k es el menor numero tal que todas las trayectorias que parten del estado i y regresan al estado i tienen una longitud múltiplo de k. Si un estado recurrente no es periódico, se llama **aperiódico.**

Para la cadena de Markov cuya matriz de transición es

$$Q = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

cada estado tiene periodo 3. Por ejemplo, si comenzamos en el estado 1, la única manera de regresar a ese estado es seguir la trayectoria 1-2-3-1 durante digamos m veces (Fig. 5). Por lo tanto, cualquier regreso al estado 1 tomará 3m transiciones, de modo que el estado 1 tiene periodo 3. Donde nos encontremos, tenemos la seguridad de regresar allí tres periodos después.

Figura 5
Cadena periódica de Markov con k = 3




DEFINICIÓN Si todos los estados de una cadena son recurrentes, aperiódicos y se comunican entre si, se dice que la cadena es **ergódica**.

El ejemplo de la ruina del jugador no es cadena ergódica porque, por ejemplo, los estados 3 y 4 no se comunican. El Ejemplo 2 tampoco es una cadena ergódica porque, por ejemplo, [2 0 0] y [0 1 1] no se comunican. El Ejemplo 4, el ejemplo de la cola, es cadena ergódica de Markov. De las siguientes tres cadenas de Markov, $P_1$ y $P_3$ son ergódicas y $P_2$ no es ergódica.

$$P_1 = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{bmatrix}$$

$$P_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{bmatrix}$$

$$P_3 = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix}$$

P2 no es ergódica porque hay dos clases cerradas de estados (la clase 1 = {1, 2} y la clase 2 = {3, 4}) y los estados en clases diferentes no se comunican entre sí.

**PROBLEMAS**

1. En el Ejemplo 1, ¿cuál es el periodo de los estados 1 y 3?

2. La cadena de Markov de la Secc. 1.3, Problema 1, ¿es ergódica?

3. Se tiene ]a siguiente matriz de transición:

| | | | | | |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 1 | 0 |
| ¼ | ¼ | 0 | ½ | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1⁄3 | 0 | 0 | 0 | 2/3 |

(si) ¿Cuáles estados son transitorios?

(b) ¿Cuáles estados son recurrentes?

(c) Identifique lodos los conjuntos cerrados de estados.

(d) ¿Es ergódica esa cadena?

4. Para cada una de las siguientes matrices, determine si la cadena de Markov es esgódica. También, para cada cadena, determine los estados recurrente, transitorio y absorbente.

a. $\begin{bmatrix} 0 & .8 & .2 \\ .3 & .7 & 0 \\ .4 & .5 & .1 \end{bmatrix}$ b. $\begin{bmatrix} .2 & .8 & 0 & 0 \\ 0 & 0 & .9 & .1 \\ .4 & .5 & .1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

5. En la Serie Mundial de Poker de 1980 participaron 54 jugadores. Cada uno de ellos comenzó con 10 000 dólares. Los juegos continuaron hasta que uno de los jugadores ganó lodo el dinero de los demás. Si se modela esta Serie Mundial como cadena de Markov, ¿cuántos estados absorbentes tendría esa cadena?

6. ¿Cuál de las siguientes cadenas es ergódica?

a. $\begin{bmatrix} .4 & 0 & .6 \\ .3 & .3 & .4 \\ 0 & .5 & .5 \end{bmatrix}$ b. $\begin{bmatrix} .7 & 0 & 0 & .3 \\ .2 & .2 & .4 & .2 \\ .6 & .1. & .1 & .2 \\ .2 & 0 & 0 & .8 \end{bmatrix}$

# 1.5 PROBABILIDADES DE ESTADO ESTABLE Y TIEMPOS MEDIOS DE PRIMER PASAJE

En nuestra descripción del ejemplo de la Cola encontramos que después de largo tiempo, la probabilidad de que la siguiente compra de una persona fuera de cola 1 tendía a .67, y la de que la compra siguiente fuera de cola 2 tendía a .33 (Tabla 1). Estas probabilidades no dependieron de si la persona era al principio tomador de cola 1 o de cola 2. En esta sección describiremos el importante concepto de probabilidades de estado estable, que se pueden usar para describir el comportamiento de una cadena de Markov a largo plazo.

El resultado siguiente es vital para comprender las probabilidades de estado estable y el comportamiento a largo plazo de cadenas de Markov.

TEOREMA 1 Sea P una matriz de transición de una cadena ergódica de s estados. Existe entonces un vector $\pi = [\pi_1, \pi_2, \cdots \pi_s]$ tal que

$$\lim_{n\to\infty} P^n = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_s \\ \pi_1 & \pi_2 & \cdots & \pi_s \\ \vdots & & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_s \end{bmatrix}$$

(La prueba de este teorema se encuentra en anexo al final de este documento)

Recuerde que el ij-ésimo elemento de $P^n$ es $P_{ij}(n)$. El teorema 1 establece que para cualquier estado inicial i,

$$\lim_{n\to\infty} P_{ij}(n) = \pi_j$$

Observe que para n grande, $P^n$ tiende a una matriz con renglones idénticos. Esto quiere decir que después de largo tiempo, la cadena de Markov se estabiliza e, independientemente del estado inicial i, hay una probabilidad $\pi_j$, de que nos encontremos en el estado j.

El vector $\pi = [\pi_1, \pi_2, \cdots \pi_s]$ a menudo se llama distribución de estado estable, o también distribución de equilibrio para la cadena de Markov. ¿Cómo podemos encontrar la distribución de probabilidades de estacionario para una cadena dada cuya matriz de transición es P? Según el teorema 1, para n grande y para toda i,

(6) $$P_{ij}(n+1) \cong P_{ij}(n) \cong \pi_j$$

Como $P_{ij}\{n + 1)$ = (renglón i de $P^n$) X (columna j de P), podemos escribir

(7) $$P_{ij}(n+1) = \sum_{k=1}^{s} P_{ik}(n) P_{kj}$$

Si n es grande, al sustituir la ecuación (6) en la (7) se obtiene

(8) $$\pi_{ij} = \sum_{k=1}^{s} \pi_k P_{kj}$$

En forma matricial, la ecuación (8) se puede escribir como:

(8') $$\pi = \pi P$$

Desafortunadamente, el sistema de ecuaciones que especifica la ecuación (8) tiene un número infinito de soluciones, porque el rango de la matriz P siempre resulta ser ≤ s- 1. Para obtener valores únicos de probabilidades de estado estable, note que para toda n y toda i,

(9) $$P_{i1}(n) + P_{i2}(n) + \cdots + P_{is}(n) = 1$$

AI hacer que n tienda al infinito en la Ecuación. (9), obtenemos

(10) $$\pi_1 + \pi_2 + \cdots + \pi_s = 1$$

Así, después de reemplazar cualquiera de las ecuaciones (8) por (10), podemos usar la ecuación (8) para despejar las probabilidades de estado estable.

Para mostrar cómo determinar las probabilidades de estado estable, las calcularemos para el Ejemplo. 4, el de la Cola. Recuerde que la matriz tic transición de ese ejemplo era

$$P = \begin{bmatrix} .90 & .10 \\ .20 & .80 \end{bmatrix}$$

Entonces las ecuaciones (8) u (8') producen

$$\begin{bmatrix} \pi_1 & \pi \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi \end{bmatrix} \begin{bmatrix} .90 & .10 \\ .20 & .80 \end{bmatrix}$$

$$\pi_1 = .90\pi_1 + .20\pi_2$$
$$\pi_2 = .10\pi_1 + .80\pi_2$$

AI reemplazar la segunda ecuación por la condición $\pi_1 + .\pi_2 = 1$, obtenemos el sistema

$$\pi_1 + .\pi_2 = 1$$
$$\pi_1 = .90\pi_1 + .20\pi_2$$

Al despejar $\pi_1$ y $\pi_2$, resulta que $\pi_1$= 2/3 y $\pi_2$=1/3. Por lo tanto, después de largo tiempo, hay probabilidad 2/3 de que una persona dada compre cola 1 y 1/3 de probabilidad de que una persona dada compre cola 2.

## ANÁLISIS DE ESTADO TRANSITORIO

Un vistazo a la Tabla 1 muestra que para el Ejemplo. 4 se alcanza el estado estable, a dos cifras decimales, sólo después de 10 transiciones. No se puede dar una regla general acerca de que tan rápido alcanzan las cadenas de Markov el estado estable pero si P contiene muy pocos elementos que queden cerca de 0 o de 1, en general, se alcanza en forma muy rápida el estado estable. El comportamiento de una cadena de Markov antes de alcanzar el estado estable se llama comportamiento transitorio (o a plazo corto). Para estudiar el comportamiento transitorio de una cadena de Markov, tan sólo se usan las fórmulas para $P_{ij}(n)$ de las ecuaciones (4) y (5). Sin embargo es bueno saber que para n grande, las probabilidades de estado estable describen con exactitud la probabilidad de encontrarse en un estado determinado.

## INTERPRETACIÓN INTUITIVA DE LAS PROBABILIDADES DE ESTADO ESTABLE

Se puede dar una interpretación intuitiva de las ecuaciones (8) de probabilidad de estado estable. Al restar $\pi_j p_{ij}$, de ambos lados de (8) se obtiene

(11) $$\pi_j(1 - p_{ij}) = \sum_{k \neq i} \pi_j p_{kj}$$

La ecuación (11) dice que en el estado estable,

Probabilidad de que una transición determinada deje el estado j

= probabilidad de que una transición determinada entre al estado j (12)

Recuérdese que en el estado estable, la probabilidad de que el sistema este en el estado j es $\pi_j$, Según esta observación se concluye que

Probabilidad de que una transición particular deje el estado j
= (probabilidad de que el periodo actual comience en j)
x (probabilidad de que la transición actual deje j)
$= \pi_j(1 - p_{ij})$

y

Probabilidad de que determinada transición entre al estado j

= $\sum_k$ (probabilidad de que el periodo actual comience en k≠i)

x (probabilidad de que la transición actual entre a j)

= $\sum_{k \neq i} \pi_j p_{kj}$

Es aceptable la ecuación (11 ). Si fuera violada para cualquier estado, entonces para un estado j el lado derecho de (11) sería mayor que el lado izquierdo. Esto ocasionaría una probabilidad de "acumulación" en el estado j y no existiría una distribución de estado estable. Se puede considerar que la ecuación (11) dice que en el estado estable, el "flujo" de probabilidad hacia cada estado debe ser igual al flujo de probabilidad que sale de cada estado. Esto explica por qué las probabilidades de estado estable se llaman con frecuencia probabilidades de equilibrio.

## USO DE LAS PROBABILIDADES DE ESTADO ESTABLE PARA TOMAR DECISIONES

EJEMPLO 5 Suponga, en el Ejemplo 4, que cada cliente hace una compra de cola durante cualquier semana (52 semanas = 1 año). Suponga que hay 100 millones de clientes de cola. La producción de una unidad de venta de cola cuesta 1 dólar y se vende a 2 dólares. Una empresa de publicidad garantiza, por 500 millones de dólares al año, un decremento al 5% de la fracción de consumidores de cola 1, que se cambian a cola 2 después de una compra. ¿Debe contratar a la empresa de publicidad la compañía que fabrica la cola 1?

En la actualidad, una fracción $\pi_1 = 2/3$ de todas las compras son de cola 1. Cada compra de cola 1 le deja al fabricante 1 dólar. Como hay un total de 52(100 000 000) = 5 200 000 000 de compras de cola cada año, las ganancias actuales del fabricante de cola 1, al año, son

2/3 * (5 200 000 000) = 3 466 666 667 dólares

La empresa de publicidad ofrece cambiar la matriz P a

$$P_1 = \begin{bmatrix} .95 & .05 \\ .20 & .80 \end{bmatrix}$$

Para $P_1$, las ecuaciones de estado estable se transforman en

$$\pi_1 = .95\pi_1 + .20\pi_2$$
$$\pi_2 = .05\pi_1 + .80\pi_2$$

Al reemplazar la segunda ecuación por $\pi_1 + .\pi_2 = 1$ y despejar, obtenemos $\pi_1 = .8$ y $\pi_2 = .2$. En este caso, la ganancia anual de la productora de cola 1 será

(.80)(5200000000) - 500000000 = 3 660 000 000 dólares

Por lo tanto, el fabricante de cola 1 debe contratar la agencia de publicidad.

EJEMPLO 6. Bajo la hipótesis de que el jugador de Monopoly que vaya a la cárcel se quede allí hasta que se saque números dobles o pasen tres turnos, la probabilidad de estado estable que tiene un Jugador de caer en cualquier cuadro de Monopoly fue determinado por Ash y Bishop (1972) (Tabla 2) Ests probabilidades de estado estable se pueden emplear para medir la eficacia de diversos monopolios con respecto al costo. Por ejemplo, cuesta 1 500 dólares construir hoteles en el monopolio anaranjado. Cada vez que un jugador cae en Tcnnessee Ave o en un hotel de St. James Place, el propietario del monopolio recibe 950 dólares, y cada vez que un jugador cae en un hotel de New York Ave., el propietario recibe 1 000 dólares. De la Tabla 3 podemos calcular la renta esperada por tirada de dados que gana el monopolio anaranjado:

950(.0335) + 950(.0318) + 1 000(.0334) = 95.44 dólares

Así, por cada dólar invertido, el monopolio anaranjado da 95.44/1500 = 0.064 de dólar por tirada de dados.

Veamos ahora al monopolio verde. Poner hoteles en el monopolio verde cuesta 3 000 dólares. Si un Jugador cae cu un hotel de North Carolina Ave. o de Pacific Ave., el propietario recibe 1 275 dólares. Si un jugador cae en un hotel de Pennsylvania Ave. el propietario recibe 1 400 dólares. De acuerdo con la Tabla 3, la ganancia promedio por tirada de dados que se gana con los hoteles en el monopolio verde es

1275(.0294) + 1275(.0300) + 1400(.0279) = 114.80 dólares

Así, por cada dólar invertido, el monopolio verde da sólo 114.80/3000 = 0.038 de dólar por tirada de dados.

Este análisis muestra que el monopolio anaranjado es mejor que el verde. Por cierto ¿por que caen los jugadores con tanta frecuencia en el monopolio anaranjado?

Tabla 2 Probabilidades de estado estable para Monopoly

| | |
|---|---|
| 0 Go | 0.0346 |
| 1 Mediterranean Ave. | 0.0237 |
| 2 Community Chest 1 | 0.0218 |
| 3 Baltic Ave. | 0.0241 |
| 4 Income tax | 0.0261 |
| 5 Rcading RR | 0.0332 |
| 6 Oriental Ave. | 0.0253 |
| 7 Chance 1 | 0.0096 |
| 8 Vermont Ave, | 0.0258 |
| 9 Connecticut Ave. | 0.0237 |
| 10 Visiting jail | 0.0254 |
| 11 St. Charles Place | 0.0304 |
| 12 Electric Co. | 0.0311 |
| 13 Statc Ave. | 0.0258 |
| 14 Virginia Ave. | 0.0288 |
| 15 Pennsylvania RR | 0.0313 |
| 16 St. James Place | 0.0318 |
| 17 Community Chest 2 | 0.0272 |
| 18 Tennessee Ave. | 0.0335 |
| 19 New York Ave. | 0.0334 |
| 20 Free parking | 0.0336 |
| 21 Kentucky Ave. | 0.031 |
| 22 Chance 2 | 0.0125 |
| 23 Indiana Ave. | 0.0305 |
| 24 Illinois Ave. | 0.0355 |
| 25 B and 0 RR | 0.0344 |
| 26 Atlantic Ave. | 0.0301 |
| 27 Ventnor Ave. | 0.0299 |
| 28 Water works | 0.0315 |
| 29 Marvin Gardens | 0.0289 |
| 30 Jail | 0.1123 |
| 31 Pacific Ave. | 0.03 |
| 32 North Carolina Ave. | 0.0294 |
| 33 Community Chest 3 | 0.0263 |
| 34 Pennsylvania Ave. | 0.0279 |
| 35 Short Une RR | 0.0272 |
| 36 Chance 3 | 0.0096 |
| 37 Park Place | 0.0245 |
| 38 Luxury tax | 0.0245 |
| 39 Boardwalk | 0.0295 |

**TIEMPOS PROMEDIO DE PRIMER PASAJE**

En una cadena ergódica, sea $m_{ij}$ = número esperado de transiciones antes de alcanzar por primera vez el estado j, dado que estamos actualmente en el estado i. $m_{ij}$ se llama tiempo promedio de primer pasaje del estado i al estado j. En el Ejemplo 4, $m_{12}$ sería el número esperado de botellas de cola que adquiere un comprador de cola 1, antes de comprar una botella de cola 2. Suponga que estamos ahora en el estado i. Entonces, con probabilidad $p_{ij}$, necesitaremos una transición para pasar del estado i al estado j. Para $k \neq j$ pasamos a continuación, con probabilidad $p_{ik}$, al estado k. En este caso, se necesitara un promedio de $1 + m_{kj}$, transiciones para pasar de k a j. Este modo de pensar indica que

$$m_{ij} = p_{ij}(1) + \sum_{k \neq j} p_{ik}(1 + m_{kj})$$

Como

$$p_{ij} + \sum_{k \neq j} p_{ik} = 1$$

podernos reformular la última ecuación como

(13) $$m_{ij} = \sum_{k \neq j} p_{ik}(m_{kj}) + 1$$

Al resolver las ecuaciones lineales representadas en (13), podemos encontrar todos los tiempos promedios de primer pasaje. Se puede demostrar que

$$m_{ii} = \frac{1}{\pi_i}$$

Con ello se puede simplificar el uso de las ecuaciones (13).

Para mostrar el uso de ellas, despejaremos los tiempos promedio de primer pasaje en el Ejemplo 4. Recordemos que . $\pi_1 = 2/3$ y $\pi_2 = 1/3$. Entonces

$$m_{11} = \frac{1}{\frac{2}{3}} = 1.5 \qquad m_{22} = \frac{1}{\frac{1}{3}} = 3$$

Entonces (13) da en las dos ecuaciones siguientes:

$$m_{12} = 1 + p_{11}m_{12} = 1 + .9m_{12}$$
$$m_{21} = 1 + p_{22}m_{21} = 1 + .8m_{21}$$

Resolviendo esas ecuaciones encontramos que $m_{12}$ = 10 y $m_{21}$ = 5. Esto quiere decir que, por ejemplo, una persona que había tomado cola 1 tomará un promedio de diez botellas de refresco antes de cambiar a cola 2.

**PROBLEMAS**

1. Determine las probabilidades de estado estable para el Problema 1 de la Sección 1-3.

2. En el problema de la ruina del Jugador, ¿por que no es razonable hablar de probabilidades de estado estable?

3. Para cada una de las siguientes cadenas de Markov, determine la fracción de las veces, a largo plazo, que se ocupará cada estado.

(a) $\begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$ (b) $\begin{bmatrix} .8 & .2 & 0 \\ 0 & .2 & .8 \\ .8 & .2 & 0 \end{bmatrix}$

(c) Determine todos los tiempos promedio de primer pasaje de! inciso (b).

4. Al principio de cada año, mi automóvil está en buen, regular o mal estado. Un buen automóvil será bueno al principio del año siguiente, con probabilidad .85, regula con probabilidad .10 y mal con probabilidad .05. Un automóvil regular estará regular al principio del año siguiente con probabilidad .70 y mal con probabilidad .30. Cuesta 6000 dólares comprar un buen automóvil, uno regular se puede conseguir por 1000 dólares; uno malo no tiene valor de venta, y se debe reemplazar de inmediato por uno bueno. Cuesta 1 000 dólares al año el funcionamiento de un buen automóvil, y 1 500 dólares el de uno regular. ¿Debe reemplazar mi automóvil tan pronto como se vuelve regular, o debo esperar hasta que se descomponga?
Suponga que el costo de funcionamiento de un automóvil durante un año depende del tipo de vehículo que se tiene a la mano a principio del año (después de llegar cualquier auto nuevo, si es el caso).

5. Se dice que una matriz cuadrada es doblemente estocástica si todos sus elementos son no negativos y los elementos de cada renglón y cada columna suman 1.
Para cualquier matriz ergódica y doblemente estocástica, demuestre que todos los estados tienen la misma probabilidad de estado estable.

6. Este problema mostrará porque las probabilidades de estado estable se llaman a veces probabilidades estacionarias. Sean $\pi_1, \pi_2, \cdots \pi_s$ las probabilidades de estado estable para una cadena ergódica con matriz P de transición. Suponga también que la cadena de Markov comienza en el estado i con probabilidad $\pi_i$.

a) Cual es la probabilidad que después de una transición el sistema se encuentre en el estado i? Sugerencia: Usar la ecuación (8).

b) Para cualquier valor de n (n = 1, 2,. . .), ¿cuál es la probabilidad de que una cadena de Markov se encuentre en el estado i después de n transiciones?

(c) ¿Por que a las probabilidades de estado estable se les llama a veces probabilidades estacionarias?

7. Se tienen dos acciones. Las acciones 1 siempre se venden a 10 dólares o 20 dólares. Si hoy las acciones 1 se venden a 10 dólares, hay una probabilidad .80 de que mañana se vendan a 10 dólares. Si las acciones 1 se venden hoy a 20 dólares, hay una probabilidad .90 de que mañana se vendan a 20 dólares. Las acciones 2 siempre se venden a 10 dólares o a 35 dólares. Si se venden hoy a 10 dólares, hay una probabilidad .90 de que se vendan mañana a 10 dólares. Si se venden hoy a
25 dólares, hay una probabilidad .85 de que mañana se vendan a 25 dólares. En promedio, ¿que acciones se venden a mayor precio? Determine e interprete todos los tiempos promedio de primer pasaje.

8. La compañía de seguros Payoff cobra a sus clientes de acuerdo a su historia de accidentes. Un cliente que no haya tenido accidentes durante los últimos dos años paga 100 dólares de prima anual. Quien haya tenido un accidente en cada uno de los dos últimos años paga una prima anual de 400 dólares. A los que hayan tenido un accidente durante sólo uno de los últimos dos años se les cobra una prima anual de 300 dólares. Un cliente que tuvo un accidente durante el último año tiene una probabilidad de 10% de accidentarse durante este año.
Si un cliente no ha tenido un accidente durante el último año, tiene una probabilidad de 3% de sufrir un accidente durante este año. Durante un año dado, ¿cuál es la prima que paga en promedio un cliente de Payoff?
{Sugerencia: En caso de dificultad, pruebe con una cadena de Markov de cuatro estados.)

9. Se tiene la siguiente cadena no ergódica;

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix}$$

a) ¿Por qué esta cadena es no ergódica?

b) Explique porqué falla el teorema 1 en esta cadena.
Sugerencia: Determine si es cierta la siguiente ecuación;

$$\lim_{n\to\infty} P_{12}(n) = \lim_{n\to\infty} P_{32}(n)$$

(c) A pesar del hecho que falla el teorema 1, determine

$$\lim_{n\to\infty} P_{13}(n) \qquad \lim_{n\to\infty} P_{21}(n)$$

$$\lim_{n\to\infty} P_{43}(n) \qquad \lim_{n\to\infty} P_{41}(n)$$

10. Se tiene la siguiente cadena no ergódica:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

(a) ¿Por que esta cadena es no ergódica?

(b) Explique por que el teorema 1 falla para esta cadena. Sugerencia: Demuestre que no existe $\lim_{n\to\infty} P_{11}(n)$ al hacer una lista del comportamiento que sigue $P_{11}(n)$ a medida que aumenta n.

## 1.6 CADENAS ABSORBENTES

Muchas aplicaciones interesantes de las cadenas de Markov incluyen cadenas en las que algunos de los estados son absorbentes y el resto son transitorios. A esas cadenas se les llama cadenas absorbentes. Veamos una cadena absorbente de Markov: si comenzamos en un estado transitorio, entonces al final tendremos la seguridad de dejar el estado transitorio y terminar en uno de los estados absorbentes. Para ver por qué nos interesan las cadenas absorbentes, describiremos las siguientes dos:

EJEMPLO 7 Cuentas por cobrar El estado de cuentas por cobrar en una empresa se modela con frecuencia como cadena absorbente de Markov. Suponga que una empresa supone que una cuenta es incobrable si han pasado más de dos meses de su fecha de vencimiento. Entonces, al principio de cada mes, se puede clasificar cada cuenta en uno de los siguientes estados específicos:

**Estado 1** Cuenta nueva.

**Estado 2** Los pagos de la cuenta están retrasados un mes.

**Estado 3** Los pagos de la cuenta están retrasados dos meses.

**Estado 4** Los pagos de la cuenta están retrasados tres meses.

**Estado 5** Se ha saldado la cuenta.

**Estado 6** Se ha cancelado la cuenta por ser mal pagado

Supongamos que los últimos datos indican que la siguiente cadena de Markov describe como cambia el estado de una cuenta de un mes al siguiente:

$$\begin{array}{r} \textit{Nueva} \\ 1\ \textit{mes} \\ 2\ \textit{meses} \\ 3\ \textit{meses} \\ \textit{Pagada} \\ \textit{Incobrable} \end{array} \begin{bmatrix} 0 & .6 & 0 & 0 & .4 & 0 \\ 0 & 0 & .5 & 0 & .5 & 0 \\ 0 & 0 & 0 & .4 & .6 & 0 \\ 0 & 0 & 0 & 0 & .7 & .3 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Por ejemplo, si al principio de un mes una cuenta lleva dos meses de vencida, hay 40% de probabilidades de que no se pague al principio del mes siguiente y, por lo tanto, que tenga tres meses de retraso y una probabilidad de 60% de que se pague.

Para simplificar el ejemplo, supondremos que después de tres meses, la cuenta o se cobra o se considera incobrable.

Una vez que una deuda se paga o se considera incobrable, se cierra y no se tienen más transiciones. Por lo tanto. Pagada e Incobrable son estados absorbentes. Como toda cuenta al final o se paga o se considera incobrable, las cuentas Nueva, 1 mes, 2 meses y 3 meses son estados transitorios. Por ejemplo, una cuenta vencida hace 2 meses puede seguir la trayectoria 2 meses-pagada, pero no hay regreso posible de Pagada a 2 meses.

Una cuenta nueva normal será absorbida ya sea como pagada o como incobrable. Una pregunta de mayor interés es: ¿cuál es la probabilidad de que una cuenta nueva Finalmente se pueda cobrar? Más adelante en esta sección se encontrará la respuesta.

Ejemplo. Planificación de personal la empresa de abogados Masón y Burger emplea a tres categorías de abogados: principiantes, con experiencia y socios. Durante un año determinado hay una probabilidad .15 que un abogado principiante sea ascendido a abogado con experiencia y una probabilidad .05 que deje la empresa. También, hay una probabilidad .20 que un abogado con experiencia sea ascendido a socio y una probabilidad .10 que deje la empresa. También hay una probabilidad .05 que un socio deje la empresa. La empresa nunca degrada a un abogado.

Surgen muchas preguntas interesantes que la empresa podría contestar. Por ejemplo, ¿cuál es la probabilidad que un abogado principiante recién contratado se vaya antes de ser socio? En promedio, ¿cuánto tiempo permanece- un abogado principiante recién contratado con la empresa? Las respuestas se deducirán después en esta sección.

Modelaremos la trayectoria de un abogado en Masón y Burger como cadena absorbente de Markov con la siguiente matriz de probabilidad cíe transición:

| | Pr*incipiante* | *Experimentad* | *Asociado* | *Sale* sin *ser socio* | *Sale siendo ser socio* |
|---|---|---|---|---|---|
| Pr*incipiante* | .80 | .15 | 0 | .05 | 0 |
| *Experimentado* | 0 | .70 | .20 | .10 | 0 |
| *Asociado* | 0 | 0 | .95 | 0 | .05 |
| *Sale* sin *ser socio* | 0 | 0 | 0 | 1 | 0 |
| *Sale siendo ser socio* | 0 | 0 | 0 | 0 | 1 |

Los dos últimos estados son estados absorbentes y los demás son transitorios. Por ejemplo. Experimentado es estado transitorio, porque hay una trayectoria de Experimentado a Sale sin ser socio, pero no hay trayectoria que regrese de Sale sin ser socio a Experimentado. Suponemos que una vez que un ahogado sale de la empresa nunca regresa.

Para toda cadena absorbente se desea conocer: (1) Si la cadena comienza en un estado determinado transitorio, y años de alcanzar un estado absorbente, ¿cuál es el número esperado de veces que se llegara a un estado? ¿Cuántos periodos esperamos pasar en un determinado estado transitorio antes que se efectué la absorción?
(2) Si una cadena inicia en un estado transitorio dado, ¿cuál es la probabilidad de terminar en cada uno de los estados absorbentes?

Para contestar estas preguntes necesitamos formular la matriz de transición con los estados en una lista con el siguiente orden: primero los estados transitorios y después los absorbentes. Para precisar, se supondrá que hay s - m estados transitorios ($t_1$, $t_2$,..., $t_{s-m}$) y m estados absorbentes ($a_1$, $a_2$, ... , $a_m$). Entonces la matriz de transición para la cadena de absorción puede escribirse como sigue:

| | | *s – m columnas* | *m columnas* |
|---|---|---|---|
| *s – m* | *renglones* | *Q* | *R* |
| *m* | *renglones* | 0 | *I* |

En este formato, los renglones y las columnas de P corresponden, en orden, a los estados ($t_1$, $t_2$,..., $t_{s-m}$) ($a_1$, $a_2$, ... , $a_m$). En este caso, I es una matriz identidad m x m, que refleja el hecho de que nunca podemos dejar un estado absorbente; Q es una matriz (s - m) x (s - m) que representa las transiciones entre los estados transitorios; R es una matriz (s - m) x m que representa las transiciones desde los estados transitorios a los estados absorbentes; 0 es una matriz m x {s - m) que consta de ceros. Esto refleja el hecho de que es imposible ir de un estado absorbente a uno transitorio.

Aplicando esta notación al Ejemplo 7, tenemos que
$t_1$= Nueva
$t_2$= 1 mes
$t_3$= 2 meses
$t_4$= 3 meses
$a_1$= Pagada
$a_2$= Incobrable

Entonces, para ese ejemplo, las partes de la matriz de probabilidad de transición se puede expresar como (s = 6, m =2)

$$Q = \begin{bmatrix} 0 & .6 & 0 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & 0 & .4 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad R = \begin{bmatrix} .4 & 0 \\ .5 & 0 \\ .6 & 0 \\ .7 & .3 \end{bmatrix}$$

Para el Ejemplo 8, sean

$t_1$= Principiante
$t_2$= Experimentado
$t_3$= Socio
$a_1$= Sale sin ser socio
$a_2$= Sale siendo socio

y podemos escribir las partes de la matriz de probabilidad de transición como

$$Q = \begin{bmatrix} .80 & .15 & 0 \\ 0 & .70 & .20 \\ 0 & 0 & .95 \end{bmatrix} \quad R = \begin{bmatrix} .05 & 0 \\ .10 & 0 \\ 0 & .05 \end{bmatrix}$$

Podemos ahora investigar algunos hechos acerca de las cadenas absorbentes (Keineny y Snell(1960)): (1) Si la cadena comienza en un determinado estado transitorio, y antes de alcanzar un estado absorbente, ¿cuál es entonces el número esperado de veces que se entrará en cada estado? ¿Cuántos periodos esperamos pasar en un estado transitorio dado antes de que se lleve a cabo la absorción? Respuesta: Si en este momento estamos en el estado transitorio $t_i$ el número esperado de periodos que pasarán en un estado transitorio $t_j$, antes de la absorción es el ij-ésimo elemento de la matriz $(I- Q)^{-1}$. Para una demostración vea e! Problema 8 al final de esta sección. (2) Si una cadena inicia en un estado transitorio dado, ¿qué probabilidad hay de terminar en cada uno de los estados absorbentes? Respuesta: Si en este momento estamos en un estado transitorio i, la probabilidad de ser absorbidos finalmente por un estado absorbente $a_j$ es el ij-ésimo elemento de la matriz $(I –Q)^{-1}R$. Para una demostración vea el Problema 9 al final de esta sección.

La matriz $(I- Q)^{-1}$ a menudo se llama matriz fundamental de la cadena de Markov. El lector que se interese en proseguir el estudio de cadenas de absorción debe consultar Kemeny y Snell (1960).

Continuación del ejemplo de Cuentas por cobrar

1. ¿Cuál es la probabilidad que una cuenta nueva sea cobrada alguna vez?

2. ¿Cuál es la probabilidad que una cuenta atrasada un mes se vuelva finalmente incobrable?

3. Si las ventas de la empresa son 100 000 dólares en promedio mensual, ¿cuánto dinero será incobrable cada año?

Solución De la descripción anterior, recuerde que

$$Q=\begin{bmatrix}0 & .6 & 0 & 0\\ 0 & 0 & .5 & 0\\ 0 & 0 & 0 & .4\\ 0 & 0 & 0 & 0\end{bmatrix} \quad R=\begin{bmatrix}.4 & 0\\ .5 & 0\\ .6 & 0\\ .7 & .3\end{bmatrix}$$

Entonces

$$I-Q=\begin{bmatrix}1 & -.6 & 0 & 0\\ 0 & 1 & -.5 & 0\\ 0 & 0 & 1 & -.4\\ 0 & 0 & 0 & 1\end{bmatrix}$$

$$(I-Q)^{-1}=\begin{bmatrix}1 & .60 & .30 & .12\\ 0 & 1 & .50 & .20\\ 0 & 0 & 1 & .40\\ 0 & 0 & 0 & 1\end{bmatrix}$$

Para contestar las preguntas 1 a 3 necesitamos calcular

$$(I-Q)^{-1}R=\begin{bmatrix}.964 & .036\\ .940 & .060\\ .880 & .120\\ .700 & .300\end{bmatrix}$$

Entonces

1. $t_1$ = Nueva, $a_1$ = Pagada. Así, la probabilidad de que una cuenta nueva se pague finalmente es el elemento 11 de $(I-Q)^{-1}R$ =.964

2.$t_2$ = 1 mes, $a_2$ = Incobrable. Entonces, la probabilidad que una cuenta atrasada un mes se vuelva incobrable es el elemento 22 de $(I-Q)^{-1}R$ = .06.

3. De la respuesta 1, sólo el 3.6% de todas las deudas son incobrables. Como las cuentas totales del año son 1 200 000 dólares en promedio, (.036)(1 200 000) =43 200 dólares serán impagables al año.

Ejemplo Planificación del personal (continuación)

1. ¿Cuál es la duración promedio de un abogado joven recién contratado en la empresa?

2. ¿Cual es la probabilidad de que un abogado joven llegue a ser socio?

3. ¿Cuál es la duración promedio que pasa un socio en el bufete?

Solución Recordemos que en el Ejemplo

$$Q = \begin{bmatrix} .80 & .15 & 0 \\ 0 & .70 & .20 \\ 0 & 0 & .95 \end{bmatrix} \quad R = \begin{bmatrix} .05 & 0 \\ .10 & 0 \\ 0 & .05 \end{bmatrix}$$

Entonces.

$$I - Q = \begin{bmatrix} .20 & -.15 & 0 \\ 0 & .30 & -.20 \\ 0 & 0 & .05 \end{bmatrix}$$

luego

$$(I - Q)^{-1} = \begin{bmatrix} 5 & 2.5 & 10 \\ 0 & \frac{10}{3} & \frac{40}{3} \\ 0 & 0 & 20 \end{bmatrix}$$

$$(I - Q)^{-1} R = \begin{bmatrix} .50 & .50 \\ \frac{1}{3} & \frac{2}{3} \\ 0 & 1 \end{bmatrix}$$

Por lo tanto,

1. El tiempo esperado que un abogado principiante permanece en la empresa = (duración esperada del abogado principiante en la empresa como principiante) + (tiempo esperado que el abogado principiante permanece en la empresa como abogado con experiencia) + (tiempo esperado que el abogado principiante permanece en la empresa como socio). Entonces

Tiempo esperado como principiante = $(I - Q)^{-1}_{11} = 5$

Tiempo esperado como con experiencia = $(I-Q)^{-1}_{12}$ = 2.5
Tiempo esperado como socio = $(I-Q)^{-1}_{1.3}$ = 10

Por lo tanto, el tiempo total esperado que un abogado principiante permanece en la empresa es 5 + 2.5 + 10 == 17.5 años.

2. La probabilidad de que un abogado principiante recién ingresado llegue a ser socio es tan solo la probabilidad de que salga de la empresa siendo socio. Como $t_1$ = Principiante y $a_2$ = Sale siendo socio, la respuesta es el elemento 12 de $(I-Q)^{-1}R$ = .50.

3. Como t3 = Socio, buscamos el numero esperado de años que pasa en t3 dado que comenzamos en t3. Esto es justamente el elemento 33 de $(I-Q)^{-1}$ = 20 años. Es razonable, porque durante cada año hay una probabilidad en 20 que un socio deje el bufete y. por lo tanto, debe lardar un promedio de 20 años en dejar la empresa.

**PROBLEMAS**

l. El departamento de admisión del colegio estatal ha modelado la trayectoria de un estudiante en esa institución corno cadena de Markov:

| | 1er año | 2do año | 3er año | 4to año | Termina | Sale |
|---|---|---|---|---|---|---|
| 1er año | .10 | .80 | 0 | 0 | .10 | 0 |
| 2do año | 0 | .10 | .85 | 0 | .05 | 0 |
| 3er año | 0 | 0 | .15 | .80 | .05 | 0 |
| 4to año | 0 | 0 | 0 | .10 | .05 | .85 |
| Sale | 0 | 0 | 0 | 0 | 1 | 0 |
| Tea | 0 | 0 | 0 | 0 | 0 | 1 |

Se observa el estado de cada estudiante al principio de cada semestre de otoño. Por ejemplo, si un estudiante es de 3er año al principio de este semestre de otoño, habrá 80% de probabilidades de que al principio del siguiente semestre de otoño sea de cuarto año, 15% de probabilidad de que aun sea de 3er año y 5% de que salga. Suponemos que una vez que sale un estudiante ya nunca vuelve a inscribirse.

(a) Si un estudiante entra al colegio a primer año, ¿cuántos años se espera que pasen siendo estudiante?

(b) ¿Cuál es la probabilidad de que se gradúe un estudiante de nuevo ingreso?

2. El Herald Tribble obtuvo la siguiente información acerca de sus suscriptores: durante el primer año como suscriptores, el 20% cancelan sus suscripciones. De los que se han suscrito por un año, el 10%

cancelan durante el segundo año. De los que se han suscrito por más de dos años, el 4% cancelan durante cualquier año dado. En promedio, ¿cuánto tiempo se suscribe una persona al Herald Tribble ?

3. Un bosque consta de dos tipos de árboles: los que tienen de 0 a 1.50 m de alto, y los que son más altos. Cada año muere el 40% de los árboles que tienen menos de 1.50 m, el 10% se venden a 20 dólares cada uno, 30% permanecen entre 0 y 1.50 in, y el 20% crecen mas de 1.50 m. Cada año, el 50% de los árboles de más de 1.50 m se venden a 50 dólares, el 20% se venden a 30 dólares, y el 30% permanecen en el bosque.

a) ¿Cuál es !a probabilidad de que muera un árbol de 0 a 1.50 m antes de venderse?

b) Si se planta un árbol de menos de 1.50 m, ¿cuál es el ingreso esperado que se va a tener con ese árbol?

4. Las cadenas absorbentes de Markov se usan en ventas para modelar la probabilidad de que un cliente que se localiza por teléfono compre finalmente algún producto. Considere un cliente posible a quien nunca le ha llamado acerca de comprar un producto. Después de una llamada, hay una probabilidad de 60% de que tenga poco interés en el producto, de 30% que muestre un gran interés en el producto, y 10% de que sea borrado de la lista de los posibles clientes de la compañía. Se tiene un cliente que actualmente tiene poco interés en el producto- Después de otra llamada, hay 30% de probabilidades de que compre el producto, 20% de probabilidades de que sea borrado de la lista, 30% de que el cliente aún tenga poco interés y 20% de que exprese un interés alto. Para un cliente que actualmente expresa alto interés, después de otra llamada hay 50% de probabilidades de que compre el producto, 40% de probabilidades de que siga teniendo gran interés y 10% de probabilidades que tenga poco interés.
a) ¿Cuál es ]a probabilidad de que un nuevo posible cliente al final compre el producto?
b) ¿Cuál es la probabilidad de que un posible cliente con poco interés sea borrado de la lista finalmente?
c) En promedio, ¿cuántas veces habrá que llamar por teléfono a un nuevo posible cliente para que compre el producto, o para que sea borrado de la lista?

5. En el problema de la ruina del jugador (Ejemplo 1), suponga que p = .60.

a) ¿Qué probabilidad hay de que alcance a ganar 4 dólares?
b) ¿Cuál es la probabilidad de que salga sin dinero?
(c) ¿Cuál es la duración esperada del juego?

6. En el cuidado de pacientes ancianos en un hospital psiquiátrico, una meta principal es la colocación correcta de los pacientes en pensiones u hospitales para ancianos. El movimiento de pacientes entre el hospital, los hogares externos y el estado absorbente (la muerte) se puede describir mediante la siguiente cadena de Markov. La unidad de tiempo es un mes:

| | Hospital | Hogares | Muerte |
|---|---|---|---|
| Hospital | .991 | .003 | .006 |
| Hogares | .025 | .969 | .006 |
| Muerte | 0 | 0 | 1 |

Cada mes que pasa un paciente en el hospital cuesta 655 dólares al estado, y cada mes que pasa en una pensión le cuesta 226 dólares, también al estado. Para mejorar la ocurrencia de éxitos de colocación de pacientes, el estado recientemente comenzó un "programa de resocialización geriátrica" (GRP) para preparar a los pacientes a desempeñarse en las pensiones. Algunos pacientes se colocan en el GRP y a continuación pasan a pensiones. Es menos probable que estos pacientes no se puedan ajustar a sus pensiones. Otros pacientes continúan pasando en forma directa del hospital a las pensiones sin haber tomado parte en el (GRP). El estado paga 680 dólares cada mes lo que cuesta el paciente en el GRP. El movimiento de los pacientes está gobernado por la siguiente cadena de Markov:

| | GRP | Hospital | Pensiones (GRP) | Pensiones (directo) | Muerte |
|---|---|---|---|---|---|
| GRP | .854 | .028 | .112 | 0 | .006 |
| Hospital | .013 | .978 | 0 | .003 | .006 |
| Pensiones (GRP) | .025 | 0 | .969 | 0 | .006 |
| Pensiones (directo) | 0 | .025 | 0 | .969 | .006 |
| Muerte | 0 | 0 | 0 | 0 | 1 |

(a) El GRP, ¿ahora fondos al estado?

(b) Bajo el sistema anterior y bajo el GRP, calcule el número esperado de meses que pasa un paciente en el hospital.

7. Freezco, Inc.. vende refrigeradores. La fábrica otorga una garantía en todos los refrigeradores que especifica cambio gratis de cualquier unidad que se descomponga antes de tres años. Se nos da la siguiente información: (1) el 3% de todos los refrigeradores nuevos falla durante su primer ano de funcionamiento; (2) el 5% de todos los refrigeradores con 1 año de funcionamiento falla durante su segundo año de trabajo, y (3) el 7% de todos los refrigeradores con dos años de funcionamiento falla durante su tercer año. La garantía no vale para el refrigerador de repuesto.

a) Use la teoría de endonas de Markov para predecir la fracción de lodos los refrigeradores que deberá cambiar Freezco.

b) Suponga que a Freezco le cuesta 500 dólares cambiar un refrigerador y que vende 10 000 refrigeradores al año. Si la fabrica redujera el plazo de garantía a dos años, ¿cuánto dinero se ahorraría en costos de reemplazo?

8. Para una matriz Q que, represente las transiciones entre estados transitorios en una cadena absorbente de Markov, se puede demostrar que

$$(I-Q)^{-1} = I + Q \ + Q^2 + \cdots$$

a) Explique por que es posible esta expresión de $(I-Q)^{-1}$

b) Defina a $m_{ij}$ = número esperado de periodos pasados en el estado transitorio t, antes de la absorción, si se sabe que iniciamos en el estado $t_i$¿. Suponga que el periodo inicial se pasa en el estado $t_i$. Explicar por qué $m_{ij}$ = (probabilidad de que estemos al principio en el estado $t_j$} + (probabilidad que estemos en el estado $t_j$ después de la primera transición) + (probabilidad que estemos en el estado $t_j$ después de la segunda transición) + . . . + (probabilidad que estemos en el estado $t_j$ después de la n-ésima transición) + • • • .

c) Explique por qué la probabilidad de que estemos inicialmente en el estado $t_j$ = elemento ij-ésimo de la matriz identidad (s - m} x (s - m). Explique por qué la probabilidad de que estemos en el estado $t_j$ después de la ij-ésima transición = elemento ij-ésimo de $Q^n$.

d) Ahora explique por qué $m_{ij}$ = elemento ij de $(I-Q)^{-1}$

9. Defina

$b_{ij}$ = probabilidad de terminar en un estado absorbente $a_j$ dado que iniciamos en un estado transitorio $t_j$
$r_{ij}$ = ij-ésimo elemento de R.
$q_{ij}$ = ij-ésimo elemento de Q.
B = matriz (s –m) x m cuyo ij-ésimo elemento es $b_{ij}$

Suponga que iniciamos en el estado $t_i$. En nuestra primera transición, pueden suceder tres tipos de eventos:

Evento 1 Pasamos al estado absorbente $a_j$, con probabilidad $r_{ij}$.

Evento 2 Pasamos al estado absorbente que no es $a_j$, con probabilidad $\sum_{k \neq j} r_{ik}$

Evento 3 Pasamos al estado transitorio $t_k$, con probabilidad $q_{ik}$

a) Explique por qué

$$b_{ij} = r_{ij} + \sum_{k \neq j} q_{ik} b_{kj}$$

b) Ahora demuestre que $b_{ij}$ es el ij-ésimo elemento de (R +QB) y que B = R+ QB

c) Demuestre que B = $(I-Q)^{-1}$ y que $b_{ij}$ = ij-ésimo elemento de B = $(I-Q)^{-1}$.

10, General Motors tiene tres divisiones automotrices (división 3, división 2 y división 3). También tiene una división de contabilidad y una de consultora de administración. La pregunta es: ¿Que fracción del costo de las divisiones de contabilidad y de consultoría de administración se debe cargar a cada división automotriz? Suponemos que el costo total de los departamentos de contabilidad y

consultora se deben repartir entre las tres divisiones automotrices. Durante un ano determinado, el trabajo de las divisiones de contabilidad y consultoría se asigna como se ve en la siguiente tabla.

| | Contabilidad | Consultoría de Adm. | | División 2 | División 3 |
|---|---|---|---|---|---|
| Contabilidad | 10% | 30% | 20% | 20% | 20% |
| Administración | 30% | 20% | 30% | 0% | 20% |

Por ejemplo, contabilidad emplea el 10% de su tiempo en problemas generados por el departamento de contabilidad, 20% en trabajos generados por la división 3, etc. Cada año, cuesta 63 millones de dólares la operación de! departamento de contabilidad, y 210 millones de dólares la del departamento de consultoría de administración. ¿Que fracción de esos costos se debe asignar a cada división automotriz?

Imaginar 1 dólar en costos incurridos en trabajos de contabilidad. Hay una probabilidad .20 de que estos costos se asignen a cada división automotriz, probabilidad .30 de que se asigne a consultoría y probabilidad .10 que se asigne a contabilidad. Si el dólar se asigna a una división automotriz, sabemos a qué división se debe cargar ese dólar. Por ejemplo, si el dólar se carga a consultaría, repetimos el proceso hasta que, por último, el dólar se cargue a una división automotriz.

Use el conocimiento de cadenas de Markov para establecer cómo asignar los costos de funcionamiento de los departamentos de contabilidad y asesoría entre las tres divisiones automotrices.

# 2.TEORÍA DE COLAS

Todos nosotros hemos esperado mucho tiempo en líneas o colas. En este capitulo diseñaremos modelos matemáticos de colas de espera. En la sección 2.3 comenzaremos describiendo parte de la terminología que se usa con frecuencia para describir las colas. En la Sección 2.2 veremos algunas distribuciones (exponencial y de Erlang) que se necesitan para describir los modelos de colas. En la Sección 2.3 introduciremos la idea de un proceso de nacimiento y muerte básico para muchos modelos de colas donde interviene la distribución exponencial. En el resto del capítulo examinaremos algunos modelos de sistemas de colas que se pueden usar para responder preguntas como las siguientes:

1. ¿Que fracción del tiempo está ocioso cada servidor?

2. ¿Cuál es el número esperado de clientes presentes en la cola?

3. ¿Cuál es el tiempo esperado que un cliente pasa en la cola?

4. ¿Cuál es la distribución de probabilidad del número de clientes presentes en una cola?

5. ¿Cuál es la distribución de probabilidad del tiempo de espera de un cliente?

6. Si un gerente de banco desea asegurar que sólo el 17% de los clientes tenga que esperar más de 5 minutos su turno ¿cuántas ventanillas debe emplear?

# 2.1 TERMINOLOGÍA PARA LA TEORÍA DE COLAS

Para describir una cola se deben especificar un proceso de entrada y uno de salida. En la Tabla 1 se presentan algunos ejemplos de procesos de entrada y salida.

| CASO | PROCESO DE ENTRADA | PROCESO DE SALIDA |
|---|---|---|
| Banco | Los clientes llegan al banco | Los cajeros despachan a los clientes |
| Pizzería | Se reciben pedidos de pizzas | La pizzería manda una motocicleta para entregar pizzas |
| Banco de sangre en hospital | Llegan bolsas de sangre | Los pacientes usan las bolsas de sangre |
| Astillero naval | Se descomponen los barcos en el mar y se mandan al astillero a reparación | Se reparan los barcos y regresan al mar |

**EL PROCESO DE LLEGADA**

El proceso de entrada se conoce, por lo general, por proceso de llegada. La llegada son los clientes. En todos los modelos que estudiamos, suponemos que solo

hay una llegada en un instante dado. En el caso de un restaurante, es una suposición muy poco real. Si sucede que hay más de una llegada en un instante dado, decimos que se permiten llegadas en masa.

En general, suponemos que el proceso de llegada no es afectado por el numero de clientes presentes en el sistema. En el contexto de un banco, esto significaría que el proceso que gobierna las llegadas no cambia, haya 5 o 500 personas en la cola.

Hay dos casos comunes en los que el proceso de llegada puede depender del número presente de clientes. El primero se da cuando las llegadas se toman de una población pequeña. Suponga que solo hay cuatro barcos en un astillero naval. Si los cuatro están en reparación, entonces ningún barco se puede descomponer en el futuro cercano. Por otro lado, si los cuatro barcos están en el mar, en el futuro cercano hay una probabilidad relativamente alta de que suceda una descompostura.

Los modelos en los que las llegadas se toman de una población pequeña se llaman modelos de origen finito. Otro caso en el que el proceso de llegada depende del número de clientes presentes, se tiene cuando la rapidez a la que llegan los clientes a la instalación disminuye cuando está demasiado concurrida. Por ejemplo, si el lector ve que el estacionamiento del banco está lleno, podría pasar e ir otro día al banco. Si un cliente llega, pero no puede entrar al sistema, decimos que el cliente ha declinado. El fenómeno de declinación (denegar o no aprovechar su turno) fue explicado por Yogi Berra cuando dijo: "Nadie va ya al restaurante; está demasiado concurrido."

Si al proceso de llegadas no lo afecta el numero de clientes presentes, lo expresamos normalmente especificando una distribución de probabilidad que gobierne el tiempo entre llegadas sucesivas.

**PROCESO DE SALIDA O DE SERVICIO**

Para describir el proceso de salida, que con frecuencia se llama proceso de servicio, de un sistema de cola, en general especificamos una distribución de probabilidad; la distribución del tiempo de servicio, que gobierna el tiempo de servicio a un cliente. En la mayor parle de los casos suponemos que la distribución de tiempo de servicio es independiente del número de clientes presentes. Esto significa, por ejemplo, que el servidor no trabaja más rápido cuando hay más clientes.

En este capítulo estudiaremos dos acomodos de servidores: servidores en paralelo y servidores en serie. Los servidores están en paralelo si todos ellos dan el mismo tipo de servicio y un cliente sólo necesita pasar por un servidor para completar su servicio. Por ejemplo, los cajeros de un banco están organizados generalmente en paralelo; cualquier cliente sólo necesita ser atendido por un cajero, y cualquier cajero puede llevar a cabo el servicio deseado. Los servidores están en serie si un cliente debe pasar por varios de ellos antes de completar el servicio.

Un ejemplo de un sistema de cola en serie es una línea de ensamble.

**DISCIPLINA DE LA COLA**

Para explicar por completo un sistema de cola, también debemos explicar la disciplina de la cola y la manera en la que los clientes se unen a ella.

La disciplina de la cola es el método que se usa para determinar el orden en el que se sirve a los clientes. La disciplina más común es la disciplina FIFO (el primero en llegar primero en ser servido), en la que los clientes son servidos en el orden de su llegada. Bajo la disciplina LIFO (último en llegar, primero en ser servido), las llegadas mas recientes son las primeras en recibir el servicio. Si pensamos que la salida de un elevador sea un servicio, entonces un elevador muy lleno ilustra una disciplina LIFO. A veces, el orden en el que llegan los clientes no tiene efecto alguno sobre el orden en que se les sirve. Este sería el caso si el siguiente cliente en llegar se selecciona al azar de entre los que están esperando para ser atendidos. A este caso se le llama disciplina SEOA (servicio en orden aleatorio). Cuando al llamar a una aerolínea se les pide que esperemos, la suerte de la selección determina con frecuencia al siguiente interlocutor que es atendido por un operador.

Por último, veremos las disciplinas de prioridad en espera. Esta disciplina clasifica cada llegada en alguna de las categorías que tiene. Luego se le da un nivel de prioridad, a cada categoría y dentro de cada nivel de prioridad los clientes entran al servicio según una disciplina LIFO. Las disciplinas de prioridad se usan con frecuencia en las salas de urgencia para determinar el orden en el que los clientes deben recibir tratamiento, y en las instalaciones de copiado y de tiempo compartido de cómputo, cuando la prioridad se da en general, a trabajos que reciben poco tiempo.

**MÉTODO UTILIZADO POR LAS LLEGADAS PARA UNIRSE A LA COLA**

Otro factor que tiene un importante efecto sobre el comportamiento de un sistema de cola es el método que usen los clientes para determinar a cuál cola unirse. Por ejemplo, en algunos bancos los clientes deben hacer una sola cola, pero en otros, pueden escoger la cola donde formarse. Cuando hay varias colas, con frecuencia los clientes se forman en la mas corla. Desafortunadamente en muchos casos, como en un supermercado, por ejemplo, es difícil definir la cola más corta. Si hay varias colas en una instalación, es importante conocer si se permite a los clientes cambiar de cola o no. En la mayor parte de los sistemas con colas múltiples se permite el cambio, pero no se recomienda el cambio en una casilla para peaje.

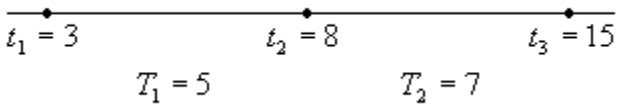
# 2-2 MODELADO DE LOS PROCESOS DE LLEGADA Y DE SERVICIO

**MODELADO DEL PROCESO DE LLEGADA**

Como se mencionó antes, suponemos que puede suceder una llegada cuando mucho en un instante dado de tiempo, Definimos a $t_i$ como el tiempo en el cual llega el i-ésimo cliente. Para ilustrarlo, veamos la Fig. 1. Para $i \geq 1$, definimos $T_i = t_{i+1} - t_i$; como el i-ésimo tiempo entre las llegadas. Así, en esa figura, $T_1 = 8 - 3 = 5$, y $T_2 = 15 - 8 = 7$. Al modelar el proceso de llegada, suponemos que las $T_i$ son variables independientes, aleatorias y continuas descritas por la variable aleatoria A. La hipótesis de independencia quiere decir, por ejemplo, que el valor de $T_2$ no tiene efecto sobre el valor de $T_3$,$T_4$; o cualquier otra $T_i$, posterior. La hipótesis de que cada T, es continua en general es buena aproximación a la realidad. Después de todo, un tiempo entre llegadas no necesita ser exactamente de 1 o 2 minutos; podría ser igualmente de, por ejemplo 1.55892 minutos. La hipótesis de que cada tiempo entre llegadas está gobernado por la misma variable significa que la distribución de llegadas es independiente de la hora del día o del día de la semana. Es la hipótesis de tiempos estables entre llegadas. A causa de fenómenos

como horas de alto tránsito, con frecuencia la hipótesis de tiempos estables entre llegadas no es real, pero con frecuencia podemos aproximar el caso real descomponiendo las horas del día en segmentos. Por ejemplo, si estuviéramos modelando el flujo de tránsito, podríamos descomponer el día hasta en tres segmentos: uno de transito elevado por la mañana, un segmento de medio día y uno de prisas por la tarde. Durante cada uno de esos segmentos los tiempos entre llegadas pueden ser estables.

Figura 1 Definición de tiempos entre llegadas




Suponga que A tiene una función de densidad a(t}. Para $\Delta$t pequeña, P( t $\leq$ A $\leq$ t + $\Delta$t) es aproximadamente $\Delta$t a(t). Naturalmente que es imposible un tiempo negativo entre llegadas. Esto nos permite escribir

(1) $P(A \leq c) = \int_0^c a(t)dt$ y $P(A > c) = \int_c^\infty a(t)dt$

Definimos $\frac{1}{\lambda}$ como el tiempo promedio entre llegadas. Sin pérdida de generalidad, suponemos que el tiempo se mide en horas. Entonces $\frac{1}{\lambda}$ tendrá unidades de horas por llegada. Podemos calcular $\frac{1}{\lambda}$ a partir de a(t) empleando la siguiente ecuación:

(2) $\frac{1}{\lambda} = \int_0^\infty a(t)dt$

A $\lambda$ se le llama rapidez de llegadas y es el tiempo promedio entre llegadas. Sus unidades son llegadas por hora.

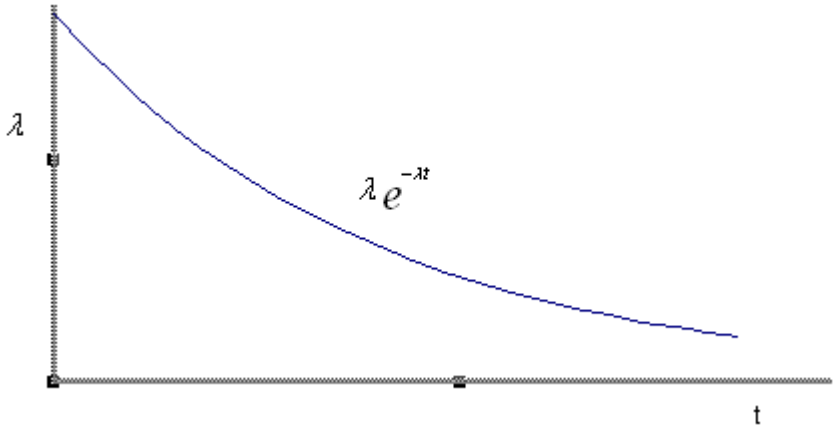
En la mayor parte de las aplicaciones de colas, un asunto impórtame es como escoger A para que refleje la realidad y que, al mismo tiempo, se pueda manejar matemáticamente. La selección mas común de A es la distribución exponencial Una distribución exponencial con parámetro $\lambda$ tiene una densidad a(t} = $\lambda e^{-\lambda t}$ la Fig. 2 se muestra la función de densidad para una distribución exponencial. Vemos que a{t) disminuye en forma muy rápida para t pequeña. Esto indica que son poco probables tiempos muy grandes entre llegadas. Con la ecuación(2) e integrando por partes podemos demostrar que el tiempo medio o promedio entre llegadas E(A) está dado por

(3) E(A) = $\frac{1}{\lambda}$

De acuerdo con el hecho que var A = E($A^2$) - E(A)$^2$, podemos demostrar que

(4) var A = $\left(\frac{1}{\lambda}\right)^2$

Figura 2 Función de densidad para la distribución exponencial




Propiedad de Amnesia de la distribución exponencial. La razón que se use con frecuencia la distribución exponencial para modelar los tiempos entre llegadas se encuentra en el siguiente lema:

LEMA 1. Si A tiene distribución exponencial, entonces para todo valor no negativo de t y h

(5) $$P(A > t + h / A \geq t) = P(A > h)$$

Demostración: Notaremos primero que según la Ecuación (1),

(6) $$P(A > h) = \int_h^\infty a(t)dt = \int_h^\infty \lambda e^{-\lambda t} dt = e^{-\lambda h}$$

Luego

$$P(A > t + h / A \geq t) = \frac{P(A > t + h \cap A \geq t)}{P(A \geq t)} = \frac{P(A > t + h)}{P(A \geq t)}$$ utilizando (6) =

$$\frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = P(A > h)$$

Se puede demostrar que no hay otra función de densidad que satisfaga la Ecuación (5). Por razones que se hacen evidentes, se dice que una densidad que satisfaga la Ecuación (5) tiene la propiedad de amnesia, o de no memoria. Suponga que sabemos que no ha habido llegadas durante las últimas t horas, que equivale a que nos digan que A ≥ t y que nos pregunten cuál es la probabilidad que no haya llegadas durante las siguientes h horas, es decir, A > t + h. Entonces, la Ecuación (5) quiere decir que esta probabilidad no depende del valor de t, y que para

todos los valores de t esta probabilidad es igual a P(A > h). En resumen, si conocemos que han pasado al menos t unidades de tiempo desde la última llegada, entonces la distribución del tiempo que queda hasta la siguiente llegada, h, no depende de t. Por ejemplo. Si h = 4, entonces la Ecuación (5) produce, para t = 5 , t = 3,t = 2 y t = 0,

$$P(A > 9 / A \geq 5) = P(A > 7 / A \geq 3) = P(A > 6 / A \geq 2) = P(A > 4 / A \geq 0) = e^{-4\lambda}$$

La propiedad de amnesia de la distribución exponencial es importante porque si deseamos conocer la distribución de probabilidad del tiempo para la siguiente llegada, no importa cuanto tiempo haya pasado desde la ultima llegada. Para decirlo en términos concretos, suponga que los tiempos entre llegadas se distribuyen exponencialmente con λ = 6. Entonces la propiedad de amnesia significa que no importa cuánto tiempo haya pasado desde la última llegada, la distribución de probabilidades que rige el tiempo para la siguiente llegada tiene la función de densidad $\lambda e^{-\lambda t}$. Esto significa que para predecir los comportamientos de las llegadas futuras no necesitamos mantener registro de cuanto ha pasado desde la última llegada. Esta observación puede simplificar mucho el análisis de un sistema de colas.

Para visualizar que conocer el tiempo desde la última llegada no afecta la distribución del tiempo para la siguiente llegada en la mayor parte de los casos, suponga que A es discreta con P(A = 5) = P(A = 100) = xi Si sabemos que no ha habido llegadas durante las últimas 6 unidades de tiempo, sabemos con certeza que pasaran 100 – 6 = 94 unidades de tiempo para la siguiente llegada. Por otro lado, si nos dicen que no ha habido llegada durante la última unidad de tiempo, entonces hay cierta probabilidad de que el tiempo para la siguiente llegada sea 5-1=4 unidades y cierta probabilidad que sea 100-1 =99 unidades. Por lo tanto, en este caso, la distribución del siguiente tiempo entre llegadas no se puede predecir conociendo el tiempo que ha pasado desde la última llegada.

**RELACIÓN ENTRE LA DISIRIBUCIÓN DE POISSON Y LA DISTRIBUCIÓN EXPONENCIAL.**

Si los tiempos entre llegadas son exponenciales, la distribución de probabilidad del número de llegadas que se tienen en cualquier intervalo de tiempo t está dada por el siguiente teorema importante:

TEOREMA 1 Los tiempos entre llegadas son exponenciales con parámetro λ si y solo si el número de llegadas que suceden en un intervalo t sigue una distribución de Poisson con parámetro λt.

Una variable aleatoria discreta N tiene una distribución de Poisson con parámetro λ si, para n = 0. 1, 2, . . . ,

(7) $P(N = n) = \dfrac{e^{-\lambda}\lambda^n}{n!}$ (n = 0.1.2....)

Si N es una variable aleatoria de Poisson, se sabe que E(N) = var N = λ. Si hacemos que N, sea el número de llegadas que suceden durante cualquier intervalo de tiempo de longitud t, el Teorema 1 establece que

$P(N_t = n) = \dfrac{e^{-\lambda t}(\lambda t)^n}{n!}$ (n = 0.1.2....)

Como N, es de Poisson con parámetro $\lambda t$, E(N) = var N = $\lambda t$. Un promedio de $\lambda t$ llegadas se suceden durante un intervalo de tiempo de longitud t y, entonces se

puede pensar que λ es el número promedio de llegadas por unidad de tiempo, o rapidez de llegadas.

¿Qué hipótesis se necesitan para que los tiempos entre llegadas sean exponenciales?. El Teorema 2, mas adelante, nos da una respuesta parcial. Veamos las dos hipótesis siguientes:

1. Las llegadas definidas en intervalos de tiempo que no se traslapan son independientes (por ejemplo, el número de llegadas que se tiene entre los tiempos 1 y 10 no nos da información alguna acerca del número de llegadas entre los tiempos 30 y 50).

2. Para $\Delta t$ pequeño, y cualquier valor de t, la probabilidad de que se tenga una llegada entre los tiempos t y t + $\Delta t$ es $\lambda \Delta t$ + o($\Delta t$) donde o($\Delta t$) es cualquier cantidad que satisfaga

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

También, la probabilidad de que no haya llegada durante el intervalo t y t + $\Delta t$ es 1 $\lambda \Delta t$ + o($\Delta t$), y la probabilidad que halla i más de una llegada en ese intervalo es o($\Delta t$).

TEOREMA 2: Si son validas las hipótesis 1 y 2, entonces N, sigue una distribución de Poisson con parámetro $\lambda t$, y los tiempos entre llegadas son exponenciales con parámetro λ Esto es, a(t} = $\lambda e^{-\lambda t}$

En esencia, el Teorema 2 establece que si la rapidez de llegadas es estable, si no pueden tenerse llegadas en masa y si las llegadas del pasado no afectan las llegadas del futuro, entonces los tiempos entre llegadas seguirán una distribución exponencial con parámetro λ y el número de llegadas en cualquier intervalo de longitud t es de Poisson con parámetro $\lambda t$. Las hipótesis del Teorema 2 pueden parecer demasiado restrictivas, pero con frecuencia los tiempos entre llegadas son exponenciales aun cuando no se satisfagan las hipótesis del Teorema 2.

En la Sección 2.12 estudiamos cómo usar los datos para probar si la hipótesis de tiempos exponenciales entre llegadas es razonable. Sucede que en muchas aplicaciones, la hipótesis de tiempos exponenciales entre llegadas es una muy buena aproximación del caso real.

El Ejemplo ilustra la relación entre Lis distribuciones exponencial y de Poisson.

EJEMPLO 1 El numero de tarros de cerveza pedidos en el Dick's Pub sigue una distribución de Poisson con promedio de 30 cervezas por hora.
1. Calcule la probabilidad de que se pidan exactamente 60 cervezas entre las 10 p.m. y las 12 de la noche.

2. Determine el promedio y la desviación estándar del número de cervezas pedidas entre las 9 p.m. y la 1 a.m.

3. Calcule la probabilidad de que el tiempo entre dos pedidos consecutivos sea entre 1 y 3 minutos.

1. El número de cervezas pedido entre las 10 p.m. y las 12 de la noche sigue una distribución de Poisson con parámetro 2(30) = 60. De la Ecuación (7), la probabilidad de que se pidan 60 cervezas entre las 10 p.m. y la medianoche es

$$\frac{e^{-60}60^{60}}{60!}$$

2. $\lambda$ = 30 cervezas por hora; t = 4 horas. Entonces, el número promedio de cervezas pedidas entre las 9 p.m. y la 1 am es 4(30) = 120 cervezas. La desviación estándar del número de cervezas pedido entre las 9 p.m. y la 1 a.m. es $(120)^{1/2}$ = 10.95.

3. Sen X el tiempo en minutos entre pedidos sucesivos de cerveza. El tiempo promedio de pedidos por minuto es exponencial con parámetro, o razón, 30/60 = .5 cervezas por minuto. Entonces la función de densidad de probabilidad del tiempo ente pedidos de cerveza es $.5e^{-.5t}$. Entonces

$$P(1 \le X \le 3) = \int_1^3 a(t)dt = \int_1^3 .5e^{-.5t}\,dt = e^{-.5} - e^{-1.5} = .38$$

**DISTRIBUCIÓN DE ERLANG**

Si los tiempos entre llegadas parecen no ser exponenciales, se modelan con frecuencia con una distribución de Erlang. Una distribución de Erlang es una variable aleatoria continua, T, cuya función de densidad f(t) está especificada por dos parámetros: un parámetro de rapidez R y un parámetro de forma k (k - debe ser entero positivo). Dados valores de R y k, la densidad de Erlang tiene la siguiente función de densidad de probabilidad:

$$(8)\quad f(t) = \frac{R(Rt)^{k-1}e^{-Rt}}{(k-1)!}\left(t \ge 0\right)$$

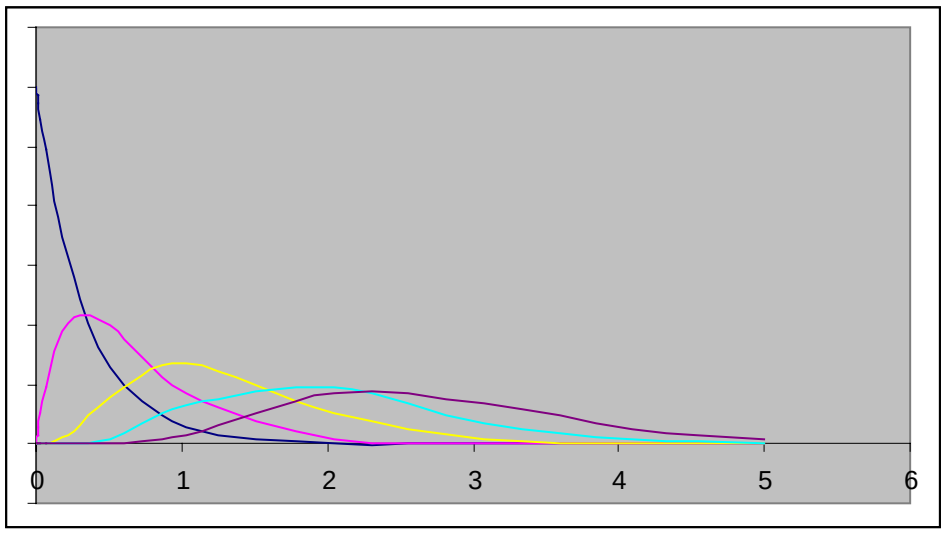
Mediante integración por parles, podemos demostrar que si T es una distribución de Erlang con parámetro R de rapidez y k de forma, entonces

(9) E(T) = k/R y Var T = $k/R^2$

Para ver cómo al variar el parámetro de forma se cambia la forma de la distribución de Erlang, consideremos un valor dado de $\lambda$, una familia de distribuciones de Erlang; con parámetro de rápido $\lambda$k y parámetro de forma k. Según la Ecuación (9), cada una de esas distribuciones de Erlang tiene un promedio de $1/\lambda$. Cuando varía k, la distribución de Erlang toma muchas formas. Por ejemplo, en la Fig. 3 se ilustra, para un valor dado de $\lambda$, las funciones de densidad para las distribuciones de Erlang que tienen parámetros de forma 1, 2, 4, 6 y 20. Para k = 1, la densidad de Erlang parece semejante a una distribución exponencial; de hecho, si hacemos k = 1 en la Ecuación (8) vemos que para este caso la distribución de Erlang es exponencial con parámetro R. Cuando aumenta k, la distribución de Erlang se comporta más y mas como distribución normal. Para valores extremadamente grandes de la distribución de Erlang tiende a una variable

aleatoria con variancia cero (es decir, tiempo constante entre llegadas). Así, al variar k podemos aproximarnos a distribuciones tanto asimétricas como simétricas.

Figura 3. Distribución Erlang para diferentes valores de k y con R=3




Se puede demostrar que una distribución de Erlang con parámetro de forma k y de rapidez $\lambda k$ tiene la misma distribución que la variable aleatoria $A_1 + A_2 + \bullet\bullet\bullet + A_k$, en la cual cada $A_i$ es una variable aleatoria exponencial con parámetro $\lambda k$ y las $A_i$ son variables aleatorias independientes.

Si modelamos los tiempos entre llegadas como una distribución de Erlang con parámetro de forma k, en realidad estamos diciendo que el proceso entre llegadas es equivalente a que un cliente pase a través de k fases, cada una de las cuales con propiedad de amnesia, antes de llegar. Por esta razón, al parámetro de forma se le llama con frecuencia numero de fases de la distribución de Erlang.

**MODELADO DEL PROCESO DE SERVICIO**

A continuación dirigiremos nuestra atención al modelado del proceso de servicio. Suponemos que los tiempos de servicio de distintos clientes son variables aleatorias independientes y que el tiempo de servicio de cada cliente está regida por la variable aleatoria S que tiene una función de densidad s(t). definimos $\frac{1}{\mu}$ como el tiempo promedio de servicio a un cliente. Naturalmente,

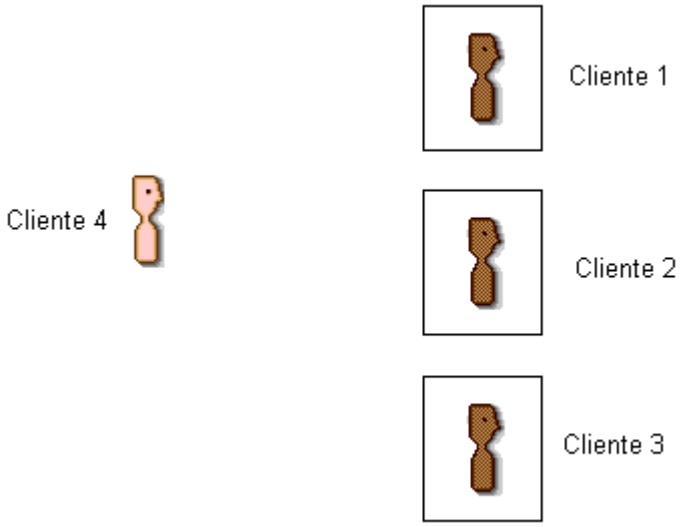
$$\frac{1}{\mu} = \int_0^\infty ts(t)dt$$

La variable $\frac{1}{\mu}$ tendrá como unidades horas por cliente; por lo tanto, $\mu$ tiene como unidades clientes por hora. Por esta razón se le llama a $\mu$ la rapidez de servicio. Por ejemplo, $\mu = 5$ quiere decir que si siempre hubiera clientes, quien atiende

podría despachar a un promedio de 5 clientes por hora, y el tiempo promedio de servicio para cada cliente sería $\frac{1}{5}$ de hora. Como en el caso de los tiempos entre llegadas, esperamos que los tiempos de servicio se pueden modelar con exactitud como variables aleatorias exponenciales. Si podemos modelar al tiempo de servicio de un cliente, como una variable aleatoria exponencial, podremos determinar la distribución del tiempo de servicio sobrante de un cliente sin tener que registrar cuánto ha durado el servicio al cliente. Nótese también que si los tiempos de servicio siguen una densidad exponencial $s(t) = \mu e^{-\mu t}$ entonces el tiempo promedio de servicio a un cliente será .

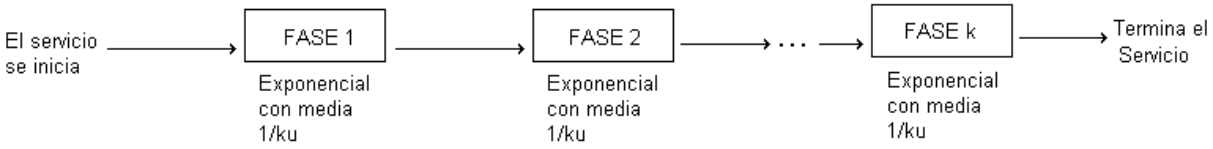
Como ejemplo de cómo la hipótesis de tiempos de servicio exponenciales puede simplificar los cálculos, veamos un sistema de tres servidores en el que cada tiempo de servicio de cliente está gobernado por una distribución exponencial $s(t) = \mu e^{-\mu t}$ Suponga que los tres servidores están ocupados y que hay un cliente esperando (Fig. 4). ¿Cual es la probabilidad que el cliente que espera sea el último de los cuatro clientes en terminar sus tramites? En la Fig. 4 está claro que sucederá lo siguiente: uno de los clientes 1 a 3, por ejemplo el cliente 3, será el primero en terminar sus trámites. El cliente 4 llegará a la ventanilla. Según la propiedad de amnesia, el tiempo de servicio del cliente 4 tiene la misma distribución que los tiempos de servicio de los clientes 1 y 2 restantes. Así, por simetría, los clientes 4, 1 y 2 tendrán la misma probabilidad de ser el último en salir. Esto significa que el cliente 4 tiene una probabilidad 1/3 de ser el último en salir. Si no hubiera propiedad de amnesia este problema sería difícil de resolver, porque sería muy difícil determinar la distribución de probabilidades del tiempo restante de servicio páralos clientes 1 y 2, después de haber salido el cliente 3.

Ejemplo de la utilidad de la distribución exponencial




Desafortunadamente, los tiempos reales de servicio pueden no ser consistentes con la propiedad de amnesia. Por esta razón, con frecuencia suponemos que s(t) es una distribución de Erlang con parámetro de forma k y parámetro de rapidez kμ. Según

la Ecuación (9), esto produce un tiempo promedio de servicio igual a $\frac{1}{\mu}$. Modelar los tiempos de servicio como distribución de Erlang con parámetro de forma k, también significa que el tiempo de servicio a un cliente puede consistir del paso través de k fases del servicio, en las que el tiempo para completar cada una tiene la propiedad de amnesia y el tiempo para completar cada fase tiene un promedio igual a $\frac{1}{\mu k}$, (véase Fig. 5). En muchos casos, la distribución de Erlang se puede ajustar estrechamente a los tiempos de servicio observados (véase Sección 2.13).
Representación del tiempo de servicios de Erlang

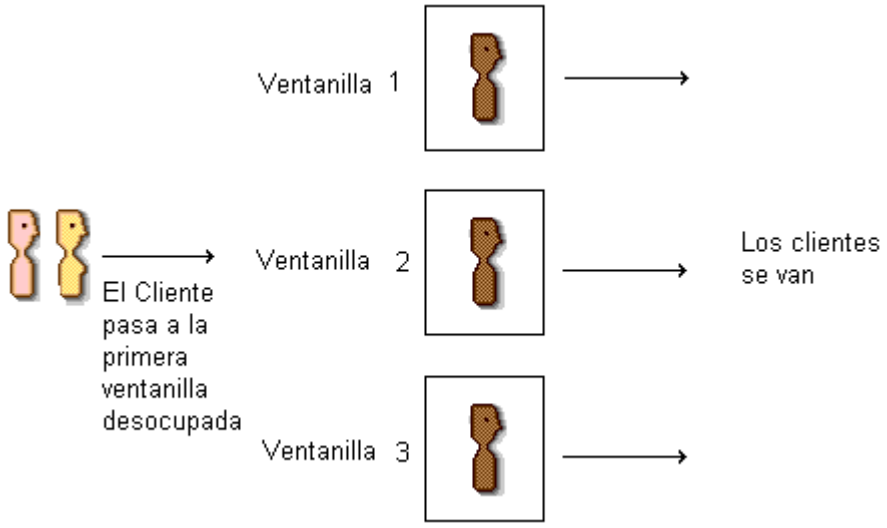



En algunos casos, los tiempos entre llegadas o los de servicio se pueden modelar como si tuvieran variancia cero; en estos casos, se considera que los tiempos entre llegadas o de servicio son deterministas. Por ejemplo, si los tiempos entre llegadas son deterministas, entonces cada tiempo entre llegadas será exactamente $\frac{1}{\lambda}$ y si los tiempos de servicio son deterministas, el tiempo de servicio de cada cliente será exactamente $\frac{1}{\mu}$

**NOTACIÓN DE KENDALL-LEE PARA SISTEMAS DE COLAS**

Hemos elaborado la terminología suficiente para describir en forma normal muchos sistemas de colas. La notación que describimos en esta sección se usa para representar un sistema de colas en el que las llegadas esperan en una sola cola hasta que uno de los s servidores idénticos en paralelo queda libre. Entonces el primer cliente de la cola entra al servicio y así sucesivamente (véase Fig. 6). Si, por ejemplo, el cliente en la ventanilla 3 es el próximo en completar el servicio, entonces, suponiendo disciplina FIFO (primero en llegar, primero en ser servido), el primero en la cola pasaría a la ventanilla 3. El siguiente cliente en la cola entraría a una ventanilla después de completar el siguiente servicio, y así sucesivamente.

Sistema de colas única con servidores en paralelo



Para representar un sistema de colas, Kendall (1951) inventó la siguiente notación. Cada sistema de colas se representa con seis características:

1/2/3/4/5/6

La primera característica especifica la naturaleza del proceso de llegada. Se usan las siguientes abreviaturas normales:

M = Los tiempos entre llegadas son independientes, distribuidos idénticamente (iid), y las variables aleatorias tienen distribución exponencial
D = Los tiempos entre llegadas son iid y deterministas.
$E_k$ = Los tiempos entre llegadas son iid con distribución de Erlang con parámetro de forma k.
GI = Los tiempos entre llegadas son iid y están gobernados por alguna distribución general.

La segunda característica especifica la naturaleza de los tiempos de servicio:

M = Los tiempos de servicio son iid y tienen distribución exponencial.
D = Los tiempos de servicio son iid y deterministas.
$E_k$ = Los tiempos de servicio son iid y con distribución de Erlang con parámetro de forma k.
G := Los tiempos de servicio son iid y siguen alguna distribución general.

La tercera característica es el numero de servidores en paralelo. La cuarta característica describe la disciplina de la cola:

FIFO = Primero en llegar, primero en ser-servido
LIFO = Último en llegar, primero en ser servido
SEOA = Servicio cu orden aleatorio
DG = Disciplina general en la cola.

La quinta característica especifica el número máximo permisible de clientes en el sistema, incluyendo los que esperan y los que están en ventanilla. La sexta característica da el tamaño de la población de la cual se toman los clientes. A menos que el número de clientes potenciales sea del mismo orden de magnitud que el número de servidores, se considera que es infinito el tamaño de la población. En muchos modelos importantes, las características 4/5/6 son DG/∞/∞. Si este es el caso. entonces con frecuencia se omite 4/5/6.

Para ilustrar esta notación, $M/E_2/8/FIFO/10/\infty$ representaría una clínica con 8 doctores, tiempos exponenciales entre llegadas, tiempos de servicio de dos fases de Erlang, disciplina de cola FIFO y una capacidad total de 10 pacientes.

**LA PARADOJA DELTIEMPO DE ESPERA**

Terminaremos esta sección con un estudio breve de una paradoja interesante llamada paradoja del tiempo de espera.

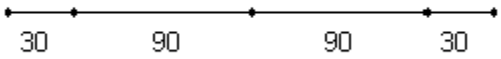
Suponga que el tiempo entre la llegada de autobuses en una terminal se distribuye en forma exponencial con promedio de 60 minutos. Si llegamos a terminal a un instante escogido al azar, ¿cual fue el tiempo promedio que tendremos que esperar para que llegue un autobús?

La propiedad de amnesia de una distribución exponencial significa que independientemente de cuánto ha transcurrido desde la última llegada de autobús,

tendríamos que esperar un promedio de 60 minutos hasta que llegara el siguiente autobús. Esta respuesta es realmente correcta, pero el siguiente argumento la parece contradecir. En promedio, alguien que llegue a un tiempo aleatorio debería llegar en la mitad de un intervalo característico entre llegadas de autobuses consecutivos. Si llegamos a la mitad de un intervalo característico, y el tiempo promedio entre las llegadas de autobuses es 60 minutos, entonces deberíamos tener que esperar, en promedio, (½)60 = 30 minutos al siguiente autobús. ¿Por que es incorrecto este argumento? Simplemente porque el intervalo característico entre autobuses es más largo que 60 minutos. La razón de esta anomalía es que es más probable que lleguemos durante un intervalo más largo que durante uno corto.

Simplifiquemos el caso suponiendo que la mitad de los autobuses llegan cada 30 minutos y la otra mitad llegan cada 9O minutos. Uno podría pensar que ya que el tiempo promedio entre autobuses es 60 minutos, el tiempo promedio de espera sería (½)60 = 30 minutos, pero esto no es cierto. Veamos una secuencia representativa de tiempos entre llegadas de autobuses (véase Fig. 7). La mitad de los tiempos entre llegadas son 30 minutos y la mitad 90 minutos. Es evidente que hay una probabilidad de $\frac{90}{30+90} = \frac{3}{4}$ de llegar durante un tiempo entre llegadas de 90 minutos y una probabilidad $\frac{30}{30+90} = \frac{1}{4}$ de llegar durante un intervalo entre llegadas de 30 minutos. Asé, el tiempo entre llegadas promedio en el que llega un cliente es ¾ (90) + ¼ (30) = 75 minutos. Como llegamos, en promedio, a la mitad de un tiempo entre llegadas, nuestra espera promedio será de ¾ ¼ 90 + ¼ ¾ 30 = 37.5 minutos, que es mas que los 30 minutos

La paradoja del tiempo de espera




Regresando al caso en el que los tiempos entre llegadas son exponenciales con promedio de 60 minutos, el tamaño promedio de un tiempo representativo entre llegadas es de 120 minutos. Entonces, el tiempo promedio que debemos esperar un autobús es ( ½ )(120) = 60 minutos. Nótese que si los autobuses llegaran siempre a intervalos de 60 minutos, entonces el tiempo promedio que una persona debería esperar sería ( ½ )(60) = 30 minutos. En general, se puede demostrar que si A es la variable aleatoria del tiempo entre autobuses, entonces el tiempo promedio para la llegada del siguiente autobús (según lo ve un cliente cuya llegada tiene igual probabilidad de entrar en cualquier momento) está dado por

$$\tfrac{1}{2}\left(E(A) + \frac{\text{var } A}{E(A)}\right)$$

Para el ejemplo de los autobuses, $\lambda = \frac{1}{60}$ entonces las Ecuaciones (3) y (4) muestran que E(A) = 60 minutos y var A = 3,600 minutos. Sustituyendo en esta fórmula obtenemos

Tiempo estimado de espera 60 minutos

**PROBLEMAS**

1. Supónganlos que llego a un sistema de cola M/M/7/FIFO/8/∞ cuando todas las ventanillas están ocupadas. ¿Cuál es la probabilidad do que termine mi trámite antes de por lo menos uno de los 7 clientes a los que están atendiendo?

2. El tiempo entre autobuses sigue la función de masa que se ve en la Tabla 2. ¿Cuál es el lapso promedio que debe uno esperar un autobús?

3. Hay cuatro grupos de tercer año en una escuela primaria. El número de alumnos en cada grupo es el siguiente: grupo 1, 20 alumnos; grupo 2, 25 alumnos; grupo 3, 35 alumnos; grupo 4, 40 alumnos. ¿Cuál es el tamaño promedio de un grupo de tercer año? Suponga que el inspector escoge al azar cualquier alumno de tercer año de esa escuela. En promedio, ¿cuántos estudiantes habrá en esa clase?

4. El tiempo entre llegadas de autobuses sigue una distribución exponencial con promedio de 60 minutos.

Tabla 2

| Tiempo entre autobuses | Probabilidad |
|---|---|
| 30 minutos | 1/4 |
| 1 hora | 1/4 |
| 2 horas | 1/2 |

(a) ¿Cuál es la probabilidad de que lleguen exactamente cuatro autobuses durante las siguientes 2 horas?

(b) ¿Cuál es la probabilidad de que por lo menos dos autobuses lleguen durante las siguientes 2 horas?

(c) ¿Cuál es la probabilidad de que no lleguen autobuses durante las próximas 3 horas?

(d) Acaba de llegar un autobús. ¿Cuál es la probabilidad de que pasen entre 30 y 90 minutos para que llegue el siguiente?

# 2.3 PROCESOS DE NACIMIENTO Y MUERTE

En esta sección analizaremos la importante idea de un proceso de nacimiento y muerte. En lo que sigue utilizaremos procesos de nacimiento y muerte para contestar preguntas acerca de diversos tipos de sistemas de colas.

Se define el número de personas presentes en cualquier sistema de cola en el tiempo t como estado del sistema en el tiempo t. Para t = O, el estado del sistema será igual al número de personas presentes al inicio en el sistema. Es de gran interés para nosotros la cantidad $P_{ij}(t)$ que se define como la probabilidad que haya j personas en el sistema de cola en el tiempo t, dado que en el tiempo O había i personas. Nótese que $P_{ij}(n)$ es análoga a la probabilidad de transición de n pasos $P_{ij}(n)$ (la probabilidad de que después de n transiciones una cadena de

Markov este en el estado j, dado que la cadena comenzó en el estado i) que se estudió cu el Capitulo anterior. Recordemos que para la mayor parle de las cadenas de Markov, la $P_{ij}(n)$ tendía a un límite $\pi_j$ que era independiente del estado inicial i. Igualmente, sucede que para muchos sistemas de cola, $P_{ij}(t)$, para valores grandes de t, tendera a un límite $\pi_j$, que es independiente del estado inicial i. A $\pi_j$, se le llama estado estable, o probabilidad de equilibrio del estado j.

Para los sistemas de cola que analizaremos, se puede imaginar que $\pi_j$ es la probabilidad que en un instante en el futuro lejano haya presentes j clientes. También, se puede pensar que $\pi_j$, es, para un futuro lejano, la fracción del tiempo que hay j clientes presentes. En la mayor parte de los sistemas de cola, el valor $P_{ij}(t)$ para t pequeña depende de gran medida de i, el número de clientes presentes al inicio. Por ejemplo, si t es pequeño podríamos esperar entonces que $P_{1\ 50}(t)$ y $P_{1\ 1}(t)$, fueran muy distintos. Sin embargo, si existen probabilidades de estado estable, entonces para t grande, tanto $P_{1\ 50}(t)$ como $P_{1\ 1}(t)$ se acercan a $\pi_1$. La pregunta que tan grande debe ser t para que se alcance el estado estable en forma aproximada es difícil de contestar. El comportamiento de $P_{ij}(t)$ antes de alcanzar el estado estable se llama comportamiento transitorio de sistema de cola. Para todos los sistemas de cola, excepto los mas sencillos, es extremadamente difícil el análisis del comportamiento transitorio del sistema. Por esta razón, cuando analizamos el comportamiento de un sistema de cola suponemos que. se ha alcanzado en el estado estable. Esto nos permite trabajar con las $\pi_j$, en lugar de con las $P_{ij}(t)$.

A continuación, analizaremos una clase especial de procesos continuos en el tiempo, que comprende muchos sistemas interesantes de cola. Para un proceso de nacimiento y muerte es fácil determinar las probabilidades de estado estable, si es que existen.

Un proceso de nacimiento y muerte es un proceso estocástico continuo en el tiempo para el que el estado del sistema en cualquier tiempo es un entero no

negativo (véase en la Sección 1.1 la definición de proceso estocástico continuo en el tiempo). Si un proceso de nacimiento y muerte está en el estado j cuando el tiempo es t, entonces el movimiento del proceso está gobernado por las leyes siguientes:

**LEYES DEL MOVIMIENTO PARA PROCESOS DE NACIMIENTO Y MUERTE**

l$^{era}$ ley Con probabilidad $\lambda_j \Delta t + o(\Delta t)$, sucede un nacimiento entre el tiempo t y t + $\Delta t$. Un nacimiento aumenta en 1 el estado del sistema, hasta alcanzar a j + 1. La variable $\lambda_j$ se llama tasa de natalidad en el estado j. En la mayor parle de los sistemas de cola, un nacimiento es simplemente una llegada.

2$^{da}$ ley Con probabilidad $\mu_j \Delta t + o(\Delta t)$, sucede una muerte entre el tiempo t y el tiempo t + $\Delta t$. Una muerte disminuye en 1 el estado del sistema, para llegar a j-1. La variable $\mu_j$, es la tasa de mortalidad en el estado j. En la mayor parte de los sistemas de colas, una muerte es el término del servicio. Nótese que se debe cumplir $\mu_0 = 0$, porque de otro modo podría tenerse un estado negativo.

3$^{era}$ ley Los nacimientos y muertes son independientes entre si.

Las leyes 1 a 3 se pueden usar para demostrar que la probabilidad que se tenga mas de un evento (nacimiento o muerte) entre los tiempos t y t + $\Delta t$ o( $\Delta t$ ), Nótese que todo proceso de nacimiento y muerte queda completamente especificado si se saben las tasas de natalidad $\lambda_j$, y tas lasas de mortalidad $\mu_j$,. Como no se puede tener un estado negativo, todo proceso de nacimiento y muerte debe cumplir con $\mu_0$ = O.

**RELACIÓN ENTRE LA DISTRIBUCIÓN EXPONENCIAL CON LOS PROCESOS DE NACIMIENTO Y MUERTE**

La mayor parte de loa sistemas de cola con tiempos exponenciales entre llegadas y tiempos exponenciales de servicio pueden modelarse como procesos de nacimiento y muerte. Para ilustrar porque es así, veamos un sistema M/M/l/FIFO/∞/∞/ de colas, en el que los Tiempos entre llegadas son exponenciales con parámetro $\lambda$ y los tiempos de servicio se distribuyen en forma exponencial con parámetro $\mu$. Si el estado (número de personas presentes cuando el tiempo es t) en el tiempo t es j, entonces la propiedad de amnesia de la distribución exponencial quiere decir que la probabilidad de un nacimiento durante el intervalo de tiempo [t,t+$\Delta t$ ]no dependerá de cuánto tiempo ha estado el sistema en el estado j. Esto quiere decir que la probabilidad que haya un nacimiento durante [[t,t+$\Delta t$ ]no dependerá de cuánto tiempo haya estado el sistema en el estado j y, por lo tanto, se puede determinar como si acabara de ocurrir una llegada en el tiempo t. Entonces, la probabilidad que se tenga un nacimiento durante [t,t + $\Delta t$ ] es

$$\int_0^{\Delta t} \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda \Delta t}$$

Según el desarrollo en series de Taylor para la exponencial

$$e^{-\lambda \Delta t} = 1 - \lambda \Delta t + o(\Delta t)$$

Esto quiere decir que la probabilidad que suceda un nacimiento durante [[t,t+$\Delta t$ ] es $\lambda \Delta t + o(\Delta t)$. Por lo anterior podemos concluir que la lasa de natalidad en el estado j simplemente es la tasa de llegadas $\lambda$.

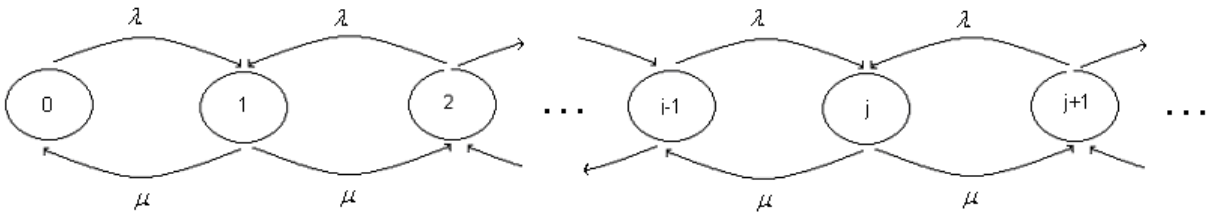
Para determinar la lasa de mortalidad cuando el tiempo es t, observe que si el estado es cero cuando el tiempo es t, entonces no hay nadie en la ventanilla y, por lo tanto, no se puede completar un servicio entre t y t+$\Delta t$. Asi, $\mu_0$ = 0. Sí el estado es j ≥ 1 cuando el tiempo es t, entonces sabemos (sólo hay un servidor) que hay exactamente un cliente en la ventanilla. La propiedad de amnesia de la distribución exponencial significa entonces que la probabilidad de que un cliente termine sus trámites entre t y t+$\Delta t$. está expresada por

$$\int_0^{\Delta t} \mu e^{-\mu t}\, dt = 1 - e^{-\mu \Delta t} = \mu \Delta t + o(\Delta t)$$

Así, $\mu_t$ = $\mu$ cuando j ≥ 1. En resumen, si suponemos que los términos de trámites y las llegadas suceden en forma independiente, entonces un sistema de colas M/M/l/FIFO/∞/∞ es un proceso de nacimiento y muerte. Las tasas de natalidad y de

mortalidad para este sistema se pueden representar en un diagrama de tasas como el de la Fig. 8.

Figura 8
Diagrama de rapidez para un sistema de colas M/M/l/FIFO/∞/∞

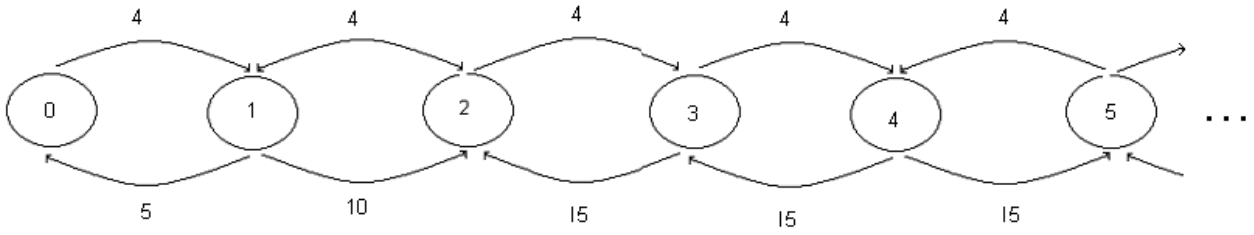



Los sistemas más complicados de colas con tiempos exponenciales entre llegadas y tiempos exponenciales de servicio se pueden modelar con frecuencia como procesos de nacimiento y muerte, sumando la rapidez de servicio en los servidores ocupados y agregando la tasa o rapidez de llegada para distintas corrientes de llegada. Por ejemplo, tenemos un sistema de colas M/M/3/FIFO/∞/∞ en el que los tiempos entre llegadas son exponenciales con $\lambda = 4$ y los tiempos de servicio son exponenciales con $\mu = 5$. Para modelar este sistema como proceso de nacimiento y muerte utilizaríamos los siguientes parámetros (véase Fig. 9):

$\lambda_j = 4 \quad (j = 0,1.2....)$

$\mu_0 = 0,\ \mu_1 = 5,\ \mu_2 = 5 + 5,\ \mu_j\ 5 + 5 + 5 = 15 \quad \{j = 3,4,5,...\}$

Figura 9 Diagrama de rapidez para un Estado sistema de colas M/M/3/FIFO/∞/∞




Si los tiempos entre llegada o los de servicio no son exponenciales, entonces el modelo de proceso de nacimiento y muerte no es adecuado. Suponga, por ejemplo, que los tiempos de servicio no son exponenciales y que estamos viendo un sistema de cola M/G/1/FIFO/∞/∞. Como los tiempos de servicio para ese sistema pueden no ser exponenciales, la probabilidad de una muerte (término del servicio) entre t y $t+\Delta t$ dependerá del tiempo que ha transcurrido desde el termino del último servicio. Esto viola la 2a ley, por lo que no podemos modelar ese sistema como proceso de nacimiento y muerte.

**DEDUCCIÓN DE LAS PROBABILIDADES DE ESTADO ESTABLE PARA PROCESOS DE NACIMIENTO Y MUERTE**

A continuación mostramos cómo se pueden calcular las $\pi_j$, para un proceso arbitrario de nacimiento y muerte. La clave es relacionar, para $\Delta t$ pequeño, a $P_{ij}(t+\Delta t)$ con $Pij(\Delta t)$. La manera de hacerlo es notar que hay cuatro modos que el estado sea j cuando el tiempo sea $t+\Delta t$. Para $j \geq 1$, los cuatro modos se muestran en la Tabla 3. Para $j \geq 1$ la probabilidad de que el estado del sistema sea j-1 cuando el tiempo es t, y j cuando el tiempo es $t+\Delta t$ es (véase Fig. 10)

$P_{ij-1}(t)(\lambda_{j-1}\Delta t + o(\Delta t))$

Con argumentos semejantes se obtienen (II) y (III). El (IV) es consecuencia porque si el sistema está en un estado que no sea j, j-1 o j+1 en el tiempo t, entonces, para finalizar en el estado j cuando el tiempo es $t+\Delta t$ debe suceder más de un evento (muerte o nacimiento) entre $t+\Delta t$. Según la 3era ley, esto tiene la probabilidad $o(\Delta t)$. Entonces,

$P_{ij}(t+\Delta t) = (I) + (II) + (III) + (IV)$

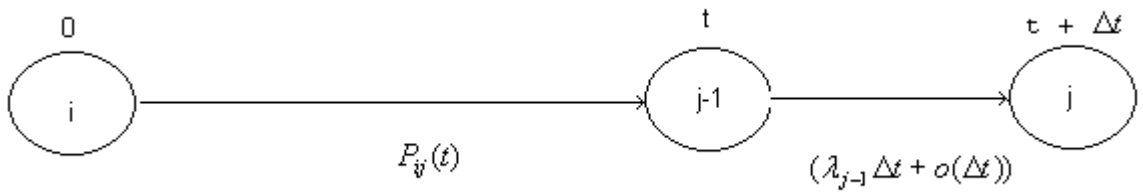
Después de reagrupar los términos en esta ecuación, obtenemos

$$(10)\quad P_{ij}(t+\Delta t) = \begin{aligned} & P_{ij}(t) + \\ & \Delta t(\lambda_{j-1}P_{ij-1}(t) + \mu_{j+1}P_{ij+1}(t) - \lambda_j P_{ij}(t) - \mu_j P_{ij}(t)) \\ & + o(\Delta t)(P_{ij-1}(t) + P_{ij+1}(t) + 1 - 2P_{ij}(t)) \end{aligned}$$

Tabla 3
Cálculos para la probabilidad que el estado sea j en el tiempo $t+\Delta t$

| ESTADO EN EL TIEMPO t | ESTADO EN EL TIEMPO $t+\Delta t$ | PROBABILIDAD DE LA SUCECCIÓN |
|---|---|---|
| j-1 | j | $P_{ij-1}(t)(\lambda_{j-1}\Delta t + o(\Delta t)) = I$ |
| j+1 | j | $P_{ij+1}(t)(\mu_{j+1}\Delta t + o(\Delta t)) = II$ |
| j | j | $P_{ij}(t)(1-\lambda_j\Delta t - \mu_j\Delta t - 2o(\Delta t)) = III$ |
| Cualquier otro | j | $o(\Delta t) = IV$ |

Figura 10
Probabilidad de que el estado sea j - 1 cuando el tiempo es t, y que
Sea j cuando el tiempo sea $t+\Delta t$. Esta probabilidad $P_{ij-1}(t)(\lambda_{j-1}(t)\Delta t + o(\Delta t))$



Como el termino subrayado se puede escribir como o( $\Delta$t), podemos volver escribir la Ecuación (10) como sigue:

$$P_{ij}(t+\Delta t) - P_{ij}(t) = \Delta t(\lambda_{j-1}P_{ij-1}(t) + \mu_{j+1}P_{ij+1}(t) - \lambda_j P_{ij}(t) - \mu_j P_{ij}(t)) + o(\Delta t)$$

Dividiendo ambos lados de esta ecuación entre $\Delta$t y haciendo que $\Delta$t tienda a cero vemos que para toda i y j $\geq$1

(10') $$P'_{ij}(t) = \lambda_{j-1}P_{ij-1}(t) + \mu_{j+1}P_{ij+1}(t) - \lambda_j P_{ij}(t) - \mu_j P_{ij}(t)$$

Como para j = O, $P_{ij-1}(0)$ = O y $\mu_j$ = O, obtenemos, para j = O,

$$P'_{i0}(t) = \lambda_0 P_{i0}(t) + \mu_1 P_{i1}(t)$$

Es un sistema infinito de ecuaciones diferenciales. (Una ecuación diferencial simplemente es una ecuación en la que aparece una derivada. En teoría, $P_{ij}$(t) se puede despejar de estas ecuaciones. Sin embargo, este sistema de ecuaciones es en realidad extremadamente difícil de resolver. Sin embargo, no todo está perdido. Podemos usar la Ecuación(10') para obtener las probabilidades $\pi_j$, de estado estable (j = 0,1,2, . . .). Como en el caso de las cadenas de Markov, definimos la probabilidad estado estable, $\pi_j$ como

$$\lim_{t\leftarrow\infty} P_{ij}(t)$$

Entonces, para t grande y cualquier estado inicial i, $P_{ij}$(t) no cambia mucho y se puede decir que es constante. Así, en el estado estable (t grande), $P'_{ij}(t)$=0. También, en el estado estable se cumplirá que $P_{ij-1}(t) = \pi_{j-1}$, $P_{ij+1}(t) = \pi_{j+1}$. Sustituyendo estas en la relaciones en la Ecuación (10') obtenemos, para j $\geq$ 1

(10") $$0 = \lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1} - \lambda_j\pi_j - \mu_j\pi_j$$
$$\lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1} = \pi_j(\lambda_j + \mu_j)$$

Para j = O se obtiene

$$\mu_1\pi_1 = \lambda_0\pi_0$$

Las Ecuación (10") son un sistema infinito de ecuaciones lineales del cual se pueden despejar con facilidad las $\pi_j$. Antes de explicar cómo se resuelven las Ecuaciones (10"), presentaremos una deducción intuitiva de dichas ecuaciones, basada en la siguiente observación: En cualquier tiempo que observemos un proceso de nacimiento y muerte, debe cumplirse que para cada estado j, el numero de veces que hemos entrado al estado j difiere cuando mucho 1 del número de veces que liemos dejado el estado j.

Suponga que cuando el tiempo es t hemos entrado tres veces al estado 6. Entonces debe haber sucedido uno de los casos de la Tabla 4. Por ejemplo, si sucede el Caso 2, iniciamos en el estado 6 y terminamos en otro estado que no sea 6. Como hemos observado tres transiciones de entrada al estado 6 cuando el tiempo es t, deben haber sucedido los eventos siguientes (entre otros):

Iniciar en estado 6
Dejar el estado 6 (primera vez)
Entrar al estado 6 (primera vez)
Dejar al estado 6 (segunda vc7)

Entrar al estado 6 (segunda vez)
Dejar el estado 6 (tercera vez)
Entrar al estado 6 (tercera vez)
Dejar el estado 6 (cuarta vez)

Tabla 4
Relación entre el número de transiciones entre el número de entrada y salidas de un estado en el tiempo 6

| ESTADO INICIAL | ESTADO EN EL TIEMPO t | NÙMERO DE TANSACIONES DE SALIDA DEL ESTADO 6 CUANDO EL TEIMPO ES T |
|---|---|---|
| Caso 1: estado 6 | Estado 6 | 3 |
| Caso 2: estado 6 | Cualquier estado menos el 6 | 4 |
| Caso 3: Cualquier estado menos el 6 | Estado 6 | 2 |
| Caso 3: Cualquier estado menos el 6 | Cualquier estado menos el 6 | 3 |

Por lo tanto, si sucede el Caso 2, entonces cuando el tiempo es t debemos haber dejado cuatro veces el estado 6.

Esta observación sugiere que para t grande y para j=0, 1, 2, ... y para cualesquiera condiciones iniciales, se cumplirá lo siguiente:

(11) $\dfrac{\text{Numero esperado de salidas del estado j}}{\text{unidad de tiempo}} = \dfrac{\text{Número esperado de entradas al estado j}}{\text{unidad de tiempo}}$

Si se supone que el sistema se ha asentado en el estado estable, sabemos que una fracción $\pi_j$ del tiempo en el estado j. Ahora podemos aplicar la Ecuación (11) para determinar las probabilidades $\pi_j$ de estado estable. Para $j \geq 1$ sólo podemos dejar el estado j pasando al estado j + 1 o al j- 1 y, entonces, cuando $j \geq 1$ obtenemos

(12) $\dfrac{\text{Número esperado de salidas al estado j}}{\text{Unidad de tiempo}} = \pi_j(\lambda_j + \mu_j)$

Como para $j \geq 1$ sólo podemos entrar al estado j desde el estado j - 1 o el j + 1

(13) $\dfrac{\text{Número esperado de entradas al estado j}}{\text{Unidad de tiempo}} = \lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1}$

Sustituyendo (12) y (13) en (11) se obtiene que

(14) $\lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1} = \pi_j(\lambda_j + \mu_j)$

Para j = 0 se obtiene

(14') $\mu_1\pi_1 = \lambda_0\pi_0$

Las Ecuaciones (14) y (14') se llaman ecuaciones de balance de flujo, o ecuaciones de conservación de flujo para un proceso de nacimiento y muerte. Observe que la ecuación (14) expresa el hecho de que en el estado estable, la rapidez a la que se tienen transiciones de entrada a cualquier estado i debe ser igual a la rapidez a la que tienen transiciones de salida del estado i. Si no se cumpliera la Ec. (14) para todos los estados, entonces la probabilidad se acumularía en algún estado, y no existiría un estado estable.

Planteando las ecuaciones para (14) y (14') obtenemos las ecuaciones de balance de flujo para un proceso de nacimiento y muerte:

(15)
$$\begin{aligned}
&(j=0) && \lambda_0\pi_0 = \mu_1\pi_1 \\
&(j=1) && \lambda_0\pi_0 + \mu_2\pi_2 = \pi_1(\lambda_1 + \mu_1) \\
&(j=2) && \lambda_1\pi_1 + \mu_3\pi_3 = \pi_2(\lambda_2 + \mu_2) \\
&\vdots \\
&j-\acute{e}sima && \lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1} = \pi_j(\lambda_j + \mu_j)
\end{aligned}$$

**SOLUCIÓN DE LAS ECUACIONES DE BALANCE DE FLUJO DE NACIMIENTO Y MUERTE**

Para resolver el sistema (15) comenzamos expresando todas las $\pi_j$ en términos de $\pi_0$. De la ecuación (j = 0) obtenemos

$$\pi_1 = \frac{\lambda_0\pi_0}{\mu_1}$$

Sustituyendo este resultado en la ecuación para j= 1, se obtiene

$$\lambda_0\pi_0 + \mu_2\pi_2 = (\lambda_1 + \mu_1)\frac{\lambda_0\pi_0}{\mu_1} \Rightarrow$$

$$\pi_2 = \frac{\lambda_0\lambda_1\pi_0}{\mu_1\mu_2}$$

Podríamos aplicar ahora para j=3, de la misma forma y así sucesivamente, obtenemos que

(16) $\pi_j = c_j\pi_0$

donde $c_j = \dfrac{\lambda_0\lambda_1\cdots\lambda_{j-1}}{\mu_1\mu_2\cdots\mu_j}$

Como en cualquier momento debemos estar en algún estado, las probabilidades de estado estable deben sumar 1:

(17) $$\sum_{j=0}^{\infty} \pi_j = 1$$

Sustituyendo (16) en (17) se obtiene

(18) $$\pi_0 (1 + \sum_{j=1}^{\infty} c_j) = 1$$

Si $\sum_{j=1}^{\infty} c_j$ es finita podemos usar ln Ecuación (18) para despejar $\pi_0$

(19) $$\pi_0 = \frac{1}{1 + \sum_{j=1}^{\infty} c_j}$$

Entonces con la Ecuación (l6) se calcula las probabilidades de estado estable. Se puede mostrar que $\sum_{j=1}^{\infty} c_j$ si finita entonces no existe distribución de estado estable. La causa más común de que no haya estado estable es que la rapidez de llegada sea cuando menos tan grande como la velocidad de atención.

En las Secciones 2.4 a 2.6, 2.9 y 2.10. aplicaremos la teoría de los proceso de nacimiento y muerte para determinar las distribuciones de probabilidad de estado estable para varios sistemas de colas.

Los modelos de nacimiento y muerte se han utilizado para modelar fenómeno distintos de los sistemas de colas. Por ejemplo, el número de empresas de la industria se puede modelar como proceso de nacimiento y muerte el estado de la industria en cualquier momento dado es el número de empresas que pertenecen a dicha industria; un nacimiento corresponde a una empresa que entra en la industria y una muerte corresponde a una empresa que sale de la industria.

## PROBLEMAS

1. En mi casa hay dos focos. En promedio, un foco dura 22 días (duración distribuida en forma exponencial). Cuando hay que cambiar un foco, me tardo, con distribución exponencial, 2 días en cambiarlo.

a) Formule un modelo de nacimiento y muerte de tres estados para este caso.
b) Determine la fracción del tiempo en que ambos focos están trabajando.
c) Determine la tracción del tiempo en la que no hay foco que funcione.

# 2.4 SISTEMA DE COLAS M/M/1/DG/∞/∞ Y LA FÓRMULA L = λW

A continuación usamos la metodología de nacimiento y muerte de la sección anterior para analizar las propiedades del sistema M/M/1/DG/∞/∞. Recuerde que en los sistemas de tiempos entre llegadas exponenciales, que suponemos que la rapidez de llegadas por unidad de tiempo es λ y que hay un solo servidor con tiempos de servicio exponenciales. Asimismo se supone que el tiempo de servicio para cada cliente es exponencial con rápido μ, bajo el esquema de nacimiento y muerte tenemos los parámetros siguientes:

$$\lambda_j = \lambda \quad (j = 0,1,2,\cdots)$$

(20) $$\mu_0 = 0$$

$$\mu_j = \mu \quad (j = 1,2,\cdots)$$

DEDUCCIÓN DE LAS PROBABILIDADES DE ESTADO ESTABLE

Podemos usar las Ecuaciones (15) a (19) para despejar las $\pi_j$, las probabilidades de estado estable de que haya j clientes. Sustituyendo la Ecuación (20) en la Ecuación (16), se obtiene

(21) $$\pi_1 = \frac{\lambda}{\mu}\pi_0, \quad \pi_2 = \frac{\lambda^2}{\mu^2}\pi_0, \quad \cdots, \quad \pi_j = \frac{\lambda^j}{\mu^j}\pi_0$$

Definimos a $\rho = \frac{\lambda}{\mu}$. Por razones que después se verán, llamaremos a ρ intensidad de trafico del sistema de colas. Sustituyendo la Ecuación (21) en la (17) se obtiene

(22) $$\pi_0(1 + \rho + \rho^2 + \cdots) = 1$$

A continuación suponemos que O < p < 1. Entonces evaluamos la suma S = $(1 + \rho + \rho^2 + \cdots)$ como sigue: Al multiplicar S por ρ se oblicué ρS = $(\rho + \rho^2 + \cdots)$. Entonces S - ρS = 1 y

(23) $$S = \frac{1}{1-\rho}$$

Sustituyendo la Ecuación (23} en la (22), se obtiene

(24) $$\pi_0 = (1-\rho) \quad (0 \le \rho < 1)$$

Sustituyendo la Ecuación (24) en la Ecuación (21) se obtiene

(25) $$\pi_j = (1-\rho)\rho^j \quad (0 \le \rho < 1)$$

Sin embargo, si ρ > 1, la suma infinita de la Ecuación (22) "explota" (trátese, por ejemplo, ρ= 1 y se obtendrá 1 + 1 + 1 + ...., así, si ρ ≥ 1 no existe

distribución de estado estable. Como $\rho = \frac{\lambda}{\mu}$ vemos que si $\lambda \geq \mu$ (esto es, la rapidez de llegadas es cuando menos tan grande como la rapidez de servicio), entonces no existe distribución de estado estable.

**DEDUCCIÓN DE L**

En el resto de esta sección suponemos que $\rho < 1$, asegurando que si existe una distribución de probabilidad de estado estable, como la que representa la Ecuación (25). Utilizamos entonces la distribución de probabilidad de estado estable dada en (25) para determinar varias cantidades de interés. Por ejemplo, si se supone que se ha alcanzado el estado estable, el número esperado de clientes presentes en el sistema de colas, L, está dado por

$$L = \sum_{j=0}^{\infty} j\pi_j = \sum_{j=0}^{\infty} j\rho^j(1-\rho) = (1-\rho)\sum_{j=0}^{\infty} j\rho^j$$

Definiendo a

$$S' = \sum_{j=0}^{\infty} j\rho^j = \rho + 2\rho^2 + 3\rho^3 + \cdots \Rightarrow \rho S' = \rho^2 + 2\rho^3 + 3\rho^4 + \cdots$$

Note que

$$S' - \rho S' = \rho + \rho^2 + \rho^3 + \rho^4 + \cdots = \frac{\rho}{1-\rho} \Rightarrow S' = \frac{\rho}{(1-\rho)^2}$$

Luego

(26) $$L = (1-\rho)S' = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu - \lambda}$$

**DEDUCCIÓN DE $L_q$**

En algunos casos nos interesa el número esperado de personas en la cola. Representamos por $L_q$, este número. Observe que si hay 0 o 1 cliente en el sistema, entonces no hay nadie en la cola, pero si hay j personas, donde $j \geq 1$, entonces habrá j-1 esperando en la cola. Entonces, si estamos en el estado estable

$$L_q = \sum_{j=1}^{\infty}(j-1)\pi_j = \sum_{j=1}^{\infty} j\pi_j - \sum_{j=1}^{\infty}\pi_j = L - (1-\pi_0) = L - \rho$$

y la última ecuación es consecuencia de la (24). Como L = $\frac{\rho}{(1-\rho)}$

escribimos

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

**DEDUCCIÓN DE $L_s$**

También $L_s$ es de interés es el número esperado de clientes en las ventanillas. Para un sistema de colas M/M/1/DG/∞/∞.

$$L_s = 0\pi_0 + \sum_{j=1}^{\infty} 1\pi_j = 1 - \pi_0 = 1 - (1-\rho) = \rho$$

Como cada cliente presente está en la cola o en la ventanilla, entonces, para cualquier sistema de colas, no solo para un M/M/1/DG/∞/∞. Así, utilizando las formulas L y $L_s$ podríamos haber determinado $L_q$ mediante

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

**FÓRMULA L =λW PARA COLAS**

Con frecuencia nos interesa el tiempo que pasa un cliente normal en una cola. Definimos W como et tiempo esperado que pasa el cliente en el sistema de colas, incluyendo el tiempo de la cola más el tiempo de servicio, y $W_q$, como el tiempo esperado que pasa el cliente en la cola. Tanto W como $W_q$, se calculan bajo la hipótesis que se ha alcanzado el estado estable. Si se utiliza un eficaz resultado, conocido como la formula de Little para colas, se calculan con facilidad W y $W_g$ a partir de L y $L_q$. Definimos primero, para todo sistema de colas o cualquier subsistema de colas, las cantidades siguientes:

λ = número promedio de llegadas que entran al sistema por unidad de tiempo
L = numero promedio de clientes presentes en el sistema de colas
$L_q$ = número promedio de clientes que esperan en la cola
$L_s$ = número promedio de clientes en el servicio (ventanilla)
W = tiempo promedio que pasa un cliente en el sistema
$W_q$ = tiempo promedio que pasa un cliente en la cola
$W_s$ = tiempo promedio que pasa un cliente en el servicio -

En estas definiciones, lodos los promedios son para estado estable. Para la mayor parte de los sistemas colas, la formula de Litlle se puede resumir en el Teorema 3.

**TEOREMA 3** Para cualquier sistemas de colas en que exista una distribución de estado estable, se cumplen las siguientes ecuaciones:

(28) $L = \lambda W$

(29) $L_q = \lambda W_q$

(30) $L_s = \lambda W_s$

Observe primero que ambos miembros de dicha ecuación tienen las mismas unidades (suponemos que la unidad de tiempo es la hora). Esto se debe a que L está expresada en términos de número de clientes, λ. en términos de clientes por hora y W en horas. Así, λW tiene las mismas unidades que L, clientes.

Partiendo del hecho que L = $\frac{\rho}{(1-\rho)}$, se tiene entonces que

(31) $$W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda}$$

De la ecuación (27) tenemos que $L_q = \frac{\rho^2}{(1-\rho)}$, luego obtenemos que

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

Nótese que, W y $W_q$, se hacen muy grandes cuando ρ se acerca a 1. Para ρ cercano a cero, Wq tiende a cero, pero para ρ pequeño, W tiende a λ el tiempo promedio de servicio,

Los tres ejemplos siguientes muestran aplicaciones de las formulas que hemos deducido.

EJEMPLO 2.
A un cajero bancario o automático sólo llega un promedio de 10 vehículos por hora. Suponga que los tiempos promedio de servicio para cada cliente es 4 minutos, y que los tiempos entre llegadas y los de servicio son exponenciales Conteste las siguientes preguntas

1. ¿Cuál es la probabilidad de que el cajero automático se encuentre vacío?
2. ¿Cuál es el número promedio de automóviles que esperan en la cola su turno? Se considera que un vehículo qu4e está ocupando el cajero automático, no está en la cola esperando.
3. ¿Cuál es el tiempo promedio que un cliente pasa en el estacionamiento del banco, incluyendo el tiempo en el servicio? |
4. En promedio, ¿cuántos clientes por hora serán atendidos por el cajero automático

Solución
Suponemos que nos ocupa un sistema de colas M/M/1/DG/∞/∞. para el cual λ = 10 automóviles por hora y μ= 15 automóviles por hora. Entonces ρ = 2/3

1. Según la Ec, (24), $\pi_0$= 1-ρ = 1- 2/3= 1/3. Entonces el cajero automático se encontrará sin clientes un prom3edio de la tercera parte del tiempo.

2. Queremos conocer Lq, según la ecuación (27),

$$L_q = \frac{\rho}{(1-\rho)} - \rho = \frac{\rho^2}{(1-\rho)} = \frac{\frac{2}{3}^2}{(1-\frac{2}{3})} = \frac{4}{3} \text{ clientes}$$

3. Buscamos saber W. Según la ecuación (28). W= L/λ, entonces según la ecuación (26),

$$L = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu - \lambda} = 2 \text{ clientes}$$

Así, W= 2/10 = 1/5 hora, unos 12 minutos.

4. Si el cajero automático estuviera ocupado siempre, podría atender un promedio de $\mu$=15 clientes por hora. De la parte 1 sabemos que sólo está ocupado dos tercer partes de su tiempo. Así, durante cada hora, el cajero atenderá aun promedio de 2/3 *15 = 10 clientes . Debe ser así, porque el estado estable llegan 10 clientes cada hora y por lo tanto deben salir 10 clientes del sistema cada hora.

EJEMPLO 3 Supongamos que todos los propietarios de automóviles llenan sus tanques de gasolina cuando están exactamente a la mitad.' En la actualidad, llega un promedio de 7.5 clientes por hora a una gasolinera que tiene una sola bomba. Se necesita un promedio de 4 minutos para atender un automóvil. Suponga que tanto los tiempos entre llegadas como los tiempos entre llegadas y de servicio son exponenciales.

1- Para el caso actual calcule L y W.
2. Suponga que se presenta escasez de gasolina y que hay compras de pánico. Para modelar este fenómeno. suponga que todos los propietarios de automóvil compran gasolina cuando a sus tanques les fallan exactamente ¾ partes. Como cada conductor pone menos gasolina al tanque durante cada visita a la gasolinera, suponga que el tiempo promedio de servicio se ha residuo a $3\frac{1}{3}$ minutos. ¿Cómo afecto la compra de pánico a L y a W? U'?

Solución

1. Tenemos un sistema de colas M/M/1/DG/$\infty$/$\infty$ con $\lambda$ = 7.5 vehículos/h y $\mu$ = 15 vehículos/h. Así $\rho$ = 7.5/15 = ½. De la Ecuación (26). L = ½ / (1 – ½ ) = 1 y de la Ecuación (28) W = $\frac{L}{\lambda}=\frac{2}{15}$= 0.13 h. Por lo tanto en este caso todo esta bajo control y son
improbables las largas colas.

2. Ahora tenemos un sistema M/M/1/DG/$\infty$/$\infty$ con $\lambda$ = = 2(7.5) = 15 vehículos/h. Esto es consecuencia de que cada automóvil llenará con el doble de frecuencia. Ahora $\mu$ = 60/(1/3) = 18 vehículos /h y $\rho$ = 15/18 = 5/6. Entonces L = $\frac{\frac{5}{6}}{1-\frac{5}{6}}$ = 5 automóviles y W = $\frac{L}{\lambda}=\frac{5}{15}$ = 1/3 horas = 20 minutos
Así, las compras de pánico han originado largas colas.

Cuando $\rho$ tiende a 1, L y W aumentan de rapidez como lo muestra la siguiente tabla

Tabla 5 Relación entre $\rho$ y L para un sistema M/M/1/DG/$\infty$/$\infty$

| $\rho$ | L para un sistema M/M/1/DG/$\infty$/$\infty$ |
|---|---|
| 0.30 | 0.43 |
| 0.40 | 0.67 |
| 0.50 | 1.00 |
| 0.60 | 1.50 |
| 0.70 | 2.33 |

| | |
|---|---|
| 0.80 | 4.00 |
| 0.90 | 9.00 |
| 0.95 | 19.00 |
| 0.99 | 99.00 |

MODELO PARA OPTIMIZAR UN SISTEMA DE COLAS

El Ejemplo 4 ilustra cómo se puede usar la teoría de colas como auxiliar para la toma de decisiones.

EJEMPLO 4 Los mecánicos que trabajan en una l planta de troquelado deben sacar herramientas de un almacén. Llega un promedio de diez mecánicos por hora buscando partes. En la actualidad el almacén esta a cargo de un empleado a quien se les pagan 6 dólares/h y gasta un promedio de 5 min. para entregar las herramientas de cada solicitud. Como a los mecánicos se les paga 10 dólares/h, cada hora que un mecánico pasa en el almacén de herramientas le cuesta 10 dólares a la empresa. Está ha de decidir si vale la pena contratar, a 4 dólares/h, un ayudante del almacén. Si se contrata al ayudante, el almacenista sólo tardara un promedio de 4 min. para atender las solicitudes de herramientas. Suponga que son exponenciales tantos los tiempos de servicio como el tiempo entre llegadas. ¿Se debe contratar al ayudante?

Solución
A los problemas en los que un tomador de decisiones debe escoger entre sistema alternativos de cola se les llama problemas para optimizar sistemas colas. En este problema, la meta de la empresa es minimizar la suma del costo horario de servicio y del costo horario esperado debido a los tiempos de inactividad de los mecánicos. En los problemas de optimización de colas, el componente del costo debido a clientes que esperan en la cola se llama costo de demora. Así, la empresa desea minimizar

$$\frac{\text{Costo esperado}}{\text{Hora}} = \frac{\text{costo de servicio}}{\text{hora}} + \frac{\text{costo esperado de demora}}{\text{hora}}$$

En general, el cálculo del costo horario de servicio es sencillo. El modo más fácil de calcular el costo horario de demora es tomando nota de que

$$\frac{\text{Costo esperado de demora}}{\text{hora}} = \frac{\text{costo esperado de demora}}{\text{cliente}} * \frac{\text{clientes esperados}}{\text{hora}}$$

En nuestro problema,

$$\frac{\text{Costo esperado de demora}}{\text{Clientes}} = \frac{\text{l0 dólares}}{\text{mecánico - hora}} \frac{\text{Promedio d horas que pasa}}{\text{el mecánico en el sistema}}$$

Asi,

$$\frac{\text{Costo esperado de demora}}{\text{cliente}} = 10W$$

costo esperado de demora = 10Wλ
hora

Ahora podemos comparar el costo esperado por hora, si no se contraía al ayudante con el correspondiente si se le contrata. Si no se contrata, λ = 10 mecánicos/h, y μ = 12 mecanicos/h. De la Ecuación (31), $W = \frac{1}{12-10} = \frac{1}{2}$. Como al despachador le paga 6 dólares por hora, tenemos que

Costo de servicio = 6 dólares
hora

costo esperado de demora = 10 ½ 10 = 50 dólares
hora

Así. sin el ayudante, el costo esperado por hora es 6 + 50 = 56 dólares. Con el ayudante. μ = 15 clientes por hora. Por lo tanto $W = \frac{1}{15-10} = \frac{1}{5}$ hora y

costo esperado de demora = 10 1/5 10 = 20 dólares
hora

Como el costo horario de servicio ahora es 6 + 4 = 10 dólares/h, el costo esperado de servicio con el ayudante es 20 + 10 = 30 dólares. Por lo tanto, se debe contratar al ayudante porque se ahorran 50 - 20 = 30 dólares/h en costos de demora, lo cual mas que compensa su salario de 4 dólares/h.

## PROBLEMAS

l. En una aerolínea se debe revisar cada pasajero, así como su equipaje, para ver si trae armas. Suponga que el aeropuerto Gotham Sity llega un pormdedio de 10 pasajeros/min. Los tiempos entre llegadas son exponenciales. Para revisar a los pasajeros, el aeropuerto debe tener una estación que consiste en un detector de metales y una máquina de rayos X para el equipaje. Cuando esta trabajando la estación se necesitan dos empleados. Una estación puede revisar un promedio de 12 pasajeros /min. El tiempo para revisar un pasajero es exponencial. Con la hipótesis que el aeropuerto sólo tiene una esta estación de verificación.

(a) ¿Cuál es la probabilidad de que un pasajero tenga que esperar para ser revisado?
(b) En promedio. ¿cuantos pasajeros esperan en la cola para entrar en la estación?
(c) En promedio cuanto tiempo pasa el pasajero en la estación de verificación?

2. El Departamento de Ciencias de la Decisión trata de determinar si renta una copiadora lenta o una rápida. El departamento cree que el tiempo de un empleado vale 15 dólares/h. El arrendamiento de la copiadora lenta cuesta 4 dólares/ h y un empleado tarda un promedio de 10 minutos en terminar sus copias, distribuido exponencialmente. La copiadora rápida cuesta 14 dólares/ h en arrendamiento y un empleado tarda un promedio de 6 minutos en termina r sus copias. Un promedio de 4 empleados/h son los que necesitan usar la copiadora. Los tiempos entre llegada son exponenciales. ¿Qué máquina debe rentar el Departamento?

3. Para un sistema de cola M/M/1/DG/∞/∞ suponga que se duplican tanto $\lambda$ y como $\mu$
(a) ¿Cambia L?
(b)¿Cambia W?
(c) ¿Cambia la distribución de probabilidad de estado estable?

4. En el Problema 1, suponga que la aerolínea desea determinar cuántas estaciones de revisión deben trabajar para reducir al mínimo los costos de operación y los costos de demora en un periodo de diez anos. Suponga que el costo de demora de un pasajero durante una hora es 10 dólares y que el aeropuerto abre todos los días, 16 horas al día. Cuesta un 1 millón de dólares comprar, operar y mantener un detector de metales y una máquina de rayos X para revisión de equipajes durante un periodo de 10 años. Por último, suponga que es igualmente probable que un pasajero entre a una estación dada.

5.1 Cada máquina de la línea de montaje de Widgetco sale de funcionamiento un promedio de una vez por minuto. Hay trabajadores asignados para restablecer una máquina con problemas. La empresa paga a cada trabajador $c_s$ dólares/h y calcula que cada hora de maquina sin funcionar le cuesta $c_m$ dólares por perdida de producción. Los dalos indican que el tiempo entre descomposturas sucesivas de una máquina y el tiempo para restablecerla son exponenciales. Widgetco desea asignar a cada trabajador un determinado número de máquinas que supervise y repare. Sea M = número total de máquinas de Widgelco, w = número de trabajadores contratados, y R = M/w máquinas asignadas a cada uno de ellos.

(a) Exprese el costo horario de Widgetco en términos de R y M
(b) Demuestre que el valor óptimo de R no depende del valor de M.
(c) Con calculo diferencial demuestre que los costos se reducen al mínimo si se escoge

$$R = \frac{\frac{\mu}{60}}{1+\left(\frac{C_m}{C_s}\right)^{1/2}}$$

(d) Suponga que $c_m$ = 78 centavos de dólar y que $c_s$ = 2.75 dólares. Widgetco tiene 200 máquinas y un trabajador puede restablecer una cualquiera de ellas en un promedio de 7.8 segundos. ¿Cómo puede Widgetco minimizar costos?

(e) En los incisos (a) a (d),hemos supuesto tácitamente que en cualquier momento, la rapidez con que se descomponen las máquinas asignadas a un trabajador no depende del número de las máquinas que trabajan bien y que están asignadas a el. ¿Parece razonable esta hipótesis?

6. Se tiene un aeropuerto donde los taxis y los clientes llegan con frecuencias respectivas de 1 y 2 por minuto. Los tiempos entre llegadas son exponenciales. Independientemente de cuántos otros taxis haya un taxi debe esperar. Si un cliente que llega no encuentra un taxi, se va de inmediato.

(a) Modele este sistema como proceso de nacimiento y muerte. Sugerencia: determine cuál es el estado del sistema en cualquier momento determinado y haga un diagrama de frecuencias,
(b) Determine el número promedio de taxis libres que esperan un cliente.
(c) Suponga que todos los clientes que usan taxi pagan 2 dólares por viaje como tarifa. Durante una hora normal, ¿cuántos ingresos recibirán los taxis?

7. Un banco Iraní de escoger entre dos máquinas cuál debe remar para comprobar cheques. La máquina 1 se alquila a 10000 dólares por año y procesa 1000 revisiones por hora. La tarifa de la máquina 2 es 15000 dólares por año y procesa 1600 revisiones por hora.

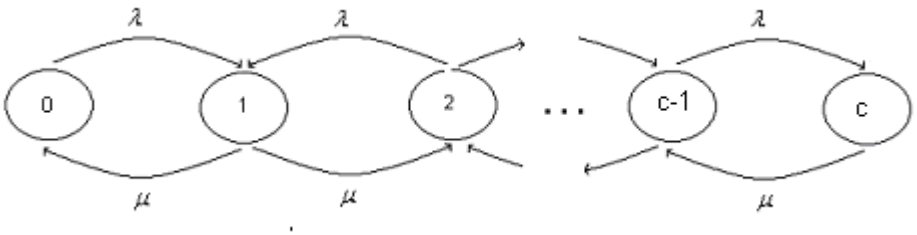
Suponga que las máquinas trabajan 8 horas diarias, 5 días por semana y 50 semanas por año. El banco debe procesar un promedio de 800 cheques por hora y el cheque procesado, en promedio es por 100 dólares. Suponga una tasa de interés anual de 20%. Luego calcule lo que cuesta al banco, en intereses perdidos, cada hora que tarda un cheque en esperar y procesarse.
Si se supone que los tiempos entre llegadas y los de servicio son exponenciales, cuál maquina se debe rentar?

8.- La planta de neumáticos debe producir un pedido de 100 neumáticos por día. La fabrica produce neumáticos en lotes de tamaño x. El gerente de planta debe determinar el tamaño de lote x que minimice el tiempo que pasa un lote en la planta. Desde que llega un lote de neumáticos, la fábrica larda un promedio de 1/20 de día en preparar la producción. Una vez que el equipo está preparado, se toma un promedio de 1/150 de día producir cada neumático. Suponga que el tiempo de producción de un lote de neumáticos tiene distribución exponencial y que el tiempo para que "llegue" un lote de neumáticos también tiene distribución exponencial. Calcule el tamaño de lote que minimice el tiempo esperado que pasa un lote en la fábrica, desde que llega un lote hasta que se termina la producción del lote.

## 2-5 SISTEMA DE COLAS M/M/1/DG/c/∞.

En esta sección analizaremos el sistema de colas M/M/1/DG/c/∞. Recuerde que este sistema es un M/M/1/DG/c/∞ con una capacidad total de c clientes. El sistema M/M/1/DG/c/∞ es idéntico al M/M/1/DG/∞/∞ con excepción de que cuando hay presentes c clientes, todas las llegadas se regresan y el sistema las pierde para siempre. Como en la Sección 22.4, suponemos que los tiempos entre llegadas son exponenciales con rapidez $\lambda$. y que los tiempos de servicio son exponenciales con rapidez $\mu$. Entonces el sistema M/M/1/DG/c/∞ se puede modelar (véase Figura 11) como proceso de nacimiento y muerte con los siguientes parámetros:

Figura 11
Diagrama de rapidez a partir sistema M/M/1/DG/c/∞



$$\lambda_j = \lambda \quad (j = 0,1,2,\cdots,c-1)$$
(35) $$\mu_0 = 0$$
$$\mu_j = \mu \quad (j = 1,2,\cdots,c)$$

Como $\lambda_c = 0$, el sistema nunca alcanzara el estado c + 1, o cualquier otro estado de numero mayor. Como en la Sección 2.4, conviene definir a $\rho = \frac{\lambda}{\mu}$. Luego aplicamos las ecuación (16) a (19) para ver que si $\lambda \neq \mu$ las probabilidades de estado estable para el M/M/1/DG/c/$\infty$ están expresadas por

$$\pi_0 = \frac{1-\rho}{1-\rho^c},$$
(34) $$\pi_j = \rho^j \pi_0 \quad j = 1,\cdots,c$$
$$\pi_j = 0 \quad j = c+1,\cdots$$

combinando la ecuación (34) con el hecho de que $L = \sum_{j=1}^{c} j\pi_j$ podemos demostrar que si $\lambda \neq \mu$

(35) $$L = \frac{\rho\left[1-(c+1)\rho^c + c\rho^{c+1}\right]}{(1-\rho)(1-\rho^{c+1})},$$

Si $\lambda = \mu$, entonces todas las $c_j$ de la Ecuación (16) son guales a 1, y todas las $\pi_j$, deben ser iguales. Por lo lanío, si $\lambda = \mu$ las probabilidades de estado estable del sistema M/M/1/DG/c/$\infty$

$$\pi_j = \frac{1}{c+1}, \quad j = 1,2,\cdots,c$$
(36)
$$L = \frac{c}{2}$$

Como el caso del sistema M/M/1/DG/$\infty$/$\infty$ $L_s = 0\pi_0 + \sum_{j=1}^{\infty} 1\pi_j = 1-\pi_0$. Como antes, podemos calcular $L_q$ mediante $L_q = L - L_s$.

El cálculo de W y $W_q$ a partir de las Ecuaciones (28) y (29) tiene sus trucos. Recuerde que en las Ecuaciones (28) y (29), $\lambda$ representa el número promedio de clientes que llegan por unidad de tiempo, quienes realmente entran al sistema. En nuestro modelo de capacidad finita, llega un promedio de $\lambda$, pero $\lambda\pi_c$ de esas llegadas encuentran al sistema lleno a toda capacidad y se van. Por lo tanto, en

realidad entrará al sistema un promedio de $\lambda-\lambda\pi_c=\lambda(1-\pi_c)$ llegadas por unidad de tiempo. Al combinar este hecho con las Ecuaciones (28) y (29) se obtiene

(37) $$W=\frac{L}{\lambda(1-\pi_c)} \qquad W_q=\frac{L_q}{\lambda(1-\pi_c)}$$

Para un sistema M/M/1/DG/c/$\infty$ existirá estado estable aun si $\lambda \geq \mu$ .Esto se debe a que aun cuando $\lambda \geq \mu$, la capacidad finita del sistema evita que "explote" el número de gentes en la cola.

Ejemplo 5.

En una peluquería hay un peluquero y un total de 10 asientos. Los tiempos de llegada tienen distribución exponenciales, y llega un promedio de 20 clientes posibles por hora. Los que llegan cuando la peluquería está llena no entran. El peluquero larda un promedio de 12 minutos en atender a cada cliente. Los tiempos de corte de pelo tienen distribución exponencial.

1. En promedio, ¿cuántos corles de pelo por hora hará el peluquero?

2. En promedio, ¿cuánto tiempo pasará un cliente en la peluquería cuando entra?

Solución

1. una fracción $\pi_{10}$ de las llegadas encuentra que está llena la peluquería. Por lo tanto, entrará a ella un promedio de $(1-\pi_{10})\lambda$ por hora. Todos los clientes que desean que se les corle el cabello, y por lo tanto, el peluquero hará un promedio de $(1-\pi_{10})\lambda$ cortes por hora. En nuestro problema, c = 10, $\lambda$=20 clientes por hora y $\mu$= 5 clientes /h. Entonces $\rho$= 20/5 = 4 y la Ecuación (34) da como resultado

$$\pi_0=\frac{1-4}{1-4^{11}}$$

$$\pi_{10}=4^{10}\left(\frac{1-4}{1-4^{11}}\right)=.75$$

Así, los cortes de pelo son en promedio 20(1 – ¾ ) =5/h. Esto significa que un promedio de 20 - 5 = 15 clientes posibles no entran cada hora.

2. Para calcular W usaremos las Ecuación (35) y (37). De la Ecuación (35),

$$L=\frac{4\left[1-(10+1)4^{10}+10\rho^{11}\right]}{(1-4\ )(1-4^{11})}=9.67 clientes$$

Entonces, la Ecuación (37) da como resultado

$$W=\frac{9.67}{20(1-\frac{3}{4})}=1.93 horas$$

Esta peluquería siempre está llena, y le debemos aconsejar al peluquero que contrate cuando menos a otro peluquero mas.

**PROBLEMAS**

1. Una instalación de servicio consiste de una persona que puede atender un promedio de 2 clientes/h. Los tiempos de servicio son exponenciales. Llega un promedio de 3 clientes por hora, y se supone que los tiempos entre llegadas son exponenciales. La capacidad del sistema es de 3 clientes.

a) En promedio, ¿cuántos clientes potenciales entran al sistema cada hora?
b) ¿Cuál es la probabilidad de que quien atiende este ocupado?

2. Hay un promedio de 40 automóviles por hora, con tiempos exponenciales entre llegadas, que desean que se les atienda en la ventanilla de "servicio en su auto" del Hot Dog King Restaurant. Si hay una cola de más de 4 coches, incluyendo el de la ventanilla, el coche que llegue se va. En promedio toman cuatro minutos en servir a un automóvil.

a) ¿Cual es el numero promedio de automóviles esperando en la cola, sin incluir al que está frente a la ventanilla?
b) En promedio, ¿a cuantos automóviles se atiende en cada hora?
c) Acabo de formarme en la cola. En promedio, ¿cuanto tiempo pasará para que reciba mis alimentos?

3. Hay dos peluquerías con un peluquero cada una, y los establecimientos están en la misma calle. Cada peluquería puede tener un máximo de 4 personas, y todo cliente potencial que encuentre que está llena no esperará. El peluquero 1 cobra 11 dólares por corte y se tarda un promedio de 12 minutos para atender a un cliente. El peluquero 2 cobra 5 dólares por corte y se tarda un promedio de 6 minutos para terminar un corte. A cada peluquería llega un promedio de 10 clientes posibles por hora. Naturalmente, un cliente posible se transforma en un cliente real sólo si encuentra que la peluquería no está llena. Si se supone que los tiempos entre llegadas son exponenciales, así como los tiempos de corte de pelo, ¿cuál peluquero gana más?

## 2-6 SISTEMA DE COLAS M/M/s/DG/∞/∞

A continuación analizaremos el sistema M/M/s/DG/∞/∞. Suponemos que los tiempos entre llegadas son exponenciales, con rapidez igual a $\lambda$, que los tiempos de servicio son exponenciales, con rapidez $\mu$ y que hay solo una cola de clientes esperando su servicio en una de las s ventanillas o servidores. Si hay $j \leq s$ clientes, entonces los j clientes están en el servicio; si hay $j > s$ clientes, entonces las s ventanillas están ocupadas, y hay $j - s$ clientes esperando en la cola. Toda llegada que encuentre una ventanilla vacía entra al servicio de inmediato, pero una llegada que no encuentre ventanilla vacía se forma en la cola de clientes que esperan su atención. Los bancos y las oficinas de correos, en los que todos los clientes esperan atención en una sola cola, se pueden modelar con frecuencia con los sistemas de cola M/M/s/DG/∞/∞.

Para describir el sistema M/M/s/DG/∞/∞ como modelo de nacimiento y muerte, observe que, como en el modelo M/M/l/DG/∞/∞ ,$\lambda_j = \lambda$ j = 0,1, 2,.... Si hay j ventanillas o servidores ocupados, entonces se terminó el servicio con una frecuencia de

$$\mu + \mu + \cdots = j\mu$$

Siempre que haya j clientes, estarán ocupadas min(j,s) ventanillas. Así, $\mu_j$ = min(j,s)$\mu$. En resumen, vemos que el sistema M/M/s/DG/∞/∞ se puede modelar como un proceso de nacimiento y muerte (véase Fig. 12) con parámetros

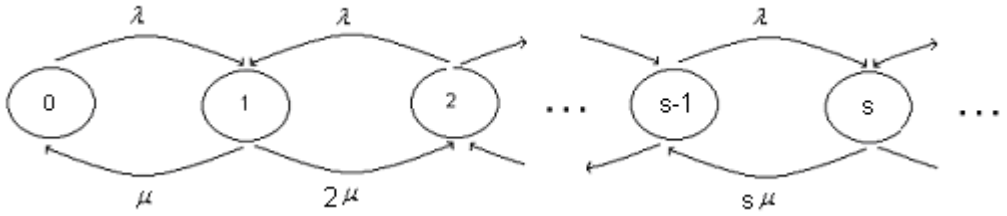
$$\lambda_j = \lambda \quad (j = 0,1,2,\cdots)$$

(38) $$\mu_j = j\mu \quad (j = 0,1,2,\cdots,s)$$

$$\mu_j = s\mu \quad (j = s+1, s+2,\cdots)$$

Finura 12

Diagrama de rapidez para un sistema de colas M/M/s/DG/∞/∞




Definimos a ρ como igual a = $\frac{\lambda}{\mu s}$. Para ρ < al sustituir la Ecuación (38) en (16) a (19) se obtienen las siguientes probabilidades de estado estable:

(39) $$\pi_0 = \frac{1}{\sum_{j=0}^{s-1} \frac{(s\rho)^j}{j!} + \frac{(s\rho)^s}{s!(1-\rho)}}$$

(39.1) $$\pi_j = \frac{(s\rho)^j}{j!}\pi_0 \quad j = 1,2,\cdots,s$$

(39.2) $$\pi_j = \frac{(s\rho)^j}{s!\,s^{j-s}}\pi_0 \quad j = s, s+1,\cdots$$

Si $\rho \geq 1$ no existe estado estable, en otras palabras, si la frecuencia o rapidez de llegadas es al menos igual a la rapidez máxima posible de servicio ( $\lambda > s\mu$), "explota" el sistema.

De la Ecuación (39.2), se puede demostrar que la probabilidad de estado estable de que todas las ventanillas estén ocupadas es

(40) $$P(j \geq s) = \frac{(s\rho)^s}{s!(1-\rho)}\pi_0$$

En la Tabla A se muestra P(j ≥ s) para diversas situaciones. También se puede demostrar que

(41) $$L_q = \frac{P(j \geq s)\rho}{(1-\rho)}$$

Tabla 6. $P(j \geq s)$ para un sistema M/M/s/DG/∞/∞

| ρ | s =2 | s=3 | s=4 | s=5 | s=6 | s=7 |
|---|---|---|---|---|---|---|
| 0.10 | 0.02 | - | - | - | - | - |
| 0.20 | 0.07 | 0.02 | - | - | - | - |
| 0.30 | 0.14 | 0.07 | 0.04 | 0.02 | - | - |
| 0.40 | 0.23 | 0.14 | 0.09 | 0.06 | 0.04 | 0.03 |
| 0.50 | 0.33 | 0.24 | 0.17 | 0.13 | 0.10 | 0.08 |
| 0.55 | 0.39 | 0.29 | 0.23 | 0.18 | 0.14 | 0.11 |
| 0.60 | 0.45 | 0.35 | 0.29 | 0.24 | 0.20 | 0.17 |
| 0.65 | 0.51 | 0.42 | 0.35 | 0.30 | 0.26 | 0.21 |
| 0.70 | 0.57 | 0.51 | 0.43 | 0.38 | 0.34 | 0.30 |
| 0.75 | 0.64 | 0.57 | 0.51 | 0.46 | 0.42 | 0.39 |
| 0.80 | 0.71 | 0.65 | 0.60 | 0.55 | 0.52 | 0.49 |
| 0.85 | 0.78 | 0.73 | 0.69 | 0.65 | 0.62 | 0.60 |
| 0.90 | 0.85 | 0.83 | 0.79 | 0.76 | 0.74 | 0.72 |
| 0.95 | 0.92 | 0.91 | 0.89 | 0.88 | 0.87 | 0.85 |

Entonces, la Ecuación (28) da como resultado

(42) $$W_q = \frac{L_q}{\lambda} = \frac{P(j \geq s)}{s\mu - \lambda}$$

Para calcular L y después W, usamos el hecho de que $L = L_q + L_s$. Como $W_s = \frac{1}{\mu}$ la Ecuación (30) muestra que $Ls = \frac{\lambda}{\mu}$ Entonces

(43) $L = L_q + \frac{\lambda}{\mu}$

También,

(44) $$W = \frac{L}{\lambda} = \frac{L_q}{\lambda} + \frac{1}{\mu} = W_q + \frac{1}{\mu} = \frac{P(j \geq s)}{s\mu - \lambda} + \frac{1}{\mu}$$

Cuando necesitemos calcular L, $L_q$, W o $W_q$, comenzamos por buscar $P(j \geq s)$ en la Tabla 6. A continuación aplicamos las Ecuación (41) a (44) para calcular la cantidad que deseemos. Si nos interesa la distribución probabilidad de estado estable, buscamos $P(j \geq s)$ en la Tabla 6 y luego usamos la Ecuación (40) para obtener $\pi_0$. Así, las Ecuaciones (39.1) y (39.2) producen la distribución completa de estado estable. Los dos ejemplos siguientes muestran el empleo de las formulas anteriores.

EJEMPLO 6.

Un banco tiene dos cajeros. Llegan al banco un promedio de 80 clientes por hora y esperan en una sola cola para que los atiendan. El tiempo promedio que se necesita para atender a un cliente es 1.2 minutos. Suponga que los tiempos entre llegadas y los de servicio son exponenciales. Calcule

1. Número esperado de clientes en el banco.
2. Tiempo esperado que pasa un cliente en el banco.
3. La fracción del tiempo que determinado cajero esta desocupado.

Solución

1. Tenemos un sistema M/M/2/DG/∞/∞. con λ = 80 clientes/h y μ = 50 clientes/h. Así, $\rho = \frac{80}{2*50} = .80 < 1$ y, por lo tanto, existe el estado estable. Si λ ≥ 100, no existiría estado estable. De la Tabal 6, P(J ≥ 2) = .71. Entonces de la Ecuación (41)

$$L_q = \frac{.80(.71)}{1-.80} = 2.84 clientes$$

y según la ecuación (43), L= 2.84+ 80/50 = 4.44 clientes.

2. Como W= $\frac{L}{\lambda}$, W=4.44/80 = .055 horas = 3.3 minutos

3. Para calcular la fracción del tiempo que determinado cajero está desocupado, nótese que esta desocupado durante lodo el tiempo que j = O, y la mitad del tiempo, por simetría, que j = 1. La probabilidad que una ventanilla esté ociosa está dada por $\pi_0 + \frac{1}{2}\pi_1$,, Aplicando el hecho de que P(j ≥ 2) = .71, obtenemos, con la Ecuación 40,

$$P(j \geq s) = \frac{(s\rho)^s}{s!(1-\rho)}\pi_0 \quad \Rightarrow \pi_0 = P(j \geq s)\frac{s!(1-\rho)}{(s\rho)^s} = .71\frac{2!(1-0.80)}{1.6^2} = .11$$

y según la Ecuación (39.1) da como resultado

$$\pi_1 = \frac{(s\rho)^1}{1!}\pi_0 = \frac{1.6}{1!}.11 = .176$$

Así, la probabilidad de que una ventanilla este vacía es $\pi_0 + \frac{1}{2}\pi_1$ = .11 + 0.5(.176) = .198. Podríamos haber calculado $\pi_{01}$directamente con la Ecuación (39) su comprobación se deja como ejercicio.

EJEMPLO 7
El gerente de un banco debe determinar cuántos cajeros deben trabajar los viernes. Por cada minuto que un cliente espera en la cola, el gerente supone que se incurre en un costo de 5 centavos de dólar. Al banco llegan un promedio de 2 clientes por minuto. En promedio, un cajero se tarda 2 minutos en tramitar la transacción de un cliente. Al banco le cuesta 9 dólares por hora la contratación de un cajero. Los tiempos entre llegadas y los tiempos de servicio son exponenciales. Para reducir al mínimo la suma de los costos de servicio y los de demora, ¿cuántos cajeros deben trabajar el banco los viernes?

Solución
Como λ = 2 clientes por minuto y μ = 0.5 clientes por minuto, ^ < 1 necesita que 4 $\frac{\lambda}{\mu s}$ < 1, o sea, s ≥ 5. Así, deben haber 5 cajeros cuando menos, porque de lo contrario "explotará" el número de clientes. A continuación calcularemos, para s = s, 6, ...

$$\frac{\text{Costo esperado de servicio}}{\text{min}} + \frac{\text{costo esperado de demora}}{\text{min}}$$

Como a cada cajero se le paga 9/60 =0.15 dólares /min,

$$\frac{\text{Costo esperado de servicio}}{\text{min}} = .015\ s \text{ dólares}$$

$$\frac{\text{Costo esperado de demora}}{\text{Min}} = \frac{\text{clientes esperados}}{\text{min}} \quad \frac{\text{costo esperado demora}}{\text{Cliente}}$$

Pero

$$\frac{\text{Costo esperado de demora}}{\text{Cliente}} = .05 \text{ dólares } W_q$$

Como llegan un promedio de 2 clientes por minuto,

$\frac{\text{Costo esperado de demora}}{\text{Min}}$ = 2(0.05 $W_q$) = 0.10 dólares $W_q$

Ya que s = 5 $\rho = \frac{2}{.5(5)} = .80$ P(j ≥ 5) = .55. Según la Ecuación (42),

$W_q$=.55/(.5*5-2) = 1.1 minutos

Así, para s = 5, el costo esperado de demora/ min = .10 *1.1 =.11 dólares.

y para s = 5,

Costo total esperado/ min = .15*5+.11=.86 dólares

Como s=6 tiene un costo de servicio de 6(0.15) = 0.90 dólares por minuto, 6 cajeros no pueden tener un costo total más bajo que 5 cajeros. Por lo tanto, lo óptimo es tener 5 cajeros de servicio. Visto de otro modo, si se añade un cajero más, el banco se puede ahorrar cuando más 0.11 dólares por minuto en tiempos de demora. Como un cajero más le cuesta al banco 0.15 dólares por minuto, no puede ser óptima la contratación de mas de 5 cajeros.

Además del tiempo esperado de un cliente en el sistema, es de interés la distribución del tiempo de espera de un cliente. Por ejemplo, si lodos los clientes que tienen que esperar más de 5 minutos en una caja de supermercado deciden cambiar a otra tienda, la probabilidad que un cliente dado cambie a otro almacenes igual a P(W > 5). Para calcular esta probabilidad necesitamos conocer la distribución del tiempo de espera de un cliente. Para un sistema de colas M/M/s/FIFO/∞/∞. (primero en llegar primero en ser servido), se puede demostrar que

(45) $P(W > t) = e^{-\mu t}\left(1 + P(j \geq s)\frac{1 - e^{[-\mu t(s-1-sp)]}}{s-1-sp}\right)$ (si s-1=sρ $P(W > t) = e^{-\mu t}\left(1 + P(j \geq s)\mu t\right)$

(46) $P(W_q > t) = P(j \geq s)e^{[-s\mu(1-\rho)t]}$

Para dar un ejemplo del uso de las Ecuaciones. (45) y (46), suponga que en el ejemplo 7, para s = 5, el gerente del banco desea conocer la probabilidad de que el cliente tenga que esperar en la cola durante más de 10 minutos. Para s = 5, ρ = .80, P(j ≥ 5) = .55 y μ = .5 clientes por minuto, la Ecuación (46) da como resultado

$P(W_q > 10) = .55e^{[-5*.5(1-.8)10]} = .004$

Por lo tanto, el gerente del banco puede estar seguro que la probabilidad de que un cliente espere mas de 10 minutos es muy baja.

•

**PROBLEMAS**

1. Un supermercado trata de decidir cuántas cajas deben estar funcionando. Suponga que cada hora llega un promedio de 18 clientes, y el tiempo promedio de atención a un cliente es 4 minutos. Los tiempos entre llegadas y los tiempos de servicio son exponenciales, y el sistema se puede modelar como uno M/M/s/DG/∞/∞. El funcionamiento de una caja cuesta 20 dolares/h, y se carga un costo de 15 centavos de dólar por cada minuto que el cliente pasa en la zona de cajas.
¿Cuantas cajas debe abrir el supermercado?

2. Un banco pequeño trata de determinar cuántos cajeros debe emplear. El costo total de emplear un cajero es, 100 dólares diarios y un cajero puede atender a un promedio de 60 clientes por día. Al banco llega un promedio de 50 clientes por día y de servicio y los tiempos entre llegadas son exponenciales. Si el costo de demora por cliente día es 100 dólares, ¿cuántos cajeros debe contratar el banco?

3. En este problema, todos los tiempos entre llegadas y de servicio son exponenciales,

a) En la actualidad, el departamento de finazas y el de ventas tienen sus propias mecanógrafas. Cada mecanógrafa puede hacer 25 cartas por día. El departamento de finanzas necesita que se mecanografíen 20 cartas diarias, en promedio, y el de ventas 15 cartas diarias. Para cada departamento, calcule el tiempo que pasa entre la petición de una carta y la terminación de ésta.

b) Suponga que las dos mecanógrafas se agrupan en una sección y que cada una quede disponible para mecanografiar cartas de cualquier departamento. Para este arreglo, calcule el tiempo promedio que pasa entre la petición de una carta y la terminación de esta.

c) Comente los resultados de los incisos (a) y (b).
d) Si las mecanógrafas pueden trabajar para cualquier departamento, ¿cual es la probabilidad que pase más de 0.20 de día entre la petición de carta y la terminación de esta?

4. MacBurger trata de determinar cuántos servidores, o colas, deben trabajar durante el turno del desayuno. Durante cada hora, llegan un promedio de 100 clientes al restaurante. Cada cola puede manejar un promedio de 50 clientes por hora. Un servidor cuesta 5 dólares por hora, y se carga un costo de 20 dólares por cada cliente que espere en la cola durante 1 hora. Suponiendo que se pueda aplicar un modelo M/M/s/DG/∞/∞, calcule el número de colas que minimice la suma de los costos de demora y de servicio.

5. Llega un promedio de 100 clientes por hora al banco de Gotham City. El tiempo promedio de servicio para cada cliente es 1 minuto.. Los tiempos de servicio y entre llegadas son exponenciales. El gerente desea asegurarse que no haya más del 1% de los clientes que tengan que esperar en la cola durante más de 5minutos. Si el banco tiene como política formar a todos los clientes en una cola única, ¿cuántos cajeros debe contratar?

6. Al banco de Gotham City llega un promedio de 100 clientes por hora. Un cajero se tarda un promedio de 2 minutos en atender a un cliente. Los tiempos entre llegadas y de servicio son exponenciales. El banco tiene actualmente cuatro cajeros trabajando. El gerente desea comparar los dos sistemas siguientes en

cuanto al número promedio de clientes en el banco y la probabilidad de que un cliente pase más de 8 minutos en el banco:

Sistema 1 Cada calero tiene su propia cola y no se permite cambiar de cola.

Sistema 2 Todos los clientes esperan en una cola única a que se desocupe un cajero.

Si el lector fuera el gerente del banco, ¿qué sistema escogería?

7. Un taller de silenciadores tiene tres mecánicos. Cada mecánico larda 45 minutos en instalar un silenciador nuevo. Suponga que llega un promedio de 1 cliente por hora. ¿Cuál es el número esperado de mecánicos que están desocupados en cualquier momento dado? Conteste esta pregunta sin suponer que los tiempos de servicio y los tiempos entre llegada sean exponenciales.

8. Se tienen los dos sistemas siguientes de colas:

Sistema 1 Sistema M/M/1 con frecuencia de llegada $\lambda$ y rapidez de servicio $3\mu$
Sistema 2 Sistema M/M/3 con frecuencia de llegada $\lambda$ y cada servidor trabaja a una velocidad $\mu$.

Sin hacer muchos cálculos, ¿cuál sistema tendrá las menores W y L? (Sugerencia: Escribir los parámetros de nacimiento y muerte para cada sistema.) A continuación determine cuál sistema es el más eficaz.

## 2-7 MODELOS M/M/∞/DG/∞/∞ Y GI/G/∞/DG/∞/∞

Hay muchos ejemplos de sistemas en los cuales un cliente nunca tiene que esperar para que se inicie su atención. En un sistema de éstos, se puede pensar que todo el tiempo que pasa el cliente en el sistema es su tiempo de servicio, o de trámite.
Como un cliente nunca tiene que esperar para que lo atiendan hay, en esencia, un servidor disponible para cada llegada, y podemos pensar que ese sistema es de cantidad infinita de servidores, o sistema de autoservicio. En la Tabla 7 se dan dos ejemplos de sistema de despachador infinito.

Tabla 7
Ejemplos de sistemas de colas con una cantidad infinita de servidores

| CASO | LLEGADA | TIEMPO EN EL SERVICIO (TIEMPO EN EL SISTEMA) | ESTADO DEL SISTEMA |
|---|---|---|---|
| Industria | La empresa entra a la industria | Tiempo hasta que la empresa sale de la industria | Número de empresas en la industria |
| Programa escolar | El estudiante entra al programa | Tiempo que el estudiante permanece en el programa | Número de estudiantes en el programa |

Mediante la notación Kendall-Lee, un sistema de servidores infinitos en el que los tiempos entre llegada y de servicio pueden seguir distribuciones arbitrarias de probabilidad, se representan como sistema de colas M/M/∞/DG/∞/∞. Un sistema de estos trabaja como sigue:

1. Los tiempos de llegadas son iid con distribución común A. Se define E(A)= $\frac{1}{\lambda}$. Así, $\lambda$ es la frecuencia o rapidez de llegadas.

2. Cuando llega un cliente, inmediatamente pasa al servicio. El tiempo de cada cliente en el sistema esta gobernado por la distribución S con E(S) = $\frac{1}{\mu}$.

Sea L el número esperado de clientes en el sistema en el estado estable, y W tiempo esperado que un cliente pasa en el sistema. Por definición, W = $\frac{1}{\mu}$. Entonces, la Ecuación (30) significa que

(47) $$L = \frac{\lambda}{\mu}$$

La Ecuación (47) no requiere hipótesis de exponencialidad. Si los tiempos entre, llegadas son exponenciales, se puede demostrar, aun para una distribución arbitraria de tiempos de servicio, que la probabilidad de estado estable $\pi_j$, de

que haya j clientes sigue una distribución de Poisson con promedio $\frac{\lambda}{\mu}$. Esto implica que

$$\pi_j = \frac{\left(\frac{\lambda}{\mu}\right)^j e^{-\lambda/\mu}}{j!}$$

El ejemplo que sigue es una aplicación representativa de un sistema M/M/∞/DG/∞/∞.

EJEMPLO 8
Durante cada año abren 3 neverías en el pueblo. El tiempo promedio que una nevería permanece en el negocio es 10 años. El 1ero de enero de 2525, ¿cuál es el número promedio de neverías que habrá en el pueblo? Si el tiempo entre inauguraciones de neverías es exponencial, ¿cuál es la probabilidad de estado estable que el 1ero de enero de 2525 haya 25 neverías en el pueblo?

Solución

Sabemos que λ= 3 neverías por año y $\frac{1}{\mu}$ = 10 años por nevería. Suponiendo que se haya alcanzado el estado estacionario, habrá un promedio de L = $\frac{\lambda}{\mu}$ = 3(10) = 30 neverías en el pueblo. Si los tiempos entre llegadas de neverías son exponenciales, entonces

$$\pi_{25} = \frac{(30)^{25} e^{-30}}{25!} = .05$$

## PROBLEMAS

1. Cada semana, el Columbus Record Club atrae a 100 miembros nuevos. Los miembros permanecen en el club un promedio de 1 año (52 semanas). En promedio, cuántos miembros tiene el club?

2. El programa de doctorado de la universidad del estado admite un promedio de 25 estudiantes de doctorado cada año. Si un estudiante de doctorado pasa un promedio de 4 años de residencia en la universidad, ¿cuantos estudiantes de doctorados cabe esperar encontrar en ella?

3. En la actualidad hay 40 empresas de construcción especializadas en energía solar en el estado de Indiana. Cada año se abren un promedio de 20 constructoras especializadas en dicho estado, y una empresa promedio permanece en funciones durante 10 años. Si la tendencia actual continúa, ¿cuál es el número esperado de constructoras especializadas en energía solar que habrá en Indiana? Si el tiempo entre inauguraciones de constructoras se distribuye en forma exponencial, ¿cuál es la probabilidad de que, en el estado estable, haya más de 300 constructoras especializadas en el negocio? (Sugerencia: Para λ grande, la distribución de Poisson es aproximadamente una distribución normal.)

# 2-8 SISTEMA DE COLAS M/G/1/DG/∞/∞

En esta sección veremos un sistema de colas de un solo servidor en el que los tiempos de llegada son exponenciales, pero en el que la distribución del tiempo de servicio (S) no necesita ser exponencial. Sea $\lambda$ la frecuencia de llegadas, que suponemos se mide en llegadas por hora. También, definimos a $\frac{1}{\mu}$=E(S) y a al $\sigma^2$= var S.

En la notación de Kendall, a este sistema de colas se le represénla con M/G/1/DG/∞/∞. Un sistema de éstos no es un proceso de nacimiento y muerte, porque la probabilidad de que la terminación de un servicio suceda entre t y t + $\Delta t$ cuando el estado del sistema es j al tiempo t, depende del tiempo transcurrido desde la terminación del último servicio, porque tos tiempos de servicio ya no tienen la propiedad de amnesia. Así, no podemos formular la probabilidad de una terminación de servicio entre t y t + $\Delta t$ en la forma $\mu\Delta t$; por ello no es adecuado el modelo de nacimiento y muerte.

El cálculo de las probabilidades de estado estable para un sistema M/G/1/DG/∞/∞ de colas es asunto difícil. Como ya no valen las ecuaciones de estado estable para nacimiento y muerte, se debe aplicar un método distinto. Se usa la teoría de las cadenas de Markov para determinar a $\pi_j'$ la probabilidad que después de que el sistema haya trabajado durante largo tiempo se encuentren j clientes en el instante inmediato posterior a aquel en que ha terminado un servicio (véase Problema 4 al final de esta sección). Se puede demostrar que $\pi_j' = \pi_j$ donde $\pi_j$, es la fracción del tiempo después de que el sistema haya funcionado largo tiempo y que haya j clientes presentes.

Por fortuna, podemos calcular $L_q$, L, $L_s$, $W_q$, W y $W_s$ usando los resultados de Pollaczek y Khinchin. Ellos demostraron que para el sistema de colas M/G/1/DG/∞/∞,

(48) $$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}$$

donde $\rho = \frac{\lambda}{\mu}$.Como $W_s = \frac{1}{\mu}$ la Ecuación (30) significa que $L_s = \frac{\lambda}{\mu} = \rho$.Como L = $L_s$ + $L_q$ se llega a

(49) $L = L_q + \rho$

Entonces, las Ecuaciones (29) y (28) quieren decir que

(50) $$W_q = \frac{L_q}{\lambda}$$

(51) $$W = W_q + \frac{1}{\mu}$$

También se puede demostrar que $\pi_0$, la fracción del tiempo que el servidor está ocioso, es 1 – ρ. Véase el Prob. 2 al final de esta sección. Este resultado es semejante al del sistema M/G/1/DG/∞/∞

Para ilustrar el. uso de las Ecuaciones (48) a (51), veamos un sistema M/M/1/DG/∞/∞ con λ = 5 clientes/h y μ = 8 clientes/h. Según el estudio del modelo M/G/1/DG/∞/∞, sabemos que

$$L = \frac{\lambda}{\mu - \lambda} = \frac{5}{3} clientes$$

$$L_q = L - \rho = \frac{5}{3} - \frac{8}{5} = \frac{25}{24} clientes$$

$$W = \frac{L}{\lambda} = \frac{1}{3} horas$$

$$W_q = \frac{L_q}{\lambda} = \frac{5}{24} horas$$

De las Ecuaciones (3) y (4), sabemos que E(S) = 1/8 horas y que var S = 1/64 horas$^2$. Entonces la Ecuaciones (48) da como resultado

$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)} = \frac{5^2\,\frac{1}{64} + (\frac{5}{8})^2}{2(1-\frac{5}{8})} = \frac{25}{24} clientes$$

$$L = L_q + \rho = \frac{25}{24} + \frac{5}{8} = \frac{5}{3} clientes$$

$$W_q = \frac{L_q}{\lambda} = \frac{5}{24} horas$$

$$W = \frac{L}{\lambda} = \frac{1}{3} horas$$

Para demostrar cómo puede afectar significativamente la variancia del tiempo de servicio a la eficacia de un sistema de colas, veremos un sistema M/G/1/DG/∞/∞ con λ y μ idénticas al sistema M/M/l/DG/∞/∞ que acabamos de analizar. Para el modelo M/D/l/DG/∞/∞ E(S) = $\frac{1}{\mu}$ de hora y var S = 0. Entonces

$$L_q = \frac{(\frac{5}{8})^2}{2(1-\frac{5}{8})} = \frac{25}{48} clientes$$

$$W_q = \frac{L_q}{\lambda} = \frac{5}{48} horas$$

En este sistema M/D/l/DG/∞/∞, un cliente normal sólo estará la mitad del tiempo en la cola en comparación con un sistema M/M/l/DG/∞/∞, con rapideces idénticas de servicio y de llegada. Como muestra este ejemplo, aun que no se disminuyan los tiempos promedios de servicio, una disminución de la variabilidad de estos puede reducir mucho el tamaño de la cola y el tiempo de tránsito del cliente.

**PROBLEMAS**

1. A la ventanilla de un restaurante de servicio en su automóvil llega un promedio de 10 automóviles por hora. Si el tiempo de servicio de cada automóvil es de 2 min, ¿cuántos coches, en promedio, estarán esperando en la cola? Suponga tiempos exponenciales entre llegadas.

2. De acuerdo con el hecho de que de $L_s = \frac{\lambda}{\mu} = \rho$ demuestre que para un sistema de colas M/D/l/DG/∞/∞, la probabilidad de que la ventanilla este ocupada es $\frac{\lambda}{\mu} = \rho$.

3. Se tiene un sistema de colas M/D/l/DG/∞/∞, en el cual hay un promedio de 10 llegadas por hora. Suponga que el tiempo de trámite de cada cliente sigue una distribución de Erlang con parámetro de rapidez 1 cliente/min y con parámetro de forma 4.
(a) Calcule el número estimado de clientes que esperan en la cola.
(b) Calcule el tiempo esperado que un cliente está en el sistema.
(c) ¿Qué fracción del tiempo el servidor estará ocioso?

4. Se tiene un sistema de colas M/D/l/DG/∞/∞: en e1 que los tiempos entre llegadas se distribuyen exponencialmente con parámetro λ y los tiempos de servicio tienen una función de densidad de probabilidad s(t). Sea X, el número de clientes que hay un instante después s que sale el i-ésimo cliente.

(a) Explique porque $X_1, X_2, \cdots, X_k, \cdots,$ es una cadena de Markov.

(b) Explique porque $P_{ij} = P(X_{k+1} = j / X_k = i)$ es cero para j < i-1

(c) Explique porqué para i > 0, $P_{ii-1}$ =(probabilidad de que no haya llegadas durante un tiempo de servicio); $P_{ii}$=(probabilidad de que haya una llegada durante un tiempo de servicio) y para j ≥ i ; $P_{ij}$= (probabilidad de que j - i + 1 llegadas entren durante un tiempo de servicio).
(d) Explique porque, para j ≥ i-1 y para i > 0.

$$P_{ij} = \int_0^\infty \frac{s(x) e^{-\lambda x} (\lambda x)^{j-i+1} dx}{(j-i+1)!}$$

Sugerencia: La probabilidad de que un tiempo de servicio esté entre x y x + $\Delta x$ es s(x) $\Delta x$. Dado que el tiempo de servicio es igual a x, la probabilidad de que se tengan j - i + 1 llegadas durante el tiempo de servicio es

$$\frac{e^{-\lambda x} (\lambda x)^{j-i+1}}{(j-i+1)!}$$

# 2-9 MODELOS DE FUENTE FINITA: EL MODELO DE REPARACIÓN DE MÁQUINA

A excepción del modelo M/M/1/DG/c/∞, todos los modelos que hemos estudiado han tenido frecuencias de llegada independientes de estado del sistema. Como se dijo antes, hay dos donde quizá no sea válida la hipótesis de independencia de estado.

1. Si los clientes no desean formarse en colas largas, la frecuencia de llegadas puede ser una función decreciente del numero de personas presentes en el sistema de colas. Véanse los Prob. 4 y 5 al final de esta sección, donde se trata este caso.

2. Si las llegadas a un sistema se forman de una población pequeña, la frecuencia de llegadas puede depender mucho del estado del sistema. Por ejemplo, si un banco sólo tiene 10 cuentahabientes, entonces en un momento en que 10 estén en el banco, la frecuencia de llegada debe ser cero, mientras que si hay menos de 10 personas en el banco, la frecuencia de llegadas puede ser positiva.

Los modelos en los que las llegadas se loman de una población pequeña se llaman modelos de origen finito. Ahora analizamos un modelo importante de origen finito que se conoce como el de la reparación de máquinas, o de interferencia de máquinas.

En el problema de reparación de maquinas, el sistema consta de K máquinas y R técnicos. En cualquier momento, una máquina determinada está en buen estado o en mal estado. El intervalo durante el cual una máquina permanece a buen estado sigue una distribución exponencial con rapidez $\lambda$. Siempre que se descompone una máquina, se manda a un técnico de reparación que tiene R técnicos. El centro de reparación da servicio a las maquinas descompuestas como si llegaran conforme un sistema M/M/R/DG/∞/∞,

Así, si hay J $\leq$ R máquinas en mal estado, una máquina que apenas se acaba de descomponer será enviada de inmediato para que la reparen; si j > R máquinas están descompuestas, j - R máquinas estarán esperando en una cola para que un técnico se desocupe. Se supone que el tiempo que se necesita para completar la compostura de una máquina es exponencial con rapidez $\mu$, o sea, que el tiempo promedio de reparación es $\frac{1}{\mu}$. Una vez que se ha reparado una máquina regresa al buen estado y de nuevo es susceptible de descomponerse. El modelo de reparación de máquinas se puede considerar como proceso de nacimiento y muerte en el que el estado j en cualquier momento es el número de máquinas en mal estado. Con la notación de Kendall-Lee, el modelo que acabamos de explicar se puede expresar como sistema M/M/R/DG/K/K. La primera K indica que en cualquier momento puede haber no más de K clientes (o máquinas), y la segunda K quiere que las llegadas se toman de una fuente finita de tamaño K.

En la Tabla 8 se da la interpretación de cada estado para un modelo de reparación de máquinas con K = 5 y R = 2 (G = máquina en buen estado, y B = máquina descompuesta). Para determinar los parámetros de nacimiento y muerte para el modelo de reparación de máquinas (véase Fig. 13), nótese que un nacimiento le corresponde a una máquina que se descompone, y una muerte es una máquina que acaba de llegar ya reparada. Para calcular la frecuencia de natalidad en el estado j,

debemos determinar la rapidez a la que se descomponen las máquinas cuando el estado del sistema es j. Cuando es asó, hay K - j máquinas en buen estado. Como cada máquina se descompone con una frecuencia o rapidez $\lambda$, la frecuencia total a la que suceden las descomposturas cuando el estado j es

$$\lambda_j = \underbrace{\lambda + \cdots + \lambda}_{K-j \ \ veces} = (K-j)\lambda$$

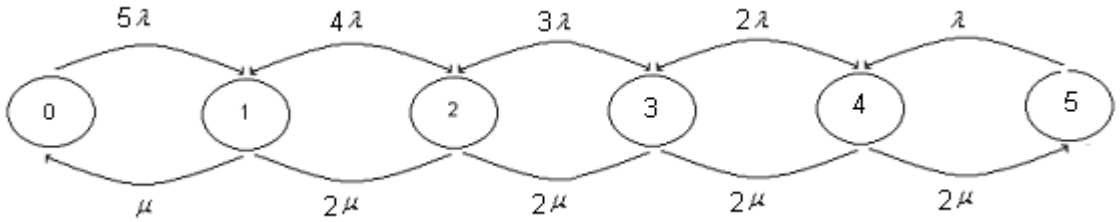
Para calcular la frecuencia de muertes para el modelo de reparación de máquinas procederemos como lo hicimos cuando estudiamos el modelo de colas M/M/s/DG/∞/∞,. Cuando el estado es j entonces estarán ocupados min(j,R) técnicos. Como cada técnico ocupado termina el trabajo con una rapidez $\mu$, la tasa de mortalidad $\mu_j$ está dada por

Tabla 8 Estados posibles en un problema de reparación de máquinas cuando K=5 y R=2

| ESTADO | NÚMERO DE ÁQUINAS EN BUEN ESTADO | COLA DE EPARACIÓN | NÚMERO DE TÉCNICOS OCUPADOS |
|---|---|---|---|
| 0 | G G G G G | | 0 |
| 1 | G G G G | | 1 |
| 2 | G G G | | 2 |
| 3 | G G | B | 2 |
| 4 | G | B B | 2 |
| 5 | | B B B | 2 |

Figura 13

Diagrama de rapidez para un sistema de colas M/M/R/DG/K/K,. cuando R = 2 y K = 5




Estado es el número de máquinas en malas condiciones.

$$\mu_j = j\mu \quad (j = 0,1,\cdots,R)$$
$$\mu_j = R\mu \quad (j = R+1,\cdots,K)$$

Si definimos que $\rho = \frac{\lambda}{\mu}$ aplicando las Ecuaciones (16) a (18) se obtiene la siguiente distribución de probabilidades de estado estable

$$(52)\quad \pi_j = \begin{cases} \binom{K}{j}\rho^j \pi_0 & (j = 0,1,\cdots,R) \\ \dfrac{\binom{K}{j}\rho^j j!\pi_0}{R!R^{j-R}} \end{cases}$$

Para usar la Ecuación (52) comenzamos calculando $\pi_0$ partiendo del hecho de que $\pi_0 + \pi_2 + \cdots + \pi_K = 1$. Mediante las probabilidades de estado estable de la Ecuación (52), podemos calcular las siguientes cantidades de interés:

L = número esperado de máquinas descompuestas
$L_q$ = número esperado de máquinas que esperan servicio
W = tiempo promedio que una máquina está descompuesta
$W_q$ = tiempo promedio que una máquina espera servicio

Desafortunadamente, no hay fórmulas sencillas para L, $L_q$, W y $W_q$. Lo mejor que podemos hacer es expresar estas cantidades en términos de las $\pi_j$ s:

$$(53)\quad L = \sum_{j=1}^{K} j\pi_j$$

$$(54)\quad L_q = \sum_{j=R}^{K} (j-R)\pi_j$$

Podemos emplear ahora las Ecuaciones (28) y (29) para calcular W y $W_q$. Como la frecuencia de llegadas depende del estado, el número promedio de llegadas por unidad de tiempo está dado por $\bar{\lambda}$. donde

$$(55)\quad \bar{\lambda} = \sum_{j=0}^{K} \lambda_j \pi_j = \sum_{j=0}^{K} \lambda(K-j)\pi_j = \lambda(K-L)$$

Si se aplica la Ecuación (28) a las máquinas que se reparan y a las que esperan su turno, obtenemos

$$(56)\quad W = \frac{L}{\bar{\lambda}}$$

Aplicando la Ecuación (29) a las máquinas que esperan campos fuera, obtenemos

$$(57)\quad W_q = \frac{L_q}{\bar{\lambda}}$$

El ejemplo que sigue ilustra el empleo de las fórmulas anteriores.

EJEMPLO 9 La policía de Gotham City tiene 5 patrullas. Una patrulla se descompone y debe repararse una vez cada 30 días. La policía tiene dos mecánicos, y cada uno de ellos tarda un promedio de 3 días en reparar un autopatrulla. Los tiempos entre descomposturas y los de reparación son exponenciales.

1. Calcule el número promedio de patrullas en buen estado.

2. Calcule el tiempo muerto promedio que pasa una patrulla en reparaciones.
3. Calcule la fracción del tiempo que está ocioso determinado mecánico.

Solución Se trata de un problema de reparación de máquinas para K = 5, R = 2, $\lambda$ = 1/30 patrulla por día, y $\mu$ = 1/3 auto por día. Entonces,

$$\rho = \frac{\frac{1}{30}}{\frac{1}{3}} = \frac{1}{10}$$

de la ecuación (52)

$$\pi_1 = \binom{5}{1}\left(\frac{1}{10}\right)\pi_0 = .5\pi_0$$

$$\pi_2 = \binom{5}{2}\left(\frac{1}{10}\right)^2\pi_0 = .1\pi_0$$

(58)
$$\pi_3 = \binom{5}{3}\left(\frac{1}{10}\right)^3 3!\pi_0/(2!*2) = .015\pi_0$$

$$\pi_4 = \binom{5}{4}\left(\frac{1}{10}\right)^4 4!\pi_0/(2!*2^2) = .0015\pi_0$$

$$\pi_5 = \binom{5}{5}\left(\frac{1}{10}\right)^5 5!\pi_0/(2!*2^3) = .000075\pi_0$$

Entonces $\pi_0$ (l + .5 + .1 + .015 + .0015 + .000075) =l, o sea $\pi_0$ = .619. Entonces, las Ecuación (58)da como resultado $\pi_1$= .310, $\pi_2$ = .062, $\pi_3$ = .009, $\pi_4$ = .001, y $\pi_5$ = 0.

1. El número esperado de patrullas en buen estado es K - L, y es

$$K - L = 5 - \sum_{j=1}^{5} j\pi_j = 5 - [0(.6t9) + 1(.310) + 2(.062) + 3(.009) + 4(.001) + (0)]$$

$$= 5 - .465 = 4.535 \text{ autos en buen estado}$$

2. Buscamos W= = $\frac{L}{\bar{\lambda}}$ Según la Ecuación (55),

$$\bar{\lambda} = \sum_{j=0}^{K} \lambda_j \pi_j = \sum_{j=0}^{5} \lambda(5-j)\pi_j = \frac{1}{30}\ [5(.6t9) + 4(.310) + 3(.062) + 2(.009) + 1(.001) + (0)]$$

= 0.151 autos por día

Como L = 0.465 patrulla, vemos que W = .0465/.151= 3.08 días.

3. La fracción de tiempo que esta desocupado determinado mecánico es $\pi_0 + \frac{1}{2}\pi_1$=.619+.5(.310)=.774.

Si hubiera tres mecánicos, la fracción de tiempo que determinado mecánico se hubiera encontrado desocupado sería $\pi_0 + \frac{2}{3}\pi_1 + \frac{1}{3}\pi_2$ y si el equipo es de R mecánicos, la probabilidad de que uno de ellos se encuentre desocupado será

$$\pi_0 + \frac{(R-1)}{R}\pi_1 + \frac{(R-2)}{R}\pi_2 + \cdots + \frac{1}{R}\pi_{R-1}$$

## PROBLEMAS

1. Una lavandería tiene 5 lavadoras. Una máquina normal se descompone una vez cada 5 días. Un técnico puede reparar una máquina en un promedio de 2.5 días. En la actualidad, hay tres técnicos en servicio. El dueño de la lavandería tiene la opción de cambiarlos por un supertécnico que puede reparar una máquina en un promedio de 5/6 de día. El sueldo del supertécnico es igual al de los tres técnicos juntos. Los tiempos entre descomposturas y los de servicio son exponenciales ¿Debe cambiar la lavandería los tres técnicos por el supertécnico?

2. Mi perra acaba de tener tres cachorros. Los perritos son vivarachos y entran y salen brincando de su caja. Un cachorrito pasa un promedio de 10 min, distribuidos exponencialmente, en la caja y salta hacia afuera. Una vez fuera, pasa un promedio de 15 min, distribuidos exponencialmente, para saltar y entrar de nuevo.

(a) En un momento dado, ¿cuál es la probabilidad de que haya más cachorritos fuera de la caja que dentro de ella?

(b) En promedio, ¿cuantos cachorritos habrá dentro de la caja?

3. Gotham City tiene 10 000 arbotantes. Se ha determinado que en un momento dado, hay un promedio de 1 000 luminarias quemadas. Una luminaria se quema después de transcurrir un promedio de 100 días de uso. La ciudad ha contratado a la empresa Mafia Inc., para cambiar las luminarias quemadas. El contrato con Mafia dice que se deben tardar un promedio de 7 días en cambiar una luminaria quemada. ¿Cree el lector que Mafia Inc. se aprovecha en este contrato?

4. Este problema es un ejemplo de la denegación. La nevería Oryo Cookie Ice Cream tiene tres competidores. Como a las personas no les gusta esperar mucho su pedido, la frecuencia de llegadas a Oryo Cookie Ice Cream depende del número de clientes adentro. En forma más específica, cuando hay $j \leq 4$ clientes dentro de la nevería, llegan clientes con una frecuencia de (20 –5) por hora. Si hay más de 4 personas, la frecuencia de llegada es cero. Por cada cliente, las ganancias menos los costos de materias primas son 50? de dólar. A cada mesero se le paga 3 dólares/h. Un mesero puede atender a un promedio de 10 clientes por hora. Para maximizar las utilidades esperadas (ingresos menos costos de materiales y mano de obra), ¿cuántos meseros debe contratar Oryo? Suponga que los tiempos entre llegadas y de servicio son exponenciales.

5. Suponga que los tiempos entre llegadas a un sistema con ventanilla única son exponenciales, pero que cuando hay n clientes, hay una probabilidad de n /n+1 de que una llegada sea denegada y deje el sistema antes de entrar al servicio. También suponga tiempos de servicio exponenciales.

(a) Determine la distribución de probabilidad del número de personas que hay en el estado estable.
(b) Calcule el número esperado de personas que hay en el estado estable. (Su gerencia: puede ayudar el hecho de que

$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$ tal vez sea usado.

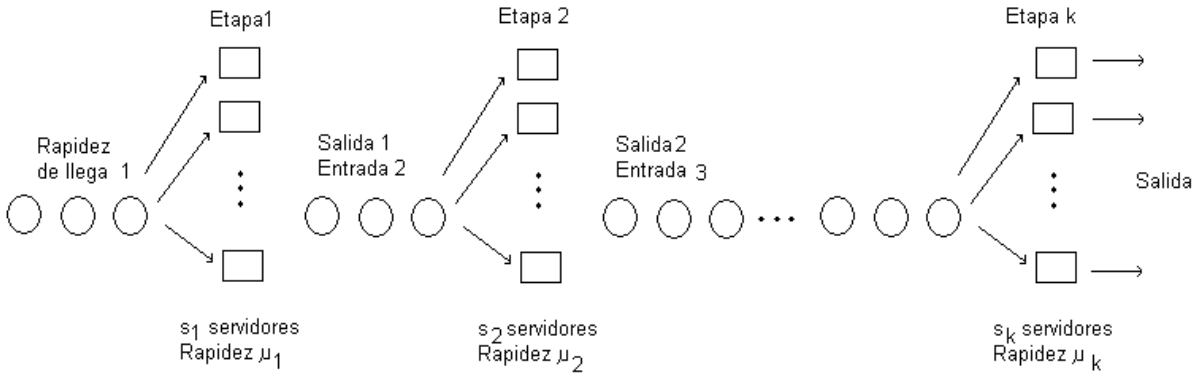
6. Bectol, Inc. construye una presa. Se necesita un total de 10000 pies$^3$ de tierra para formar la cortina. Para extraer esa tierra se usa una pala mecánica. A continuación, la tierra se transporta mediante camiones de volteo al lugar de la cortina. Tan sólo se cuenta con una pala mecánica, y su renta es 100 dólares/h. Bectol puede rentar, a 40 dólares/h, tantos camiones de volteo como desee. Cada camión puede manejar 1 000 pies de tierra. Para cargarlo, la pala mecánica tarda un promedio de 12 min para ir a la presa y regresar a la pala mecánica, el camión de volteo tarda un promedio de 5 minutos. Plantee las hipótesis adecuadas acerca de exponencialidad y determine cómo puede Bectol minimizar el cosió total esperado del movimiento de la tierra que se necesita para formar la cortina. (Sugerencia: ¡En algún lado debe haber un problema de reparación de máquinas!).

## 2-10 COLAS EXPONENCIALES EN SERIE Y REDES ABIERTAS DE COLAS

En los modelos de cola que hemos estudiado hasta aquí, todo el tiempo de servicio a un cliente se gasta con un solo servidor. En muchos casos, como en la producción de un artículo en una línea de montaje, los trámites del cliente no se terminan sino hasta que haya intervenido mas de un servidor. Véase, por ejemplo, la Fig. 14.

Al entrar al sistema de la Fig. 14, el cliente pasa por la etapa 1 de servicio, después de esperar en una cola si lodos los servidores de la etapa 1 están ocupados cuando llega. Después de terminar la etapa 1 de servicio, el cliente espera y pasa a la etapa 2 de servicio. Este proceso sigue hasta que el cliente termina la etapa k de servicio. Un sistema como el de la Fig. 14 se llama sistema de colas de k etapas en serie, o en tándem. Contamos con un notable teorema debido a Jackson (1957), que es el siguiente

Figura 14
Colas exponenciales en serie




TEOREMA 4 Si (1) los tiempos de llegada a un sistema de colas en serie son exponenciales con rapidez $\lambda$, (3) los tiempos de servicio para cada tramite en la etapa i son exponenciales, y (3) toda etapa tiene sala de espera de capacidad infinita, entonces los tiempos entre llegada para alcanzar cada etapa del sistema de colas son exponenciales con rapidez $\lambda$.

Para que este resultado sea válido, cada etapa debe tener capacidad suficiente para dar servicio a una corriente de llegadas que entre con rapidez $\lambda$; si no es así, el sistema "explotaría" en la etapa que tuviera capacidad insuficiente. De acuerdo con el análisis del sistema M/M/s/DG/$\infty$/$\infty$ de la Sección 2.6, vemos que cada etapa tendrá capacidad suficiente para manejar una corriente de llegadas de rapidez o frecuencia $\lambda$ si y sólo si $\lambda<s_j\mu_j$ cuando j = 1, 2, , . . , k. Si $\lambda<s_j\mu_j$ el resultado de Jackson quiere decir que la etapa j del sistema de la Fig. 14 se puede analizar como un sistema M/M/s/DG/$\infty$/$\infty$ con tiempos exponenciales entre llegadas que tienen la rapidez $\mu$ y tiempos exponenciales de servicio con un promedio $\frac{1}{\mu_j}$. En el ejemplo que sigue se apreciará la utilidad del Teorema de Jackson.

EJEMPLO 10 Las dos últimas cosas que se hacen en un automóvil para completar su ensamble son instalar el motor y poner los neumáticos. Llega un promedio de 54 automóviles/h que necesitan de esas dos tareas. Un trabajador instala el motor y puede atender un promedio de 60 automóviles/h. Después de instalar el motor, el automóvil pasa a la estación de los neumáticos y espera para que le pongan los neumáticos. En esa estación trabajan tres personas. Cada una trabaja en un automóvil a la vez y puede colocar neumáticos en un automóvil a un promedio de 3 minutos por cada auto. Los tiempos entre llegadas y los de servicio son ambos exponenciales.

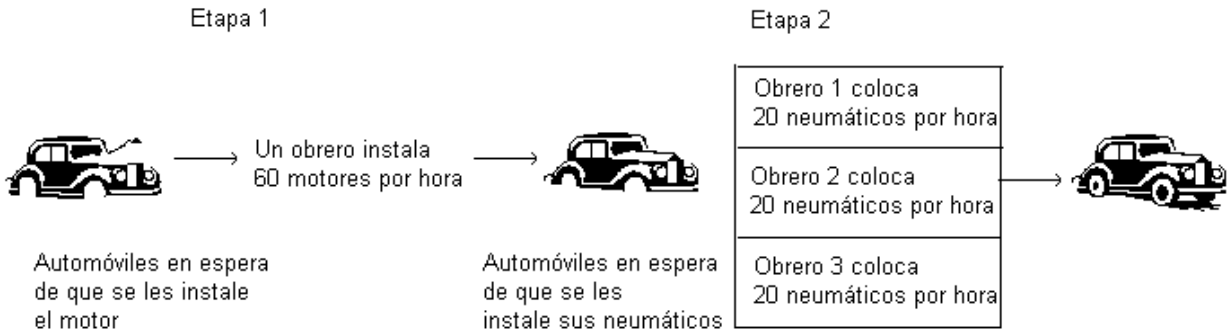
1. Calcule la longitud promedio de la cola en cada estación de trabajo.

2. Determine el tiempo total esperado que pasa un automóvil esperando su turno.

Solución Se trata de un sistema de cola en serie con $\lambda$ = 54 automóviles/h, $s_1$ = 1, $\mu_1$ = 60 automóviles/h, $s_2$ = 3 y $\mu_2$ = 20 automóviles/h (véase Fig. 15). Como $\lambda < \mu_1$ y $\lambda < 3\mu_2$ ninguna de las colas "explotará", y se puede aplicar el Teorema de Jackson. Para la etapa 1 (motor), $\rho$ = 54/60 = .90. Entonces, la Ecuación (27) da como resultado

$L_q$ (para el motor) = $\frac{\rho^2}{1-\rho} = \frac{.9^2}{1-.9} = 8.1$ automóviles

Figura 15
Sistema de colas en serie para el ejemplo de la fábrica de automóviles




Luego la Ec. (32) da como resultado

Wq (para el motor) = $\frac{L_q}{\lambda} = \frac{8.1}{54} = .15$ horas

Para la etapa 2 (neumáticos), p = 54/(3*20) = .90. La Tabla 6 indica que P(j≥3) = .83. Entonces, según la Ecuación (41),

$L_q$(para neumáticos) = $\frac{.83*.90}{1-.90} = 7.47$ automóviles.

Entonces

Wq (para neumáticos ) $\frac{L_q}{\lambda} = \frac{7.47}{54} = .138$ horas.

Así, el tiempo total esperado para que a un automóvil le instalen su motor y sus neumáticos es 0.15 + 0.138 = 0.288 horas,

REDES ABIERTAS DE COLAS

A continuación describiremos las redes abiertas de colas, que son una generalización de colas en serie- Como en la Fig. 14, suponemos que la estación y consiste en $s_j$ servidores exponenciales, cada uno trabajando a una velocidad $\mu_j$. Se supone que los clientes llegan a la estación y procedentes del exterior del

sistema con una frecuencia o rápido $r_j$. Se supone que estos tiempos entre llegadas están distribuidos exponencialmente. Una vez terminado el servicio en la estación i, un cliente se forma en la cola de la estación j con probabilidad $P_{ij}$; y termina sus trámites, o servicio, con probabilidad

$$1-\sum_{j=1}^{k} P_{ij}$$

Definimos a $\lambda_j$, como la frecuencia o rapidez con la que llegan los clientes a la estación j, incluyendo las llegadas a la estación y desde el exterior del sistema y de otras estaciones $\lambda_1, \lambda_2, \ldots, \lambda_k$ /^ se pueden calcular resolviendo el siguiente sistema de ecuaciones lineales:

$$\lambda_j = r_j + \sum_{i=1}^{k} P_{ij}\lambda_i \quad (j = 1,.\cdots,k)$$

Esto es consecuencia de que una fracción $P_{ij}$ de las $\lambda_j$ llegadas a la estación i pasará a continuación a la estación j. Supongamos que para todas las estaciones es válido que $\lambda_j < s_j\mu_j$ ,. Entonces se puede demostrar que la distribución de probabilidad del numero de clientes presentes en la estación j y el número esperado de clientes presentes en la estación j se pueden calcular tratando esa estación como un sistema $M/M/s_j/DG/\infty/\infty$ con frecuencia de llegada $\lambda_j$ y rapidez de servicio $\mu_j$. Si $\lambda_j < s_j\mu_j$ no se cumple para alguna j entonces no existe distribución de estado estable para los clientes.

Para calcular L, el número esperado de clientes en el sistema de cola, tan sólo sumamos el número esperado de clientes que hay en cada estación. Para calcular W, el tiempo promedio que pasa un cliente en el sistema, sólo aplicamos la fórmula $L = \lambda W$ al sistema completo. En este caso, $\lambda = r_1 + r_2 + \bullet \bullet \bullet + r_k$, porque esto representa el número promedio de clientes por unidad de tiempo que llega al sistema. El ejemplo que sigue ilustra el análisis de redes abiertas de colas.

EJEMPLO 11 Se tienen dos servidores. Del exterior llega un promedio de 8 clientes/h al servidor 1 y un promedio de 17 clientes, también del exterior, al servidor 2. Los tiempos entre llegadas son exponenciales. El servidor 1 puede atender a una rapidez exponencial a 20 clientes/h, y el 2, a 30 clientes/h, también con rapidez exponencial.

Después de terminar sus trámites con el servidor 1, la mitad de los clientes sale del sistema y la mitad va al servidor 2. Al terminar su servicio en el servidor 2 ¾ de los clientes salen y ¼ regresa al servidor 1.

1. ¿Qué fracción de tiempo está desocupado el servidor 1?
2. Calcule el número esperado de clientes en cada servidor.
3. Calcule el tiempo promedio que pasa un cliente en el sistema.
4. ¿Cómo se modificarían los incisos (1) a (3) si el servidor 2 sólo pudiera atender a un promedio de 20 cientes/h?

Solución Tenemos una red abierta de colas con r1 = 8 clientes/h y r2 = 17 clientes/h.
También, se tiene $P_{12} = .5$; $P_{21} = .25$; $P_{11} = P_{22} = 0$. Podemos calcular $\lambda_1$ y $\lambda_2$ resolviendo sistema de ecuaciones $\lambda_1 = 8 + .25\lambda_2$ y $\lambda_1 = 17 + .5\lambda_1$. Esto da como resultado $\lambda_1 = 14$ clientes/h y $\lambda_2 = 24$ clientes/h.

1. El servidor 1 se puede considerar un sistema M/M/1/DG/∞/∞ con λ= 14 clientes/h y μ=20 clientes/h. Entonces $\pi_0$ = 1 - ρ = 1 - .7 = .3. Así, el servidor 1 está desocupado 30% del tiempo.

2. De la Ecuación (26) vemos que L= $=\frac{\rho}{(1-\rho)}=\frac{.7}{1-.7}=\frac{14}{6}$ en el servidor 1 y L = 4 en el servidor 2, Así, habrá en el sistema un promedio de $4+\frac{14}{6}=\frac{19}{3}$ clientes.

3. W = L/λ donde λ = 8 + 17 = 25 clientes/h. Entonces W = $\frac{\frac{19}{3}}{25}=\frac{19}{75}$ horas

4. En este caso $\lambda_2 > s_2\mu_2$ = 20 entonces no existe estado estable.

**PROBLEMAS**

1. Una sucursal del Seguro Social estudia las dos alternativas siguientes para procesar las solicitudes de tarjeta de seguridad:

Opción 1 Tres empleados procesan las solicitudes en paralelo que les llegan de una cola única. Cada empleado llena el machote para la solicitud en presencia del solicitante. El tiempo de proceso es exponencial con un promedio de 15 minutos. Los tiempos entre llegadas son exponenciales.

Opción 2 Cada solicítame llena primero un machote de solicitud sin el auxilio del empleado. El tiempo para llenarlo está distribuido en forma exponencial con un promedio de 65 minutos. Cuando el solicitante ha llenado su forma, se forma en una cola única para esperar a que uno de los tres oficinistas le revise el machote. Un oficinista gasta un promedio de 4 minutos, distribuidos exponencialmente, en revisar la solicitud.

El tiempo entre llegadas de los solicitantes es exponencial. y llega un promedio de 4,8 solicitantes/h. ¿Que opción hará que salgan los solicitantes más rápidamente de la oficina?

2. Se tiene una línea de ensamble de automotores en la que a cada unidad se le hacen dos tipos de servicios: pintura y, después instalación del motor. Cada hora, pasa un promedio de 22.4 chasises sin pintar a la línea. En promedio, se necesitan 2.4 min para pintar un automóvil y un promedio de 3.75 min para instalar el motor. La línea de ensamble tiene un pintor y dos obreros que instalan el motor. Suponga que los tiempos entre llegadas y los tiempos de servicio son exponenciales.

(a) En promedio, ¿cuántos automóviles ya pintados sin motor instalado estarán en la línea?

(b) En promedio, ¿cuánto tiempo esperará un automóvil pintado para que se inicie la instalación de su motor?

3. Se tienen los dos sistemas de colas siguientes:

Sistema 1 Llega un promedio de 40 clientes cada hora los tiempos entre llegadas son exponenciales. Los clientes deben terminar dos tipos de trámites para poder

dejar el sistema. El primer servidor tarda un promedio de 30 segundos, distribuidos exponencialmente, para efectuar el servicio del tipo 1. Después de esperar en una cola, cada cliente obtiene un servicio tipo 2, distribuido exponencialmente y con promedio de 1 minuto, con un solo servidor. Después de completar su tramite tipo 2, el cliente deja el sistema.

Sistema 2 El proceso de llegada al sistema 2 es idéntico al del sistema 1. En el sistema 2, un cliente debe terminar solo un tipo de trámite. El tiempo de servicio es 1.5 min en promedio y está distribuido exponencialmente. Hay dos servidores.
¿En cuál sistema un cliente representativo pasa menos tiempo?

4. Llegan un promedio de 120 estudiantes cada hora a la oficina de inscripciones de un colegio; los tiempos entre llegadas son exponenciales; para terminar su inscripción, una persona debe pasar por tres ventanillas. Cada ventanilla tiene un solo servidor. Los tiempos de servicio en cada ventanilla son exponenciales y sus duraciones promedio son 20 segundos en la ventanilla 1, 15 segundos en la ventanilla 2 y 12 segundos en la ventanilla 3. En promedio, ¿cuántos estudiantes habrá en laoficina de inscripciones para tramitar sus cursos?

5. A un taller llega un promedio de 10 trabajos/h. Los tiempos entre llegadas de los trabajos están distribuidos exponencialmente. Para hacer un trabajo se necesita un promedio de 10/3 minutos, exponencialmente distribuidos. Desafortunadamente, 1/3 de los trabajos terminados se deben volver a procesar. Así, con probabilidad 1/3 un trabajo terminado debe esperar en una cola para que se vuelva a procesar. En el estado estable, ¿cuántos trabajos puede uno esperar encontrar en el taller? ¿Cuál sería la respuesta si se necesitara un promedio de 5 min para terminar un trabajo?

6. Se tiene un sistema de colas que consiste en tres estaciones en serie. Cada estación tiene un servidor único, que puede procesar un promedio de 20 trabajos/h. Los tiempos de proceso en cada estación son exponenciales. Llega un promedio de diez trabajos por hora a la estación 1. Los tiempos entre llegadas son exponenciales. Cuando un trabajo termina su servicio en la estación 2. hay una probabilidad de .1 de que regrese a la estación 1 y una probabilidad .9 de que continúe a la estación 3. Cuando un trabajo termina su servicio en la estación 3, existen .20 de posibilidades que regrese a la estación 2 y .8O posibilidades de que deje el sistema. Todos los trabajos que terminan su servicio en la estación 1 pasan de inmediato a la estación 2.

(a) Calcule la fracción del tiempo que cada servidor está ocupado.

(h) Determine el número esperado de trábalos en el sistema.

(c) Estime el tiempo promedio que pasa un trabajo en el sistema.

# 2-11 SISTEMA M/G/s/DG/s/∞

# (SISTEMA DEPURADO, SD)

En muchos sistemas de colas, el sistema pierde para fines prácticos una llegada porque encuentra a todos los servidores ocupados. Por ejemplo, una persona que llama a una aerolínea para reservar boleto y oye señal de teléfono ocupado llamará probablemente a otra aerolínea. O bien, suponga el lector que alguien llama para dar una alarma de incendio y que no se dispone de bombas; en ese caso el incendio saldrá de control. Así, en cierto sentido, la solicitud de una bomba contra incendio que llega cuando no hay bombas disponibles se puede considerar que la pierde el sistema. Si las llegadas que encuentran a todos los servidores ocupados dejan al sistema depurado, entonces, el sistema se libera de los clientes que no tuvieron atención. Suponiendo que los tiempos entre llegada son exponenciales, un sistema de estos se puede modelar como un sistema de cola tipo M/G/s/DG/s/∞.

Para este sislema M/G/s/DG/s/∞, L, W, $L_q$ y $W_q$ tienen interés limitado. Por ejemplo, ya que nunca puede haber cola, Lq = Wq = 0. Si hacemos que $\frac{1}{\mu}$ sea el tiempo promedio de servicio y $\lambda$ la frecuencia de llegadas, entonces W = $W_s$ = $\frac{1}{\mu}$.

En la mayor parte de los sistemas depurados, el interés practico se centra en la fracción de todas las llegadas que no entran. Como las llegadas no entran sólo cuando hay presentes s clientes, una fracción $s\pi$ no entrará. Por lo tanto, el sistema perderá un promedio de $\lambda\pi_s$ llegadas por unidad de tiempo. Como la frecuencia promedio de llegadas por unidad de tiempo que entran al sistema es $\lambda(1-\pi_s)$,
llegamos a la conclusión de que

$$L = L_s = \frac{\lambda(1-\pi_s)}{\mu}$$

Figura 16. Probabilidades de Perdida de un sistema M/G/s/DG/s/∞,

Se puede demostrar que para un sistema M/G/s/DG/s/∞, $\pi_s$ depende de la distribución del tiempo de servicio tan sólo a través de su promedio $\frac{1}{\mu}$. Este hecho se conoce como formula de Erlang de pérdida. En otras palabras, todo sistema M/G/s/DG/s/∞, con una frecuencia de llegada λ y un tiempo promedio de servicio de tendrá el mismo valor de $\pi_s$. Si definimos a ρ como igual a $\frac{\lambda}{\mu}$. entonces, para determinado valor de s, se puede calcular el valor de $\pi_s$, con la Fig. 16. Tan sólo se lee el

valor de $\rho$ en el eje x. Así el valor de y de la curva de s servidores, que corresponde a $\rho$, será igual a $\pi_s$. El siguiente ejemplo ilustra el uso de la Fig. 16.

EJEMPLO 12 En el hospital de Gotham City se recibe un promedio de 20 solicitudes de ambulancia por hora. Una ambulancia necesita un promedio de 20 min para recoger un paciente y llevarlo al hospital. La ambulancia queda disponible entonces para recoger otro paciente. ¿Cuantas ambulancias debe tener el hospital para asegurar que cuando más haya el 1% de probabilidades de no poder atender de inmediato una solicitud de ambulancias? Suponga que los tiempos entre solicitudes están distribuidos exponencial mente.

Solución

Sabemos que $\lambda$ = 20 llamadas/h, y $\frac{1}{\mu}=\frac{1}{3}$ de hora. Así, $\rho$= 20/3 = 6.67. buscamos el valor mínimo de s para el cual $\pi_s$ sea .01 o menor. En la figura 16, vemos que para s = 13, $\pi_s$ = .011, y para s = 14, $\pi_s$ = .005. El hospital necesita 14 ambulancias para cumplir con las normas deseadas de servicio.

PROBLEMAS

# 2-14 MODELOS DE COLAS PRIORITARIOS

Hay muchos casos en los que a los clientes no se les atiende con el sistema "el primero que llega es el primero en ser servido" (FIFO). También analizaremos, en el servicio en orden aleatorio (SEOA) y el sistema último en llegar primero en ser servido (LIFO). Sean $W_{FIFO}$, $W_{SEOA}$ y $W_{LIFO}$ las variables aleatorias que representan al tiempo de espera de un cliente con las disciplinas FIFO, SEOA y LIFO respectivamente. Se puede demostrar que

$$E(W_{FIFO}) = E(W_{SEOA}) = E(W_{LIFO})$$

Así, el tiempo promedio en estado estable que pasa un cliente en el sistema no depende de cuál de las tres disciplinas se escoja. También se puede demostrar que

$$\text{(59) } Var(W_{FIFO}) < Var(W_{SEOA}) < Var(W_{LIFO})$$

Como en el caso general se relaciona una variancia grande con una variable aleatoria que tiene bastante probabilidad de asumir valores extremos, la Ec. (59) indica que son más probables los tiempos de espera relativamente grandes en la disciplina LIFO, y que se presentan con menos probabilidad con la disciplina FIFO. Esto es razonable, ya que en un sistema LIFO, un cliente puede tener suerte y entrar de inmediato al servicio, pero también se puede quedar al último en una cola larga. Sin embargo, en la disciplina FIFO el cliente no puede quedar atorado al final de una cola larga y, por lo tanto no es probable un tiempo de espera muy largo.

En muchas organizaciones, el orden en que se atiende a los clientes depende del "tipo" de cliente. Por ejemplo, las salas de urgencia de los hospitales dan servicio en general a pacientes muy graves antes de atender a pacientes que no lo estén. Asimismo, muchos sistemas de cómputo procesan los trabajos más largos hasta que se hayan terminado los más cortos de la cola. Los modelos en los que un tipo de cliente determina el orden en el que se atiende a las personas se llaman modelos de colas prioritarios.

El siguiente caso abarca muchos modelos de colas prioritarios, incluso todos los que se analizan en esta sección. Se supone que hay « tipos de clientes, identificados como tipo 1, tipo 2, . . ., tipo n. Los tiempos entre llegadas de clientes tipo i están distribuidos exponencialmente con frecuencia $\lambda_i$. Se supone que los tiempos entre llegadas de distintos tipos de clientes son independientes entre sí. El tiempo de servicio para un cliente del tipo i se representa mediante una variable aleatoria $S_i$, y no es necesariamente exponencial. Asimismo, se supone que los tipos de número más bajo tienen prioridad sobre los tipos de número más alto.

**MODELOS SIN PRIORIDAD ADQUIRIDA**

Comenzamos por considerar los modelos sin prioridad adquirida. En estos modelos sin privilegios no se puede interrumpir el servicio a un cliente. Después de terminar cada servicio, se escoge el siguiente cliente al que dará atención dando prioridad a los clientes de número de tipo más bajos, y se aplica la disciplina FIFO en cada categoría. Por ejemplo, si n = 3 y hay en la cola tres clientes tipo 2 y cuatro tipo 3, el siguiente cliente que entra sería el primero de tipo 2 que haya llegado.

En la notación Kendall-Lee, un modelo sin prioridad adquirida se representa poniendo en la cuarta característica "NPRP", iniciales de Non Preemptive Priority, para indicar múltiples tipos de clientes, se pone como subíndice i a las primeras dos características. Así, Mi/Gi/... represénla un caso en el que los tiempos entre llegadas para el i-ésimo tipo de cliente son exponenciales y en que los tiempos de servicio para el í-ésimo tipo de cliente tienen una distribución general. En lo que sigue haremos que

$W_{qk}$= tiempo esperado en estado estable que pasa un cliente tipo k en la cola
$W_k$= tiempo esperado en estado estable que pasa un cliente tipo k en el sistema
$L_{qk}$= número esperado en estado estable de clientes tipo k esperando en la cola
$L_k$= número esperado en estado estable de clientes tipo k en el sistema

MODELO $M_i/G_i/1/NPRP/\infty/\infty$

Los primeros resultados conciernen al sistema $M_i/G_i/1/NPRP/\infty/\infty$ con un solo servidor y sin prioridad adquirida. Definimos $\rho_i = \frac{\lambda_i}{\mu_i}$, $a_0 = 0$ y $a_k = \sum_{i=1}^{k} \rho_i$. Supondremos que

$$\sum_{i=1}^{n} \rho_i < 1$$

Entonces

$$W_{qk} = \frac{\sum_{k=1}^{n} \lambda_k E(S_k^2)/2}{(1-a_{k-1})(1-a_k)}$$

(60)
$$L_{qk} = \lambda_k W_{qk}$$
$$W_k = W_{qk} + \frac{1}{\mu_k}$$
$$L_k = \lambda_k W_k$$

El ejemplo siguiente muestra el uso de las Ec. (60)

EJEMPLO 16 Una instalación de copiado da prioridad a los trabajos cortos, sobre los trabajos largos. Los tiempos entre llegada para cada tipo de trabajo son exponenciales, y cada hora llegan 12 trabajos cortos y 6 largos. Sea trabajo tipo 1 = trabajo corto y trabajo tipo 2 = trabajo largo. Entonces los datos son

$E(S_1)$ = 2 min $E(S_1^2)$ = 6 min$^2$ = 1/600 horas$^2$

E(S2) = 4 min E{$S_2^2$) = 18 min$^2$ = 1/200 horas$^2$

Calcule el tiempo promedio que pasa un trabajo de cada tipo en la instalación de

copiado.

Solución
Sabemos que $\lambda_1$ = 12 trabajos/h, $\lambda_2$ = 6 trabajos/h, $\mu_1$ = 30 trabajos/h y $\mu_2$= 15 trabajos/h. Entonces

$$\rho_1 = \tfrac{12}{30} = .4$$
$$\rho_2 = \tfrac{6}{15} = .4$$

y como $\rho_1 + \rho_2 < 1$ existirá un estado estable. Ahora bien, $a_0 = 0, a_1 = .4, a_2 = .8$ Las Ecuaciones en (60) nos dan

$$W_{q1} = \frac{12*(1/600)/2 + 6*(1/200)/2}{(1-0)(1-.4)} = .042 horas$$
$$W_{q1} = \frac{12*(1/600)/2 + 6*(1/200)/2}{(1-.4)(1-.8)} = .208 horas$$
$$W_1 = W_{q1} + \frac{1}{\mu_1} = .042 + .033 = .075 horas$$
$$W_2 = W_{q2} + \frac{1}{\mu_2} = .0208 + .067 = .275 horas$$

Entonces, como esperábamos, los trabajos largos pasan mucho más tiempo en la instalación de copiado que los trabajos cortos.

**MODELO $M_i/G_i/NPRP/\infty/\infty$ CON COSTOS DE ESPERA QUE DEPENDEN DE LOS CLIENTES**

Considere un sistema sin prioridad adquirida con un solo servidor en el cual se carga un costo $c_k$ por cada unidad de tiempo que pasa un cliente tipo k en el sistema. Sí deseamos minimizar el costo esperado en que se incurre por unidad de tiempo, en el estado estable, ¿que prioridad se debe dar a los tipos de clientes? Suponga que los n tipos de cliente se numeran de tal modo que

(61) $$c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_n\mu_n$$

Entonces el costo esperado se minimiza dando la mayor prioridad a los clientes tipo 1, la segunda prioridad a los del tipo 2, etc. y la menor prioridad a los clientes del tipo n. Para ver por qué esta orden de prioridades es razonable, observe que cuando se atiende a un cliente tipo k, el costo deja al sistema con una rapidez $c_k\mu_k$. Así, se puede minimizar el costo dando la mayor prioridad a los tipos de cliente con los valores mayores de $c_k\mu_k$ .

Como caso especial de este resultado, supongamos que deseamos minimizar, el número esperado de trabajos en el sistema. Sean $c_1 = c_2 = \cdots = c_n = 1$. Entonces, en cualquier momento, el costo por unidad de tiempo es igual al numero de clientes en el sistema. Entonces, el costo esperado por unidad de tiempo será igual a L. Entonces la Ecuación (61) se transforma en

$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ o bien $\frac{1}{\mu_1} \leq \frac{1}{\mu_2} \leq \cdots \leq \frac{1}{\mu_n}$

De este modo podemos llegar a la conclusión de que el número esperado de trabajos en el sistema se reducirá al mínimo si se da la mayor prioridad a los tipos de clientes con tiempo promedio de servicio más corto. Esta disciplina de prioridad se conoce como disciplina de tiempo mínimo de proceso (TMP)

**MODELO $M_i$/M/s/NPRP/∞/∞**

Para obtener resultados analíticos manejables para sistemas de prioridad con servidores múltiples debemos suponer que cada cliente tiene tiempos de servicio con distribución exponencial con un promedio $\frac{1}{\mu}$, y que los clientes tipo ¿tienen tiempos entre llegadas que se distribuyen exponencialmente con frecuencia $\lambda_i$. A este sistema con s servidores se le represénla con la notación $M_i$/M/s/NPRP/∞/∞. Para este modelo,

(62) $$W_{qk} = \frac{P(j \geq s)}{s\mu(1-a_{k-1})(1-a_k)}$$

En esta ecuación,

$$a_k = \sum_{i=1}^{k} \frac{\lambda_i}{s\mu} (k \geq 1)$$

$a_0$ = O, y P(j ≥ s) se obtiene de la Tabla 6 para un sistema con s servidores que tiene

$$\rho = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_n}{s\mu}$$

El Ejemplo 17 ilustra el uso de la Ecuación (62).

EJEMPLO 17
El ayuntamiento de Gotham tiene 5 patrullas. La policía recibe dos tipos de llamadas: las de urgencia (tipo 1) y las de no urgencia (tipo 2). Los tiempos entre llegadas para cada tipo de llamadas están distribuidos exponencialmente con un promedio de 10 llamadas de urgencia y 20 de no urgencia cada hora. Cada tipo de llamada tiene un tiempo exponencial de servicio, con un promedio de 8 minutos. Suponga que, en promedio, 6 de los 8 minutos representan el tiempo que Tarda la patrulla de la estación de policía hasta el lugar de la llamada, y regresar. Se les da prioridad a las llamadas de urgencia sobre las que no son de urgencia. En promedio, ¿cuánto tiempo pasará entre una solicitud que no es de urgencia y la llegada de la patrulla?

Solución
Los datos son s = 5, $\lambda_1$ = 10 llamadas/h, $\lambda_2$ = 20 llamadas/h, $\mu$ = 7.5 llamadas/h, $\rho$ = (10+20)/(7.5*5)=.80 , $a_0 = 0, a_1 = \frac{10}{37.5} = .267, a_2 = \frac{10+20}{37.5} = .80$. Según la Tabla 6, para s = 5 y $\rho$ = .80, P(j $\geq$ 5) = .55. Entonces, la ecuación (62) da como resultado

$$W_{q2} = \frac{.55}{5*7.5*(1-.267)(1-.80)} = .10 horas = 6 \min$$

El tiempo promedio entre la recepción de una llamada que no es urgente y la llegada de la autopatrulla al lugar es $W_{q2} + \frac{1}{2}$ (tiempo total de viaje por llamada) = 6+3=9 minutos.

**PRIORIDAD ADQUIRIDA**

Terminaremos el estudio de los sistemas de colas con prioridad describiendo un sistema de colas prioritarios. En este sistema puede hacerse a un lado a un cliente de menor prioridad, por ejemplo el cliente tipo i, siempre que llegue un cliente de mayor prioridad. Una vez que no haya clientes de mayor prioridad el cliente tipo i que se hizo a un lado vuelve a entrar al servicio. En un modelo recuperable el servicio al cliente continúa a partir del punto en el que se interrumpió. En un modelo de repetición el cliente inicia el servicio de nuevo cada vez que vuelve a entrar al sistema Naturalmente, que si los tiempos de servicio tienen distribución exponencial, las disciplinas recuperable y de repetición son idénticas (¿por qué?). En la notación de Kendall-Lee representaremos un sistema de colas prioritario poniendo PRP (de Preemptive Priority}. Veremos ahora un sistema $M_i/M/1/PRP/\infty/\infty$ en el que el tiempo de servicio para cada cliente es exponencial con promedio $\frac{1}{\mu}$, y los tiempos entre llegada para el í-ésimo cliente se distribuyen en forma exponencial con frecuencia $\lambda_i$. Entonces

(63) $$W_k = \frac{\frac{1}{\mu}}{(1-a_{k-1})(1-a_k)}$$

en la cual $a_0$ = O

$$a_k = \sum_{i=1}^{k} \frac{\lambda_i}{\mu}$$

Por rabones obvias, las disciplinas de prioridad raramente se usan sí tos clientes son personas. Sin embargo, se usan a veces para "clientes" como trabajos computadora. El ejemplo siguiente ilustra la Ecuación (63).

EJEMPLO 18
En el centro de cómputo de la universidad, los trabajos de los profesores (tipo 1) tienen prioridad ante los trabajos de los alumnos (tipo 2). El tiempo de procedimiento de los trabajos de ambos tipos sigue una distribución exponencial con 30 segundos de promedio. Cada hora entran 10 trabajos de maestros y 50 de

alumnos. ¿Cuál es el tiempo promedio que pasa entre la llegada de un trabajo de estudiante y su terminación? Suponga que los tiempos entre llegadas son exponenciales.

Solución

Sabemos que $\mu$ = 2 trabajos por minuto, $\lambda_1$ = 1/6 trabajo por minuto y $\lambda_2$= 5/6 trabajos por minuto. Entonces

$a_0$ = 0 $a_1 = \frac{1}{12}$ $a_2 = \frac{1}{12} + \frac{5}{12} = \frac{1}{2}$

La Ecuación (63) da como resultado

$$W_k = \frac{\frac{1}{2}}{(1-\frac{1}{12})(1-\frac{1}{2})} = \frac{12}{11}\min$$

Pasa un promedio de 1.09 minutos desde que un estudiante lleva un trabajo y termina.

**PROBLEMAS**

1. El profesor de inglés, Jacob Bright, tiene una mecanógrafa que trabaja 8 horas diarias. El profesor le pide tres tipos de trabajos: pruebas, trabajos de investigación y material de clase. Se dispone de la información de la Tabla 11. El profesor Bright dijo a la mecanógrafa que las pruebas tienen prioridad sobre los trabajos de investigación, y que éstos tienen prioridad sobre el material de clase. Si se supone un sistema no prioritario, calcule el tiempo esperado que el profesor tendrá que esperar para que se termine cada tipo de trabajo.

2. Suponga que un supermercado utiliza un sistema en el que lodos los clientes hacen una cola para esperar al primer cajero disponible. Suponga también que el tiempo de servicio para un cliente que compra k artículos se distribuye en forma exponencial con un promedio de k segundos. Asimismo, un cliente que compra k

Tabla 11

| TIPO DR TRABAJO | FRECUENCIA (Trabajos por día) | $E(S_i)$ horas | $E(S_i^2)$ horas$^2$ |
|---|---|---|---|
| Pruebas | 2 | 1 | 2 |
| Trabajos de investigación | .5 | 4 | 20 |
| Material de clase | 5 | .5 | 0.50 |

artículos siente que el costo de esperar en la cola durante un minuto es 1 dòlar/k. Si se pudiera asignar prioridad a los clientes, ¿qué asignación de prioridades minimiza el costo esperado de espera en el que incurre los clientes del supermercado? ¿Por qué el costo de espera de un cliente por minuto sería una función decreciente de K?

3. Cuatro médicos de un hospital trabajan en una sala de urgencias que atiende a tres tipos de pacientes. El tiempo que pasa un médico con los pacientes de los

tres tipos tiene distribución exponencial con un promedio de 15 minutos. Los tiempos entre llegadas para cada tipo de cliente son exponenciales, y el número promedio de llegadas por hora para cada tipo de paciente es como sigue: tipo 1, tres pacientes; tipo 2, cinco pacientes y tipo 3, tres pacientes. Suponga que los pacientes del tipo 1 tienen la mayor prioridad, y que los del tipo 3 tienen la menor. No hay privilegios. ¿Cuál es el tiempo promedio que espera cada tipo de cliente para que lo atienda un médico?

4. Se tiene un sistema de cómputo el cual procesa dos tipos de trabajos. El tiempo promedio de ejecución de cada tipo de trabajo es $\frac{1}{\mu}$. Los tiempos entre llegadas para cada tipo de trabajo tienen distribución exponencial, con un promedio de $\lambda_i$, trabajos tipo i que llegan cada hora. Revise los tres casos siguientes:

1. Los trabajos tipo 1 tienen prioridad sobre los tipo 2, y se permiten privilegios.

2. Los trabajos tipo 1 tienen prioridad sobre los tipo 2, y no se permiten privilegios.

3. El servicio a los trabajos es bajo la disciplina FIFO.

¿Bajo cuál sistema salen más rápido los trabajos tipo 1? ¿Bajo cuál salen más lento? Conteste las mismas preguntas para los trabajos tipo 2.

# 3-SIMULACIÓN

LA SIMULACIÓN ES una técnica muy poderosa y ampliamente usada en las ciencias administrativas, para analizar y estudiar sistemas complejos. En los capítulos anteriores nos ocupamos de la formulación de modelos que se pudieran resolver en forma analítica.

En casi todos esos modelos nuestra meta fue determinar soluciones óptimas. Sin embargo, debido a la complejidad, las relaciones estocásticas, etc./ no todos los problemas del mundo real se pueden representar adecuadamente en forma de modelo como las de los capítulos anteriores. Los intentos por usar modelos analíticos para sistemas como éstos, en general necesitan de tantas hipótesis de simplificación que es probable que las soluciones no sean buenas, o bien, sean inadecuadas para su realización. En esos casos, con frecuencia la única opción de modelado y análisis de que dispone quien toma decisiones es la simulación.

Se puede definirá la simulación como la técnica que imita el funcionamiento de un sistema del mundo real cuando evoluciona en el tiempo. Esto se hace, por lo general, al crear un modelo de simulación. Un modelo de simulación comúnmente, toma la forma de un conjunto de hipótesis acerca del funcionamiento del sistema, expresado como relaciones matemáticas o lógicas entre los objetos de interés del sistema. En contraste con las soluciones matemáticas exactas disponibles en la mayoría de los modelos analíticos, el proceso de simulación incluye la ejecución del modelo a través del tiempo, en general en una computadora, para generar muestras representativas de las mediciones del desempeño o funcionamiento. En este aspecto, se puede considerar a la simulación como un experimento de muestreo acerca del sistema real, cuyos resultados son puntos de muestra. Por ejemplo, para obtener la mejor estimación del promedio de la medición de funcionamiento, calculamos el promedio de los resultados de muestra. Es claro que tanto más puntos de muestra generemos, mejor será nuestra estimación. Sin embargo, hay otros factores que tienen influencia sobre la bondad de nuestra estimación final, como las condiciones iniciales de la simulación, la longitud del intervalo que se simula y la exactitud del modelo mismo. Estos temas se analizarán después en este capítulo.

Como en la mayoría de las otras técnicas, la simulación tiene sus ventajas y sus desventajas. La ventaja principal es que la teoría de la simulación es relativamente directa. En general, los métodos de simulación son más fáciles de aplicar que los analíticos. En tanto que los métodos analíticos necesitan de muchas simplificaciones,
los modelos de simulación tienen pocas restricciones como éstas y, por lo tanto, permiten una flexibilidad mucho mayor en la representación del sistema real. Una vez formado un modelo, se puede usar en forma repetida para analizar diversas políticas, parámetros o diseños. Por ejemplo, si una empresa tiene un modelo de simulación para su sistema de inventario, se pueden probar diversas políticas de inventario con el modelo, en lugar de arriesgar experimentando en el mundo real. Sin embargo, se debe hacer notar que la simulación no es una técnica de optimización. Se usa con más frecuencia para analizar preguntas tipo "¿Qué sucede si...". Es posible optimizar con la simulación, pero, en general es un proceso tardado. También, la simulación puede ser costosa. Sin embargo, con la creación de lenguajes especiales para simulación, costos decrecientes de cómputo y avances en metodologías de simulación, el problema del costo se hace cada vez menos importante.

En este capítulo la atención se centrará en modelos de simulación y en la técnica de simulación. Presentaremos varios ejemplos de modelos de simulación y exploraremos conceptos tales como números aleatorios/ mecanismos de flujo de tiempo/ muestreo Monte Carlo y lenguajes de simulación y temas estadísticos en la simulación.

# 3.1 TERMINOLOGÍA BÁSICA

Comenzaremos el estudio presentando algo de la terminología usada en simulación. En la mayor parte de !os estudios de simulación nos ocupamos de la simulación de algún sistema. Así, para modelar un sistema, debemos comprender el concepto de sistema.

Entre las muchas maneras de definir un sistema, la definición más adecuada para problemas de simulación es la propuesta por Schmidl y Taylor (1970).

DEFINICIÓN

Un sistema es un conjunto de entidades que actúan e interactúan para la realización de un fin lógico

Sin embargo, en la práctica esta definición tiende por lo general a ser más flexible. La descripción exacta del sistema normalmente depende de los objetivos del estudio de simulación. Por ejemplo, lo que puede ser un sistema para un estudio particular, puede ser solo un subconjunto del sistema general para otro.

Los sistemas tienden en general a ser dinámicos; su estado varía en el tiempo. Para describir este caso usamos el concepto de estado de un sistema.

DEFINICIÓN El estado de un sistema es el conjunto de variables necesarias para describir la condición del sistema en un momento determinado.

Como ejemplo de un sistema, veamos un banco. En este caso, el sistema consiste en los empleados y los clientes que esperan formados o están siendo atendidos. A medida que llegan o se van los clientes, cambia el estado del sistema. Para describir este cambio de estado se necesita un conjunto de variables, llamadas variables de estado. Por ejemplo, el numero de empleados ocupados, el número de clientes en el banco, el tiempo de llegada del próximo cliente y el tiempo de salida de los clientes en el servicio describen entre sí todo cambio posible en el estado del banco. Así, esas variables podrían emplearse como variables de estado para este sistema. En un sistema, un objeto de interés se llama entidad y cualquier propiedad de una entidad se llama atributo. Por ejemplo, los clientes del banco pueden ser descritos como entidades y las características de los clientes, como por ejemplo sus ocupaciones, pueden ser los atributos.

Los sistemas se pueden clasificar en discretos o continuos.

DEFINICIÓN Un sistema discreto es aquel en el cual las variables de estado cambian sólo en puntos discretos o contables en el tiempo.

Un banco es un ejemplo de sistema discreto, ya que las variables de estado cambian sólo cuando llega un cliente, o cuando un cliente termina sus trámites y se va. Estos cambios tienen lugar en puntos discretos en el tiempo.

DEFINICIÓN Un sistema continuo es aquel en el que las variables de estado cambian en forma continua a través del tiempo.

Un proceso químico es un ejemplo de un proceso continuo. En este caso, el estado del sistema varía continuamente a través del tiempo. Estos sistemas se modelan en general mediante ecuaciones diferenciales. En este capítulo no se estudiará ningún sistema continuo.

Hay dos tipos de modelos de simulación: estáticos y dinámicos.

DEFINICIÓN Un modelo estático de simulación es una representación de un sistema en determinado punto en el tiempo.

En general, llamaremos simulación de Monte Carlo a una simulación estática.

DEFINICIÓN Una simulación dinámica es una representación de cómo evoluciona un sistema a través del tiempo.

Dentro de estas dos clasificaciones, una simulación puede ser determinista o estocástica. Un modelo determinista de simulación es aquel que no contiene variables aleatorias; un modelo estocástico de simulación contiene una o más variables aleatorias. Los modelos discretos y continuos de simulación son semejantes a los sistemas discretos y continuos. En este capítulo nos concentraremos en modelos estocásticos discretos. A estos modelos se les llama modelos de simulación de evento discreto; la simulación de eventos discretos se relaciona con el modelado de un sistema estocástico a medida que evoluciona a través del tiempo mediante una representación en la que las variables de estado cambian sólo en puntos discretos en el tiempo.

## EJEMPLO DE UNA SIMULACIÓN DE EVENTO DISCRETO

Antes de proseguir con los detalles del modelado de la simulación, será útil trabajar con un ejemplo sencillo de simulación para ilustrar algunos de los conceptos básicos en simulación de evento discreto. El modelo que hemos escogido como ejemplo inicial es un sistema puesto en cola de espera con un solo empleado. A este sistema llegan los clientes que proceden de determinada población y son atendidos de inmediato si el empleado está desocupado, o se forman (hacen cola) para esperar si el empleado está ocupado. Ejemplos de este tipo de sistema son una peluquería con un peluquero, una tienda pequeña con sólo una caja, y una sola ventanilla de boletos en una terminal aérea.

El mismo modelo fue estudiado en el capitulo anterior en relación con la teoría dé las colas de espera. Allí usamos un modelo analítico para determinar las diversas características de funcionamiento del sistema. Sin embargo, tuvimos que hacer algunas hipótesis restrictivas para poder emplear la teoría de colas. En particular, cuando estudiamos un sistema M/M/1 tuvimos que suponer que tanto los tiempos entre llegadas como los de servicio estaban distribuidos exponencialmente. En muchos casos, estas hipótesis pueden no ser adecuadas. Por ejemplo, las llegadas al mostrador de una aerolínea tienden a

ser en lotes de personas, debido a factores tales como llegadas de autobuses de transbordo y de vuelos de conexión. Para un sistema de esos, se debe usar una distribución empírica de tiempos de llegada, lo cual significa que el modelo analítico de la teoría de colas ya no es factible. Con la simulación se puede usar cualquier distribución de tiempos entre llegadas y de tiempos de servicio, dando con ello mucho más flexibilidad al proceso de solución.

Para simular un sistema de colas de espera tenemos que describirlo primero. Para este sistema de un solo empleado, suponemos que las llegadas se toman de una población infinita que necesita el servicio. La capacidad de la sala de espera es ilimitada y los clientes se atienden en el orden que lleguen, esto es, en base al primero que llega primero que se atiende (FIFO). Además supondremos que las llegadas se efectúan una a la vez de modo aleatorio y que los tiempos entre llegadas se distribuyen como aparece en la Tabla 1. Todas las llegadas se atienden finalmente con la distribución de tiempos de servicio que se ve en la Tabla 2. También se supone que los tiempos de servicio son aleatorios, A este sistema de colas de espera se le puede representar como se muestra en la Fig. 1.

Antes de tratar los detalles de la simulación misma, debemos definir al estado de este sistema y comprender los conceptos de los eventos y la hora en el reloj con una simulación. Para este ejemplo, usaremos las siguientes variables para definir el

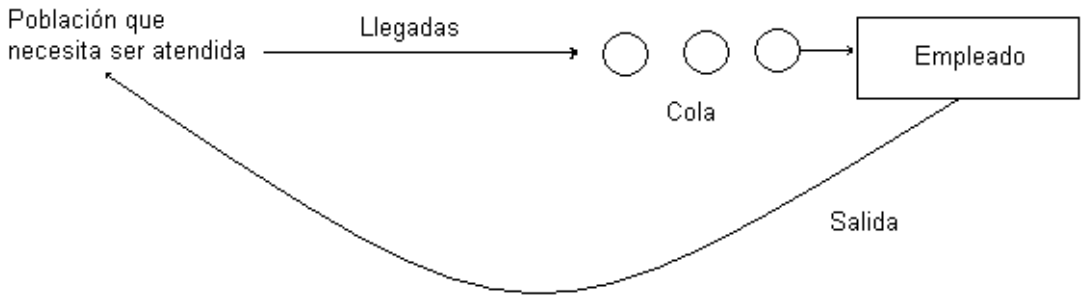
Tabla 1
Distribución de tiempos entre llegadas

| TIEMPO ENTRE LLEGADAS (Minutos) | PROBABILIDAD |
|---|---|
| 1 | .20 |
| 2 | .30 |
| 3 | .35 |
| 4 | .15 |

Tabla 2
Distribución de tiempos de servicio

| TIEMPO DE SERVICIO (Minutos) | PROBABILIDAD |
|---|---|
| 1 | .35 |
| 2 | .40 |
| 3 | .25 |

Figura 1
Sistema de cola de espera con un solo empleado

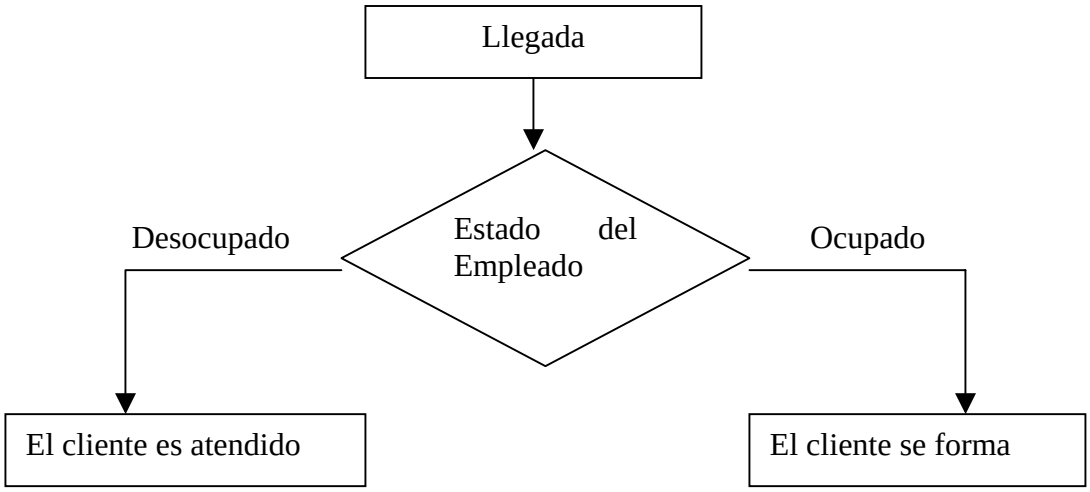



estado del sistema: (1) el número de clientes en el sistema; (2) el estado del empleado; es decir, si está ocupado o desocupado, y (3) la hora de la llegada siguiente.

Estrechamente relacionado con el estado del sistema está el concepto de un evento. Un evento se define como una situación que hace que el estado del sistema cambie en forma instantánea. En el modelo de puesta en cola de espera con un solo empleado hay dos eventos posibles que pueden cambiar el estado del sistema: una llegada al sistema y una salida de él al completar el servicio. En la simulación, estos eventos se programan para llevarse a cabo en determinados puntos en el tiempo. Toda la información acerca de ellos se mantiene en una lista llamada lista de eventos. Dentro de esta lista mantenemos registro del tipo de eventos programados y, más importante, el tiempo al cual estos eventos están programados para llevarse a cabo. Se mantiene el tiempo en una simulación mediante una variable, llamada hora o tiempo del reloj. El concepto de hora se hará más claro a medida que progresemos en el ejemplo.

Comenzaremos esta simulación con un sistema vacío y supondremos en forma arbitraria que nuestro primer evento, una llegada, se efectúa en la hora 0. Esta llegada encuentra desocupado al empleado y es atendido de inmediato. Las llegadas en otros puntos del tiempo pueden encontrar al empleado desocupado u ocupado. Si está desocupado, el cliente es atendido. Si está ocupado, el cliente se forma en la cola de espera. Estas acciones se pueden resumir en la Fig. 2.

Figura 2
Diagrama de flujo de una llegada




A continuación, programamos el tiempo de partida del primer cliente. Esto se hace al generar un tiempo de servicio a partir de la distribución de tiempos de servicio, que se describe más adelante en el capítulo, y establecer el tiempo de partida como

Tiempo de salida = hora de este momento + tiempo de servicio generado (1)

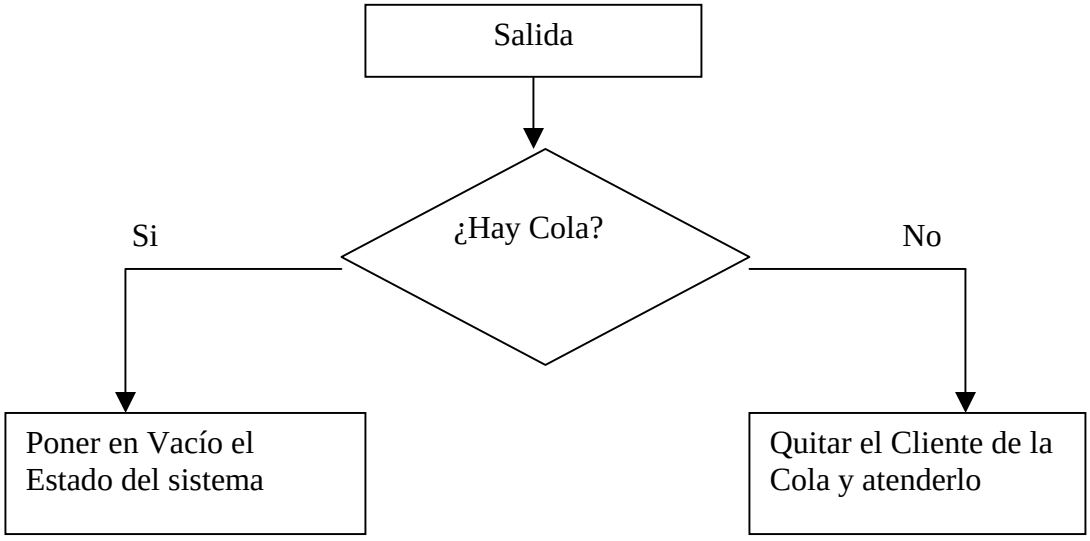
También, programaremos la siguiente llegada al sistema generando al azar un tiempo entre llegadas a partir de la distribución de tiempos entre llegadas y establecer el tiempo de llegada como

Tiempo de llegada = hora de este momento + tiempo entre llegadas que se genera (2)

Si, por ejemplo, hemos generado un tiempo de servicio de 2 minutos, entonces el tiempo de salida para el primer cliente será cuando el reloj marque 2. Igualmente, si hemos generado un tiempo entre llegadas de 1 minuto, la siguiente llegada se programará para cuando el reloj marque 1.

Ambos eventos y sus tiempos programados se mantienen en la lista de eventos. Una vez que hemos completado todas las acciones necesarias, para la primera llegada, se inspecciona la lista de eventos para determinar el siguiente evento programado y su hora. Si el siguiente evento debe ser una llegada, pasamos la hora a la hora programada de llegada y se busca en la secuencia de acciones anterior una llegada. SÍ el evento siguiente es una salida, movemos la hora del reloj a la hora de salida y procesamos una salida. Para una salida comprobamos si la longitud de la cola es mayor que cero. Sí lo es, quitamos al primer cliente de la cola e iniciamos el servicio a éste al establecer una hora de salida mediante la ecuación (1). Si nadie espera, se establece el estado del sistema en desocupado. En la Fig. 3 se resumen estas acciones de salida.

Figura 3
Diagrama de flujo de una salida ,




Este método de simulación se llama mecanismo de avance de hora hasta el siguiente evento, a causa del modo en que se actualiza la hora. Adelantamos el reloj de simulación a la hora del evento más inminente, esto es, el primer evento en la lista. Como las variables de estado sólo cambian en las horas de eventos, se omiten los periodos de inactividad entre los eventos al pasar de un evento a otro. Al hacerlo, efectuamos las acciones propias de cada evento, incluyendo cualquier programación de eventos futuros. Continuamos de esta manera hasta que se satisfaga determinada condición de paro preespecificada. Sin embargo, el procedimiento necesita que en cualquier punto de la simulación tengamos programada una llegada y una salida para el futuro. Así, una llegada futura siempre se programa al procesar una nueva llegada al sistema. Por otro lado, un tiempo de salida, sólo se puede programar cuando un cliente es atendido. Así si el sistema está desocupado no se pueden programar salidas. En esos casos, la práctica normal es programar una salida virtual al hacer que el tiempo de salida sea un número muy grande, digamos 9999 o mayor, si es probable que el reloj rebase las 9999. De este modo nuestros dos eventos consistirán en una llegada real y una salida virtual.

El salto al siguiente evento en el mecanismo del siguiente evento puede ser grande o pequeño; esto es, los saltos son de tamaño variable en este método comparamos este método con el método de avance de hora por incrementos fijos. En este caso adelantamos el reloj de simulación en incrementos de $\Delta t$ unidades de tiempo, donde $\Delta t$ es alguna unidad de tiempo adecuada, en general 1 unidad de tiempo. Después de cada actualización del reloj comprobamos si hay algún evento programado para esa hora. Si es así, llevamos a cabo las acciones adecuadas para el evento. Si no hay nada programado, o si hemos completado todas las acciones necesarias para la hora actual, adelantamos el reloj de simulación $\Delta t$ unidades y repetimos el proceso. Al igual que con el método del siguiente evento, continuamos este modo hasta llegar a la condición prescrita de paro. El mecanismo de avance por incrementos fijos con frecuencia es más sencillo de entender debido a sus etapas fijas de tiempo. Sin embargo, para la mayor parte de los modelos, el mecanismo de evento siguiente tiende a ser más eficaz desde el punto de vista

de cómputo. En consecuencia, sólo emplearemos el método de siguiente evento para la creación de los modelos durante el resto del capítulo.

A continuación mostraremos la mecánica de la simulación del sistema de cola de espera con un solo empleado mediante un ejemplo numérico. En particular, deseamos mostrar cómo se represénla el modelo de simulación en la computadora a medida que la simulación avanza en el tiempo. En la Fig. 4 presentamos el modelos completo de simulación para el modelo de cola con un solo empleado, en forma de diagrama de flujo. Todos los bloques de este diagrama están numerados para tener una referencia fácil. Por simplicidad suponemos que tanto los tiempos entre llegadas como los tiempos de servicio ya se han generado para los primeros clientes a partir de las distribuciones de probabilidad dadas en las Tablas 1 y 2. Estos tiempos se muestran en la Tabla 3, en la cual podemos ver que el tiempo entre la primera y segunda llegadas es 2 unidades, el tiempo entre la segunda y tercera llegadas también es 2 unidades, etc. De igual manera, el tiempo de servicio para el primer cliente es 3 unidades, para el segundo también es 3 unidades, y así sucesivamente.

Para demostrar el modelo de simulación, necesitamos definir algunas variables;

TM = hora de la simulación
AT = tiempo programado para la siguiente llegada
DT = tiempo programada para la siguiente salida
SS = estado del empleado (1 = ocupado, O = desocupado)
WL = longitud de la cola de espera
MX = longitud, en unidades de tiempo, de una corrida de simulación

Una vez vistos estos preliminares, comenzaremos ahora la simulación al inicializar todas las variables (bloque 1 de la Fig. 4). Como se supone que la primera llegada tiene lugar en el tiempo 0, hacemos que AT = 0. También suponemos que el sistema está vacío en el tiempo 0 y, por lo tanto, hacemos que SS = 0, WL = 0 y DT = 9999. Nótese que DT debe ser mayor que MX. Esto significa que nuestra lista de eventos ahora consiste cu dos eventos programados: una llegada en el tiempo 0 y una salida virtual en el tiempo 9999. Con esto se completa el proceso de Inicialización y obtenemos la representación de computadora que muestra la Tabla 4.

Estamos listos para nuestra primera acción en la simulación: búsqueda en la lista de eventos para determinar el primer evento (bloque 2), Como nuestra simulación consiste sólo en dos eventos, únicamente determinamos el primer evento al comparar AT y DT. En otras simulaciones podríamos tener más de dos eventos, de modo que deberíamos tener un sistema eficaz de búsqueda por la lista de eventos. Una llegada está definida por AT< DT; una salida por DT <AT. En este punto 0 es menor que DT = 9999, lo cual indica que a continuación tendrá lugar una llegada. Identificamos este evento como 1 y actualizamos la hora del reloj, TM a la hora del evento 1 (bloque 3). Esto es, hacemos que TM = 0.

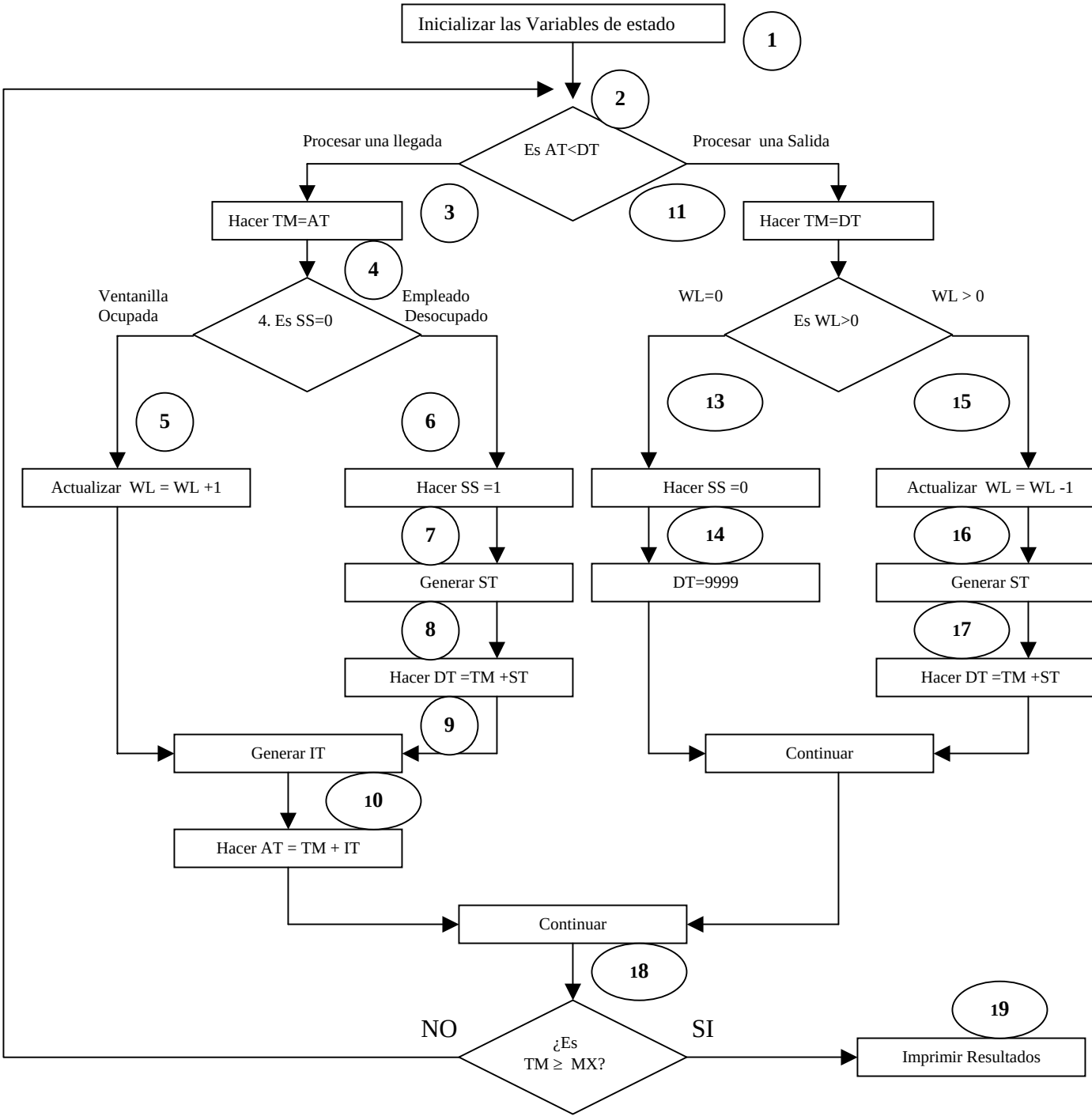
La llegada cuando el tiempo es 0 encuentra vacío al sistema, lo cual se indica por el hecho de que SS = 0 (bloque 4). En consecuencia, el cliente es atendido de inmediato. Para esta parte de la simulación hacemos primero que SS = 1 para indicar que el empleado se encuentra ocupado ahora (bloque 6). A continuación generamos un tiempo de servicio (bloque 7) y establecemos el tiempo de salida para este cliente (bloque 8). En la Tabla 3 vemos que ST para el cliente 1 es 3.

Tabla 3

Tiempos entre llegadas y de servicio, generados

| CLIENTE NUMERO | TEIMPO ENTRE LLEGADAS IT | TIEMPO DE SERVICIO ST |
|---|---|---|
| 1 | | 3 |
| 2 | 2 | 3 |
| 3 | 2 | 2 |
| 4 | 3 | 1 |
| 5 | 4 | 1 |
| 6 | 2 | 2 |
| 7 | 1 | 1 |
| 8 | 3 | 2 |
| 9 | 3 | - |

Figura 4 Diagrama de flujo para el modelo de simulación de un sistema de cola de espera con un solo empleado



Como TM = 0 en este punto, hacemos que DT = 3 para el primer cliente. En otras palabras, el cliente 1 saldrá del sistema a la hora 3. Por último, para completar todas las acciones del procesamiento de una llegada, programamos la siguiente llegada al sistema al generar un tiempo entre llegadas, IT (bloque 9) y establecer la hora de esta llegada mediante la ecuación AT = TM + IT (bloque 10). Como IT= 2, hacemos que AT = 2. Esto es, la segunda llegada tendrá lugar en la hora 2. Al final del evento 1 la representación de la simulación en computadora será como se muestra en la Tabla 4.

En esta etapa de la simulación proseguimos al bloque 18 para determinar si la hora, TM, ha rebasado el tiempo especificado de la simulación, MX. Si es así, imprimimos los resultados (bloque 19) y detenemos la ejecución del modelo de simulación. Sí no es el caso, continuamos con la simulación. A esto se le llama proceso de terminación. Ejecutamos este proceso al final de cada evento. Sin embargo, para este ejemplo, suponemos que MX es un número grande. En consecuencia, de aquí en adelante, no describiremos el proceso de terminación.

En este punto nos regresamos al bloque 2 para determinar el siguiente evento. Como AT = 2 y DT == 3, el siguiente evento, que es el 2, será una llegada a la hora 2. Una vez determinado el siguiente evento, adelantamos la simulación a la hora de esta llegada al actualizar TM a 2.

La llegada cuando el tiempo es 2 encuentra al empleado ocupado, de modo que ponemos a este cliente en la cola de espera al actualizar WL de O a 1 (bloque5). Como el evento actual es una llegada, programamos la siguiente llegada al sistema. Dado que IT = 2 para la llegada 3, la siguiente llegada tiene lugar a la hora 4. Con esto se terminan las acciones necesarias para el evento 2. De nuevo regresamos al bloque 2 para determinar el evento siguiente. De acuerdo con la representación de computadora del sistema en la Tabla 4, vemos que en este punto, final del evento 2, DT = 3 es menor que AT = 4. Esto indica que el evento siguiente, el 3, será una salida a la hora 3, Adelantamos el reloj a la hora de esa salida; esto es, actualizamos TM a 3 (bloque 11).

Cuando el tiempo es 3 procesamos la primera salida del sistema. Con la salida, el empleado queda libre. Comprobamos el estado de la cola de espera para ver si hay clientes que esperen servicio (bloque 12). Como WL == 1, tenemos un cliente en espera. Quitamos a este cliente de la cola, hacemos que WL = 0 (bloque 15) y lo atendemos al generar un tiempo de servicio, ST (bloque 16); y ajustamos el tiempo de salida mediante la relación DT = TM + ST (bloque 17). En la Tabla 3 vemos que para el cliente 2, ST = 3. Como TM = 3, hacemos que DT = 6. Hemos completado ya todas las acciones para el evento 3, obteniendo la representación de computadora que se muestra en la Tabla 4.

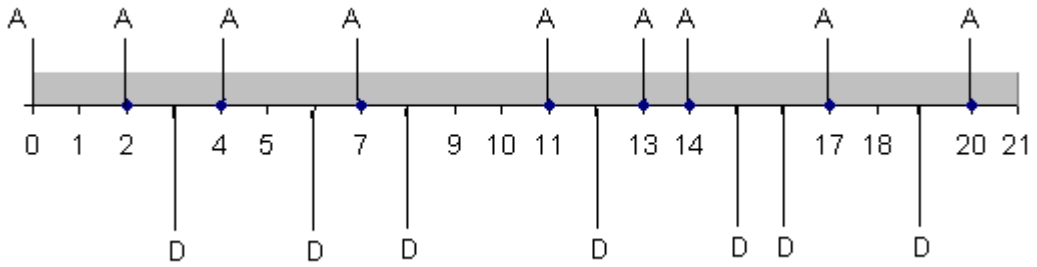
De aquí en adelante, dejamos que e) lector repase la lógica de la simulación para el resto de los eventos de este ejemplo. La Tabla 4 muestra el estado de la simulación al final de cada uno de esos eventos. Nótese que al final de los eventos 8, 10 y 14, salidas, el sistema queda desocupado. Durante la secuencia de acciones para esos eventos hacemos que SS = O (bloque 13) y DT = 9999 (bloque 14). En cada caso, el sistema permanece desocupado hasta que se tiene una llegada. Esta simulación se resume en el diagrama continuo de tiempo de la Fig. 5. En este diagrama las A representan las llegadas y las D las salidas. Nótese que las zonas sombreadas, como la que hay entre los tiempos 9 y 11, quieren decir que el sistema está desocupado.

Tabla 4Representada la simulación computadora

| Final del Evento | Tipo de Evento | Cliente Número | Variables del Sistema | | | Lista de Eventos | |
|---|---|---|---|---|---|---|---|
| | | | TM | SS | WL | AT | DT |
| 0 | Inicialización | 1 | 0 | 0 | 0 | 0 | 9999 |
| 1 | Llegada | 1 | 0 | 1 | 0 | 2 | 3 |
| 2 | Llegada | 2 | 2 | 1 | 1 | 4 | 3 |
| 3 | Salida | 1 | 3 | 1 | 0 | 4 | 6 |
| 4 | Llegada | 3 | 4 | 1 | 1 | 7 | 6 |
| 5 | Salida | 2 | 6 | 1 | 0 | 7 | 8 |
| 6 | Llegada | 4 | 7 | 1 | 1 | 11 | 8 |
| 7 | Salida | 3 | 8 | 1 | 0 | 11 | 9 |
| 8 | Salida | 4 | 9 | 0 | 0 | 11 | 9999 |
| 9 | Llegada | 5 | 11 | 1 | 0 | 13 | 12 |
| 10 | Salida | 5 | 12 | 0 | 0 | 13 | 9999 |
| 11 | Llegada | 6 | 13 | 1 | 0 | 14 | 15 |
| 12 | Llegada | 7 | 14 | 1 | 1 | 17 | 15 |
| 13 | Salida | 6 | 15 | 1 | 0 | 17 | 16 |
| 14 | Salida | 7 | 16 | 0 | 0 | 17 | 9999 |
| 15 | Llegada | 8 | 17 | 1 | 0 | 20 | 19 |

Este ejemplo sencillo muestra algunos de los conceptos básicos de la simulación y el modo en que se puede usar ésta para analizar un problema determinado. Aunque no es probable que este modelo se use para evaluar muchos casos de importancia, ha proporcionado un ejemplo específico y, más importante, ha presentado distintos conceptos claves de simulación. En el resto del capitulo analizaremos algunos de estos conceptos de simulación con más detalle. En el ejemplo no se mencionó la reunión de datos estadísticos, pero se pueden incorporar con facilidad procedimientos al modelo para determinar las medidas de desempeño de este sistema. Por ejemplo, podríamos ampliar el diagrama de flujo para calcular e imprimir el tiempo promedio de espera, el número promedio de personas en la cola de espera y la proporción del tiempo libre. Describiremos con detalle temas estadísticos más adelante en este capítulo.

Figura 5
Representación del continuo de tiempo para la simulación con un solo empleado



## 3.2 NÚMEROS ALEATORIOS Y SIMULACIÓN DE MONTE CARLO

En el ejemplo de simulación de cola de espera vimos que el movimiento fundamental a través del tiempo se logra al generar los tiempos entre llegadas y los tiempos de servicio mediante las distribuciones especificadas de probabilidad. De hecho, todos los tiempos de un evento se determinan ya sea en forma directa o indirecta por estos tiempos generados en servicio y entre llegadas. El procedimiento de generación de esos tiempos, a partir de la distribución dada de probabilidad se conoce como muestreo de acuerdo con distribuciones de probabilidad, o generación de variable aleatoria, o muestreo de Monte Carlo. En esta sección presentaremos y describiremos varios métodos de muestreo distintos a partir de distribuciones discretas. Primero demostraremos la técnica mediante una ruleta y luego la ampliaremos al realizar el muestreo con números aleatorios.

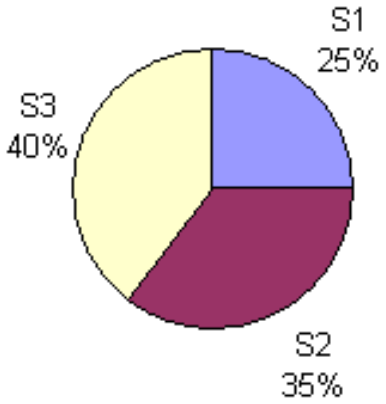
El principio de muestrear de distribuciones discretas se basa en la interpretación de frecuencia que hace la probabilidad. Esto es, a la larga, desearíamos que los resultados se presentaran con las frecuencias especificadas por las probabilidades de la distribución. Por ejemplo, si consideramos la distribución de tiempos de servicio de la Tabla 2, nos gustaría que, a la larga, se generara un tiempo de servicio de 1 minuto el 35% de las veces, uno de 2 minutos el 40% de las veces y uno de 3 minutos el 25% de las veces. Además de obtener las frecuencias correctas, el procedimiento de muestreo debe ser independiente; esto es, cada tiempo de servicio que se genera debe ser independiente de los tiempos de servicio que le anteceden y que le siguen.

Para alcanzar estas dos propiedades por medio de una ruleta, primero dividimos a la ruleta en tres segmentos, cada uno con un área proporcional a una probabilidad en la distribución (véase Fig. 6). Por ejemplo, al primer segmento, digamos $S_1$, le asignamos el 35% del área de la ruleta. Esta área corresponde a la probabilidad .35 y al tiempo de servicio 1 minuto. El segundo segmento, $S_2$, cubre el 40% del área y corresponde a la probabilidad .4 y al tiempo de servicio 2 minutos. Por último, el tercer segmento, $S_3$ se asigna al 25% restante del área y corresponde a la probabilidad .25 y al tiempo de servicio 3 minutos. Si hacemos girar ahora a la ruleta y el indicador cae en el segmento $S_1$, quiere decir que hemos generado un tiempo de servicio de 1 minuto. Si éste cae en el segmento $S_2$, asignamos un tiempo de servicio de 2 minutos. Si cae en el segmento $S_3$ hemos generado un tiempo de servicio de 3 minutos. Si la ruleta es justa, cosa que suponemos, entonces, a la larga, (1) generaremos los tiempos de servicio con la misma frecuencia, aproximadamente, que la especificada en la distribución, y (2) los resultados de cada tirada serán independientes de los resultados que se tengan antes y después.

Ahora ampliaremos esta técnica usando números para la segmentación, en lugar de áreas. Suponemos que la rueda de ruleta tiene 100 números que van del 00 al 99 inclusive. Además suponemos que la segmentación es tal que cada número tiene la misma probabilidad, .01, de salir. Con este método de segmentación asignamos 35 números, digamos del 00 al 34, a) tiempo de servicio 1 minuto. Como cada número tiene probabilidad .01 de salir, los 35 números juntos equivalen a una probabilidad de .35. Igualmente, si asignamos los números del 35 al 74 al tiempo de servicio 2 minutos y los números del

75 al 99 al tiempo de servicio 3 minutos, hemos logrado las probabilidades deseadas. Como antes, hacemos girar la ruleta para generar tos tiempos de servicio, pero con este método, los números determinan en forma directa los tiempos de servicio. En otras palabras, si generamos un número entre 00 y 34, se establece el tiempo de servicio en 1 minuto; si el número es entre 35 y 74, el tiempo de servicio será de 2 minutos, y si el número es entre 75 y 99 será de 3 minutos.

Figura 6. Segmentación de una ruleta




Con este procedimiento de segmentación y una ruleta, equivale a generar números enteros aleatorios entre 00 y 99. Esto es consecuencia del hecho de que una sucesión de números enteros aleatorios es tal que cada número en la secuencia, en este caso del 00 al 99, tiene una probabilidad igual (en este caso .01) de salir, y cada numero aleatorio es independiente de los números antes y después de él. Si ahora tuviéramos un procedimiento para generar los 100 números aleatorios entre 00 y 99, entonces, en lugar de hacer girar una ruleta para obtener un tiempo de servicio, podríamos usar un número aleatorio generado. Técnicamente, un número aleatorio, Ri, se define como una muestra aleatoria independiente tomada de una distribución continua uniforme cuya función de densidad de probabilidad está dada por

$$f(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \textit{en cualquier otrocaso} \end{cases}$$

Así, cada número aleatorio estará distribuido uniformemente sobre el intervalo entre 0 y 1. En consecuencia, es común referirse a estos números como números aleatorios U(0,1), o simplemente como números aleatorios uniformes.

Se pueden generar números aleatorios uniformes de muchos modos distintos. Como nuestro interés sobre los números aleatorios es para usarlos en simulaciones, necesitamos poder generarlos en una computadora. Esto se hace mediante funciones matemáticas llamadas generadores de números aleatorios.

La mayoría de los generadores de números aleatorios usan alguna forma de relación congruente. Los ejemplos de esos generadores comprenden al generador congruente lineal, el generador multiplicativo

y el generador mixto. El generador congruente lineal es, con mucho, el que se usa más. De hecho, la mayoría de las funciones de números aleatorios interconstruidos en sistemas de cómputo emplean este generador. Con este método se produce una sucesión de enteros $x_1$, $x_2$, $x_3$,... entre cero y m - 1 de acuerdo con la siguiente fórmula recursiva:

$$x_{i+1} = (a\, x_i + c) \text{ módulo } m \quad (i == 0,1,2,...)$$

Al valor inicial de $x_0$ se le llama la semilla, "a" es el multiplicador constante, "c" el incremento y "m" el módulo. Estas cuatro variables se llaman parámetros del generador. Con esta relación, el valor de $x_{i+1}$ es igual al residuo de la división de$(a\, x_i + c)$ entre m. El número aleatorio entre 0 y 1 se genera entonces con la ecuación $Ri = x_i / m$

Por ejemplo, si $x_0 = 35$, a = 13, c = 65 y m = 100, el algoritmo actúa como sigue:

Iteración 0 Se hace que $x_0 = 35$, a = 13, c = 65 y m = 100.

Iteración 1 Se calcula

$$x_1 = (13\, x_0 + 65) \text{ módulo } 100 = 20$$

Da como resultado

$$R_1 = 20/100 = 0.20$$

y así sucesivamente.

Cada número aleatorio que se genera con este método será un decimal entre 0 y 1. Nótese que aunque es posible generar un cero, un número aleatorio no puede ser igual a 1. Los números aleatorios que se generan con métodos de congruencia se llaman números pseudoaleatorios. No son números aleatorios verdaderos en el sentido técnico, porque quedan determinados por completo una vez que se define la relación de recurrencia y se especifican los parámetros del generador. Sin embargo, si se seleccionan con cuidado los valores de a, c, m y $x_0$, se puede hacer que los números pseudoaleatorios cumplan con todas las propiedades estadísticas de los números aleatorios. Además de las propiedades estadísticas, los generadores de números aleatorios deben tener otras características importantes si se van a usar en forma eficaz en simulaciones con computadora. Algunas de estas características son (1) la rutina debe ser rápida; (2) la rutina no debe necesitar un gran espacio de almacenamiento; (3) los números aleatorios deben ser reproducibles, y (4) la rutina debe tener un ciclo suficientemente largo; esto es, debemos poder generar una sucesión larga sin repetir los números aleatorios.

Hay un aspecto importante que vale la pena mencionar aquí. La mayor parte de los lenguajes de programación tienen funciones interconstruidas que dan números aleatorios o pseudoaleatorios en forma directa. Por lo tanto, la mayoría de los usuarios sólo necesitan conocer la función de biblioteca en determinado sistema. En algunos sistemas el usuario puede tener que especificar un valor de la semilla $x_0$, pero no es probable que tenga que elaborar un diseño de un generador de números aleatorios. Sin embargo, para más informes, el lector que se interés puede consultar el libro de Banks y Carson (1984), el de Knuth (1969) y el de Law y Kelton(1982).

Enseguida se abordará una etapa más del método de muestreo de Monte Carlo y se creará un procedimiento por medio de números aleatorios generados en una computadora. La idea es transformar los números aleatorios U(0,1) en números aleatorios enteros entre 00 y 99 y, a continuación, usarlos para lograr la segmentación por números. La transformación es un procedimiento relativamente directo. Si se multiplican los números aleatorios (0,1) por 100, quedarán distribuidos uniformemente entre los límites de 0 a 100. Entonces, si se elimina la parte fraccionaria del número, el resultado serán enteros entre 00 y 99 con igual probabilidad. Por ejemplo, si hubiéramos generado el número aleatorio 0.72365, al multiplicarlo por 100 se obtiene 72.365. Si se quita la parte decimal del numero nos quedaremos con el número aleatorio entero 72. En la computadora obtenemos esta transformación al generar primero un número aleatorio U(0,l). Luego, lo multiplicamos por 10 y por último almacenamos el producto mediante una variable entera; esta etapa final truncará la parte decimal del número. Con este procedimiento obtendremos números aleatorios entre 00 y 99.

A continuación formalizaremos este procedimiento y lo usaremos para generar valores aleatorios de una variable aleatoria discreta. El procedimiento consta dedos pasos: (1) se elaborará la distribución de probabilidad acumulada para variable aleatoria dada y (2) usamos esa distribución para asignar los números aleatorios enteros en forma directa a los diversos valores de la variable aleatoria. Para mostrar este procedimiento usaremos la distribución de los tiempos en las llegadas en el ejemplo de puesta en cola de espera de la sección anterior elaboramos la distribución de probabilidad acumulada para este caso, obtenemos las probabilidades que se ven en la Tabla 5.

El primer tiempo entre llegadas, de I minuto, se presenta con una probabilidad .20. Así, necesitamos asignar 20 números aleatorios a este resultado. Si asignan los 20 números de 00 a 19, utilizamos el intervalo de números aleatorios decimal de O a .199999. Nótese que el límite superior de este intervalo queda justo antes la probabilidad acumulada .20, Para el tiempo entre llegadas de 2 minutos asignamos 30 números aleatorios. Si asignamos los números enteros del 20 al 49, notaremos que esto cubre el intervalo de números aleatorios decimales desde 0.20 hasta la 0.49999. Como antes, el extremo superior de este intervalo queda justo antes de la probabilidad acumulada de .50, pero el límite inferior coincide con la probabilidad acumulada anterior de .20. Si ahora al tiempo de llegadas de 3 minutos asignamos los números aleatorios enteros del 50 al 84, notamos que estos números se obtuvieron del intervalo de números aleatorios decimales de 0.50, el mismo cuanto a la probabilidad acumulada asociada con un tiempo entre llegada de 2 minutos, a 0.84999, que es una fracción más pequeña que .85. Por último, se aplica el mismo análisis al tiempo de 4 minutos entre llegadas. En otras palabras, la distribución de probabilidad acumulada nos permite asignar en forma directa intervalos de números aleatorios enteros. Una vez especificados esos intervalos para una distribución dada, lo que tenemos que hacer para obtener el valor de una variable aleatoria es generar un número aleatorio entero y compararlo con asignaciones de números aleatorios. Por ejemplo, si sucediera que el numero aleatorio obtenido fuera 35, éste se traduciría en un tiempo entre llegadas de 2 minutos. Igualmente, el número aleatorio 67 se traduciría en un tiempo de 3 minutos entre llegadas, etc. A continuación demostraremos estos conceptos en un ejemplo de simulación de Monte Carlo.

Tabla 5: Función de distribución acumulada e intervalo de números aleatorios enteros por tiempo entre llegadas

| TIEMPO ENTRE LLEGADAS (Minutos) | PROBABILIDAD | PROBABILIDAD ACUMULADA | INTERVALO DE NÚMEROS ALEATORIOS |
|---|---|---|---|
| 1 | .20 | .20 | 00-19 |
| 2 | .30 | .50 | 20-49 |
| 3 | .35 | .85 | 50-84 |
| 4 | .15 | 1.00 | 85-99 |

# 3.3 SIMULACIONES CON VARIABLES ALEATORIAS CONTINUAS

Los ejemplos de simulación que se han presentado sólo usan distribuciones de probabilidades discretas para las variables aleatorias. Sin embargo, en muchas simulaciones es más realista y práctico usar variables aleatorias continuas. En esta sección presentaremos y describiremos algunos procedimientos para generar cantidades aleatorias a partir de distribuciones continuas. El principio básico es muy semejante al caso discreto. Como en el método discreto, generamos primero un número aleatorio U(0,l) y luego lo transformamos en una cantidad aleatoria de acuerdo con la distribución especificada. Sin embargo, el proceso para llevar a cabo la transformación es bastante distinto del caso discreto.

Hay muchos métodos diferentes para generar cantidades aleatorias continuas. La selección de un algoritmo determinado dependerá de la distribución de la cual deseamos generar, teniendo en cuenta factores tales como la exactitud de las variables aleatorias, las eficacias de cómputo y de almacenamiento, y la complejidad del algoritmo. Los dos algoritmos que más se usan son el método de transformación inversa y el de aceptación o rechazo. Entre esos dos, es posible generar variables aleatorias a partir de casi cualquiera de las distribuciones que más se usan. Presentaremos una descripción detallada de ambos algoritmos, junto con varios ejemplos de cada método. Además, daremos dos métodos para generar variables aleatorias a partir de una distribución normal.

### MÉTODO DE TRANSFORMACIÓN INVERSA

Este método se usa, por lo general, para distribuciones cuya función de distribución acumulada se pueda obtener en forma cerrada. Los ejemplos comprenden las distribuciones exponencial, uniforme, triangular y de Weibull. Para distribuciones cuya función de distribución acumulada no existe en forma cerrada, se podrá usar algún método numérico, como la expansión en serie de potencias, dentro del algoritmo para evaluar la función. Sin embargo, es probable que esto complique el procedimiento a tal grado que resulte mejor usar un algoritmo distinto para generar las cantidades aleatorias. El método de transformación inversa es relativamente fácil de describir y de ejecutar. Consiste en los tres pasos siguientes:

Paso 1: Dada una función de densidad de probabilidad f{x) para una variable aleatoria X, obtener la función de distribución acumulada F(x) como

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Paso 2 Generar un numero aleatorio r. '
Paso 3 Hacer que F(x) = r y despejar x. La variable x es entonces un número aleatorio procedente de la distribución cuya función de densidad de probabilidad es f(x).

A continuación describiremos la mecánica del algoritmo mediante un ejemplo. Para ello se tiene la distribución representada por

$$f(x) = \begin{cases} \frac{x}{2} & 0 \le x \le 2 \\ 0 & otro\ caso \end{cases}$$

Una función de este tipo se llama función de rampa. Se puede representar en forma gráfica como se ve en la Fig. 7. El área bajo la curva, f(x) =x/2 , representa la probabilidad de ocurrencia de la variable aleatoria X. Suponemos que en este caso que X representa los tiempos de servicio de un cajero de banco. Para obtener valores aleatorios a partir de esta distribución mediante el método de transformación inversa, calculamos primero la función de densidad acumulada como

$$F(x) = \int_0^x f(t)dt = \frac{x^2}{4}$$

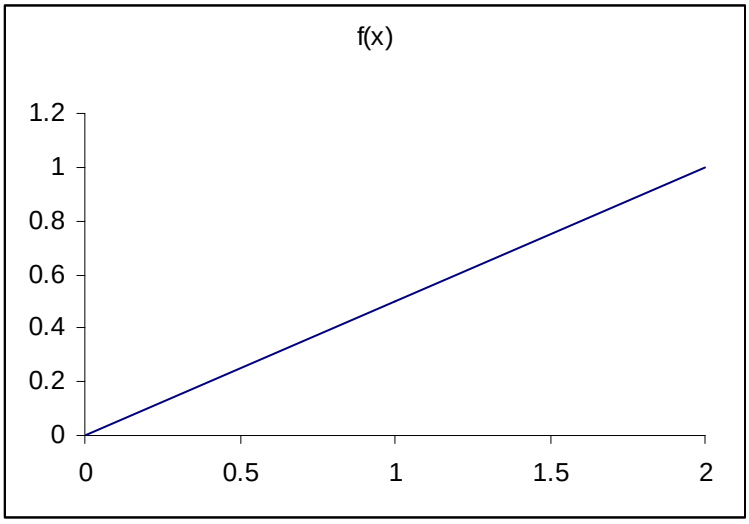
$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{4} & 0 \le x \le 2 \\ 0 & x > 2 \end{cases}$$

A continuación, en el paso 2, generamos un número aleatorio r.\ Por último, en el paso 3, hacemos que F{x) = r y calculamos x.

$$r = \frac{x^2}{4} \Rightarrow x = \pm 2\sqrt{r}$$

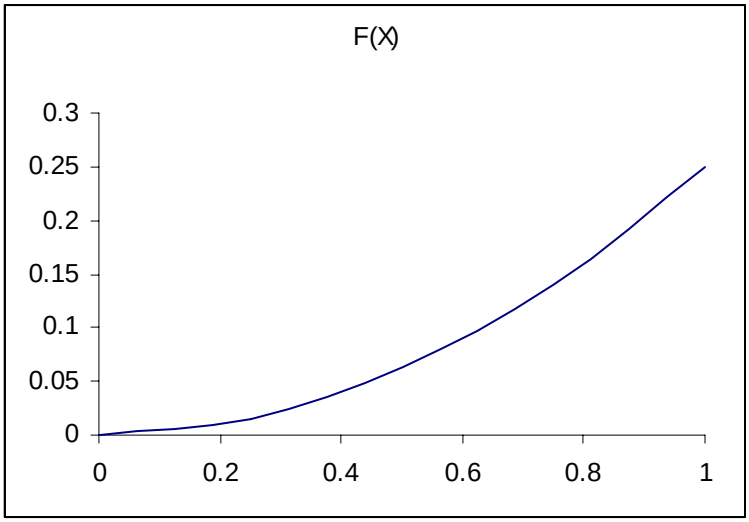
Como los tiempos de servicio sólo se definen para valores positivos de x, no es posible un valor x = - $2\sqrt{r}$ . Esto nos deja con x = $2\sqrt{r}$ como solución de x. A esta ecuación se le llama generador de valores aleatorios o generador de proceso. Así, para obtener un tiempo de servicio, generamos primero un número aleatorio y luego lo transformamos por medio de la ecuación anterior. Cada ejecución de la ecuación nos dará un tiempo de servicio de la distribución dada. Por ejemplo, si se obtiene un número aleatorio r = 0.64, se generará un tiempo de servicio x = $2\sqrt{.64}$ = 1.6.

Figura 7
Función de distribución de probabilidad en rampa




En forma gráfica, el método de la transformación inversa se puede representar como se ve en la Fig. 8. Vemos en esta gráfica que el intervalo de valores para la variable aleatoria es $0 \le x \le 2$, el cual coincide con las probabilidades acumuladas $0 \le F(x) \le 1$. En otras palabras, para cualquier valor de F(x) en el intervalo [0,1] existe un valor correspondiente de la variable aleatoria, representado por x. Como un número aleatorio también se define en el intervalo entre 0 y 1, esto indica que un número aleatorio se puede traducir en forma directa a un valor correspondiente de x mediante la relación r = F(x). La solución para x en términos de r se conoce como el cálculo de la inversa de F{x) y se representa por $x = F^{-1}\{r\}$. De aquí el nombre de transformación inversa. Nótese que si r es igual a cero, generaremos
un cantidad aleatoria igual a cero, el valor más pequeño posible de x. Igualmente, si generamos un numero aleatorio igual a 1, se transformará en 2, el valor más grande posible para x.

Figura 8. Representación gráfica del método de la inversa




EJEMPLO 2 La. distribución exponencial como se mencionó en el capitulo 2, la distribución exponencial tiene aplicaciones importantes en la representación matemática de los sistemas de cola de espera. La función de distribución de probabilidad para esta distribución exponencial está representada por

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & 0 \le x, \lambda > 0 \\ 0 & otro\ caso \end{cases}$$

Con el método de transformación inversa, genere observaciones a partir de una distribución exponencial.

En el paso 1, determinamos la función de distribución acumulada de probabilidad. Esta función esta dada por

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & 0 \le x, \lambda > 0 \\ 0 & otro\ caso \end{cases}$$

Enseguida generamos un numero aleatorio r y hacemos que F(x) = r para despejar x. Esto nos da

$$1 - e^{-\lambda x} = r \Rightarrow 1 - r = e^{-\lambda x} \Rightarrow x = -\frac{1}{\lambda} \ln 1 - r$$

pero como "r" es un número aleatorio entre cero y uno, también lo es 1-r, por lo que da lo mismo indicar que $x = -\frac{1}{\lambda} \ln r$ .

EJEMPLO 3. La distribución uniforme, se tiene una variable aleatoria X que se distribuye uniformemente en el intervalo [a, b]. La función de distribución de probabilidad para este caso está representada por

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & otro\ caso \end{cases}$$

la función de distribución acumuladada en este caso esta dada por

$$f(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

Para utilizar una distribución uniforme, primero generamos un número aleatorio r y a continuación hacemos F(x) = r para despejar x. Esto da como resultado

$$r = \frac{x-a}{b-a} \Rightarrow x = (b-a)r + a$$

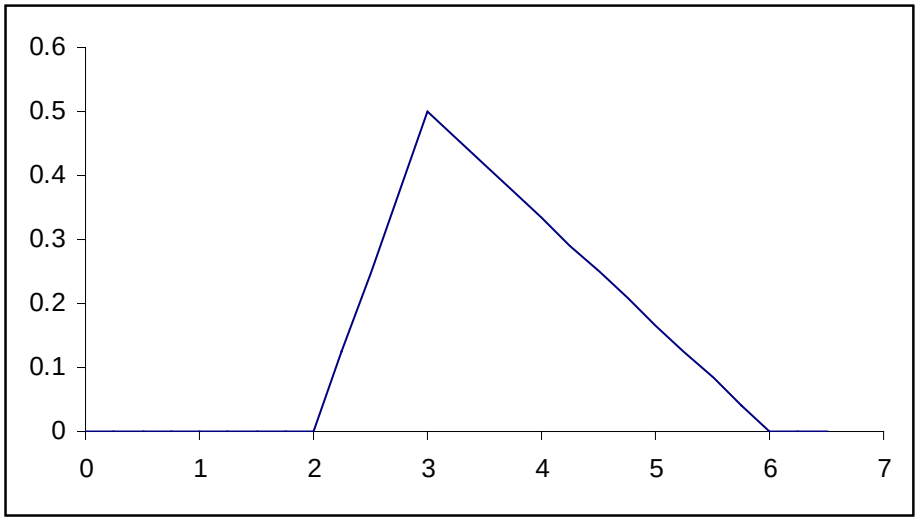
EJEMPLO 4 Se tiene una variable aleatoria X cuya función de densidad de probabilidad es

$$f(x) = \begin{cases} \frac{1}{2}(x-2) & 2 \le x \le 3 \\ \frac{1}{2}(2 - \dfrac{x}{3}) & 3 < x \le 6 \\ 0 & \text{en otro caso} \end{cases}$$

Use el método de transformación inversa para generar observaciones a partir de la distribución. Esta distribución, que se llama triangular, se representa en la Fig. 8. Tiene los puntos extremos [2,6] y su moda está en 3. Podemos ver que el 25% del área bajo la curva queda en el intervalo de x de 2 a 3, y el 75% restante queda en el intervalo de 3 a 6. En otras palabras, 25% de los valores 3e la variable aleatoria1 queda entre 2 y 3, y el otro 75% queda entre 3 y 6. La distribución triangular tiene aplicaciones importantes en simulación; se usa con frecuencia en la representación de actividades para las cuales hay pocos o ningún dato. Para una explicación detallada de esta distribución, consulte a Banks y Carson (1984) o Law y Kelton (1982).

Figura 9

Función de densidad para una distribución triangular




Solución La función de distribución acumulada para este caso es

$$F(x)=\begin{cases} 0 & x<0 \\ \frac{1}{4}(x-2)^2 & 2\le x\le 3 \\ \frac{1}{12}(x^2-12x+24) & 3<x\le 6 \\ 1 & x>6 \end{cases}$$

Luego se genera un número aleatorio r

$$r=\tfrac{1}{4}(x-2)^2 \quad 0\le r\le .25 \Rightarrow x=2\pm 2\sqrt{r}\,(\text{solo el caso positivo es posible})$$
$$x=2+2\sqrt{r}$$

$$r=\tfrac{1}{12}(x^2-12x+24) \quad .25<r\le 1 \Rightarrow 12r+12=(x^2-12x+24)+12$$
$$\Rightarrow 12r+12=(x-6)^2 \Rightarrow x=6\pm 2\sqrt{3-3r}\,(\text{solo el caso negativo es posible})$$
$$x=6-2\sqrt{3-3r}$$

## MÉTODO DE ACEPTACIÓN O RECHAZO

Hay varias distribuciones importantes, incluyendo la de Erlang que se usa en modelos de cola, de espera y la función beta, que se usa en PERT, cuyas funciones de distribución acumulada no existen en forma cerrada. Para esas distribuciones debemos recurrir a otros métodos de generación de variables aleatorias, uno de los cuales es el de aceptación o rechazo. Este método se usa en general para distribuciones cuyos dominios estén definidos en intervalos finitos. Así, dada una distribución cuya función de distribución de probabilidad f(x) esté definida en el intervalo $a \leq x \leq b$, el algoritmo consiste en los pasos siguientes:

a. Se debe seleccionar una M constante (al que M sea el valor máximo de f{x) en el intervalo [a,b].

b. Generar dos números aleatorios: $r_1$, $r_2$

c. CalcuIar x* = a + b{b - a)$r_1$. Con ello se asegura que cada miembro de [a,b] tenga la misma probabilidad de ser seleccionado como x*.

d. Evaluar la función f(x) en el punto x*. Sea este valorar f(x*).

1. e. Si $r_2 \leq f(x^*)$ tomar x* como valor aleatorio a partir de la distribución cuya función de
2. distribución de probabilidad es f{x}. En cualquier otro caso, rechazar x* y regresar al paso 3.

Nótese que el algoritmo continua de regreso al paso 2 hasta que se acepte una variable aleatoria. Para ello se pueden necesitar varías iteraciones. Por este motivo, el algoritmo puede ser relativamente ineficaz. Sin embargo, la eficacia depende mucho de la forma de la distribución. Hay varios modos mediante los cuales se puede hacer más eficaz al método.

Uno de ellos es emplear una función en el paso a), en lugar de una constante. Véase Fishman (1978) o Law y Kelton (1982), que presentan los detalles del algoritmo.

## MÉTODOS DIRECTO Y DE CONVOLUCION PARA LA DISTRIBUCIÓN NORMAL

Debido a la importancia de la distribución normal se ha dado mucha atención a generar variables aleatorias normales. Ello ha originado muchos algoritmos distintos para la distribución normal. Tanto el método de transformación inversa como el de aceptación o rechazo, son inadecuados para la distribución normal, porque (1) no existe la función de distribución acumulada en forma cerrada y (2) la distribución no está definida en un intervalo finito. Aunque es posible emplear métodos numéricos en el método de transformación inversa y truncar la distribución para el método de aceptación o rechazo, hay otros métodos que tienden a será mucho más efectivos- En esta sección describiremos dos de ellos; primero, un algoritmo que se basa en técnicas de convolución y después, un algoritmo de transformación directa que produce dos variables estándar con promedio 0 y variancia 1.

En el algoritmo de convolución se hace uso directo del teorema de límite central. Este teorema afirma que la suma Y de n variables aleatorias independientes e idénticamente distribuidas, $Y_1$, $Y_2$,..., $Y_n$, cada

una con promedio $\mu$ y variancia finita $\sigma^2$ tiene una distribución casi normal con promedio $n\mu$ y varianza $n\sigma^2$. Si aplicamos ahora esto a variables aleatorias, U(0,l), $R_1$, $R_2$,..., $R_n$ con media =0.5 y $\sigma^2$ = 1/12, entonces

$$Z = \frac{\sum_{i=1}^{n} R_i - .5n}{\left(\frac{n}{12}\right)^{1/2}}$$

es aproximadamente normal con promedio 0 y variancia 1. Deberíamos esperar que esta aproximación funcione mejor a medida que crece n. Sin embargo, la mayoría de las citas acerca de simulación sugieren que se use un valor de n = 12. Usar 12 no sólo parece adecuado sino que, más importante, tiene la ventaja de que simplifica el procedimiento de cálculo. Si ahora sustituimos n = 12 en la ecuación anterior, el generador de proceso se simplifica a

$$Z = \sum_{i=1}^{12} R_i - 6$$

Esta ecuación evita una raíz cuadrada y una división, las cuales son rutinas tardadas en una computadora.

Si deseamos generar una variable normal X con media $\mu$ y variancia finita $\sigma^2$ primero generamos Z mediante este generador de proceso y luego la transformamos por medio de la relación $X = \mu + \sigma Z$. Nótese que esta convolución es exclusiva de la distribución normal y que no se puede ampliar a otras distribuciones. Desde luego, otras distribuciones se prestan a métodos de convolución. Por ejemplo, podemos generar variables aleatorias a partir de una distribución de Erlang con parámetro de forma k y parámetro de rapidez $k\lambda$, usando el hecho que una variable aleatoria de Erlang se puede obtener mediante la suma de k variables aleatorias exponenciales iid, cada una con parámetro $k\lambda$.

El método directo para la distribución normal fue creado por Box y Muller (1958). Aunque no es tan eficaz como algunas de las técnicas más modernas, es fácil de aplicar y ejecutar. El algoritmo genera dos números aleatorios U(0,1), $r_1$ y $r_2$, para después transformarlos en dos valores aleatorios normales, cada uno con media 0 y variancia 1, por medio de las transformaciones directas

$$Z_1 = (-2\ln r_1)^{1/2} \sin 2\pi r_2$$
$$Z_2 = (-2\ln r_1)^{1/2} \cos 2\pi r_2$$

esto se puede verificar si se toma la siguiente transformación

$$r_1 = e^{-(z_1^2 + z_x^x)/2}$$
$$r_2 = \frac{1}{2\pi} \tan^{-1}\left(\frac{z_1}{z_2}\right)$$

Y el jacobiano $\frac{\partial(r_1, r_2)}{\partial(z_1, z_2)} = \left[\frac{1}{\sqrt{2\pi}} e^{-z_1^2/2}\right]\left[\frac{1}{\sqrt{2\pi}} e^{-z_2^2/2}\right]$

Como en el método de convolución, es fácil transformar estas variables normales estandarizadas en variables normales $X_1$ y $X_2$ a partir de la distribución con media y variancia al mediante las ecuaciones

$$X_1 = \mu + \sigma Z_1$$
$$X_2 = \mu + \sigma Z_2$$

El método directo produce variables aleatorias normales exactas, en tanto que el método de convolución tan sólo nos da variables aleatorias normales aproximadas. Por este motivo, el método directo se usa mucho mas. Consulte, para detalles de éstos y otros algoritmos, los libros de Fishman (1978), o bien, de Law y Kelton (1982).

## 3.4 ANÁLISIS ESTADÍSTICO EN LAS SIMULACIONES

Como se mencionó antes, los datos obtenidos de la simulación siempre presentan variabilidad aleatoria, ya que se alimentan variables aleatorias al modelo de simulación. Por ejemplo, si ejecutamos dos veces la misma simulación, cada una con distinta secuencia de números aleatorios es seguro que, las medidas estadísticas generadas en los dos casos tendrán valores distintos. Debido a ello, debemos usar métodos estadísticos para analizar los resultados de las simulaciones. Si el desempeño del sistema se mide con un parámetro, digamos $\theta$, entonces nuestro objetivo en la simulación será obtener una estimación $\bar{\theta}$ de $\theta$ y determinar la exactitud del estimador $\bar{\theta}$. Medimos esta exactitud con la desviación estándar, que también se llama error estándar, de $\bar{\theta}$. La medida general de la variabilidad se enuncia a menudo en la forma de un intervalo de confianza a determinado nivel de confianza. Así, el objeto del análisis estadístico es estimar este intervalo de confianza.

Se complica la determinación de los intervalos de confianza en las simulaciones por el hecho de que los resultados pocas veces son independientes, si es que lo llegan a ser. Esto es, los datos se autocorrelacionan. Por ejemplo, en una simulación de cola, el tiempo de espera de un cliente suele depender de los clientes anteriores. Igualmente, en una simulación de inventario, los modelos se establecen, por lo general, de tal manera que el inventario inicial en un día determinado sea el inventario final del día anterior, con lo cual se crea una correlación. Esto significa que los métodos estadísticos clásicos, en los cuales suponemos independencia, no son directamente aplicables al análisis de resultados de simulación. Por lo tanto, modificaremos los métodos estadísticos para hacer inferencias adecuadas a partir de los datos de la simulación.

Además del problema de la autocorrelación, podemos tener un segundo problema en que la especificación de las condiciones iniciales del sistema en el tiempo 0 pueden influir en los resultados. Por ejemplo, supongamos que en la simulación de la cola de espera del ejemplo de este capítulo la distribución de tiempos entre llegadas y de servicio es tal que el tiempo promedio de espera por cliente es mayor de 15 minutos. En otras palabras, el sistema está muy congestionado. Si fuéramos a iniciar esta simulación sin personas en el sistema, los pocos clientes iniciales tendrán tiempos de espera cero o muy pequeños. Estos tiempos iniciales de espera dependen mucho de las condiciones iniciales, y por lo tanto, pueden no ser representativos del comportamiento del sistema en estado estable. A este periodo inicial, antes que la simulación alcance el estado estable, se le llama periodo transitorio, o periodo de calentamiento.

Hay dos métodos para superar los problemas relacionados con el periodo transitorio. El primero es usar un conjunto de condiciones iniciales que sea representativo del sistema en estado estable. Sin embargo, en muchas simulaciones puede ser difícil establecer esas condiciones iniciales. Esto es especialmente válido en las simulaciones de colas. El otro método es dejar que la simulación se ejecute durante un rato y desechar la parte inicial de la simulación. Con este método estamos suponiendo que la parte inicial de la simulación lleva al modelo hasta un estado de equilibrio. Como no anotamos ninguna medida estadística durante la etapa de calentamiento, podemos reducir mucho del sesgo de la inicialización. Desafortunadamente, no hay una manera fácil de estimar cuántos datos iniciales borrar para reducir el sesgo de inicialización a niveles insignificantes. Como cada modelo de simulación es distinto, depende del analista determinar cuándo termina el periodo transitorio. Aunque esto es difícil, hay algunos lineamientos que se pueden usar. Para estos detalles y otros del tema, consulte a Law y Kelton(l982).

Con objeto de analizar resultados, en general clasificamos las simulaciones en dos tipos: simulaciones de terminación y simulaciones de estado estable. Una simulación de terminación es aquella que se ejecuta durante un tiempo $T_E$, donde E es un evento o eventos especificados que detienen la simulación. El evento E puede ser un tiempo especificado, en cuyo caso la simulación se ejecuta durante un lapso de tiempo determinado. O bien, si es una condición especificada, la duración de la simulación será una variable aleatoria. Una simulación de estado estable es aquella que se ejecuta durante un tiempo largo, esto es, la duración de la simulación "tiende a infinito."

Con frecuencia, el tipo de modelo determina qué tipo de análisis de resultados es el adecuado para determinada simulación. Por ejemplo, en la simulación de un banco es más probable que usemos una simulación de terminación, ya que el banco, cierra todas las tardes, lo cual nos da un evento de terminación adecuado. Cuando se simula un sistema de cómputo, puede ser más adecuada una simulación de estado estable, ya que la mayor parte de los sistemas grandes de cómputo no dejan de funcionar, excepto en casos de descompostura o de mantenimiento. Sin embargo, el sistema o modelo no siempre es el mejor indicador de qué simulación debe ser la más adecuada. Es muy posible usar el método de simulación de terminación para sistemas más adecuados a simulaciones de estado estable, y viceversa. En esta sección daremos una descripción detallada del análisis estadístico asociado con simulaciones de terminación. El análisis para las simulaciones de estado estable es mucho más complicado. Para los detalles de éste ultimo, consulte las obras de Banks y Carson (1984) y de Law y Kelton (1982).

Supongamos que hacemos n réplicas independientes con un método de simulación de terminación. Si cada una de las n simulaciones se inicia con las mismas condiciones iniciales y se ejecuta con una secuencia distinta de números aleatorios, entonces cada simulación se puede tratar como una réplica independiente. Por simplicidad, suponemos que sólo hay una medida de desempeño, representada por la variable X. Así, $X_j$ es el estimador de la medida de desempeño de la j-ésima réplica. Entonces, dadas las condiciones de las réplica (o simulaciones), la sucesión $X_1,..., X_n$ serán variables aleatorias iid. Con éstas, podemos usar el análisis estadístico clásico para formar un intervalo de confianza 100(1 -$\alpha$)% para $\theta$ = E(X) como sigue:

$$\overline{X}(n) \pm t_{(\alpha/2,n-1)}\sqrt{\frac{S^2(n)}{n}}$$

con

$$\overline{X}(n) = \sum_{i=1}^{n} \frac{X_i}{n}$$

$$S^2(n) = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}(n)\right)^2}{n-1}$$

$t_{(\alpha/2,n-1)}$ es $P(Y > t_{(\alpha/2,n-1)}) = \alpha$ , donde Y claramente sigue una distribución t-student.

# 4. MODELOS DE PREDICCIÓN

En este capítulo estudiamos dos tipos importantes de métodos de predicción: los métodos de la extrapolación y el de predicción causal. Los métodos de extrapolación que se usan para predecir valores futuros de una serie temporal a partir de valores en el pasado de otra serie temporal. Para dar un ejemplo veamos las ventas mensuales de televisores/ discos compactos (CD) y acondicionadores de aire (AA) de la empresa Lowland Appliance Company durante los últimos 24 meses que se ve en la Tabla 1. En un método de predicción por extrapolación se supone que los comportamientos y tendencias del pasado continuarán en los meses futuros. Así, los datos del pasado acerca de las ventas, sin ninguna información adicional, se usan para generar pronósticos para las ventas de aparatos electrodomésticos durante los meses futuros. Los métodos de extrapolación, a diferencia de los métodos de predicción causal no tienen en cuenta lo que "causó" los datos del pasado; tan sólo suponen que las tendencias y comportamientos del pasado continuarán en el futuro.

Los métodos de predicción causal tratan de pronosticar valores futuros de una variable, la variable dependiente, mediante datos del pasado para estimar la relación entre la variable dependiente y una o más variables independientes. Por ejemplo, Lowland debería tratar de pronosticar ventas mensuales futuras de acondicionadores de aire con datos del pasado para determinar cómo se relacionan las ventas de acondicionadores de aire con variables independientes, como el precio/ la publicidad y el mes del año.

Tabla 1 Ventas de Lowland Appliance

| MES | VENTAS DE TV | VENTAS DECD | VENTAS DEAA | MES | VENTAS DE TV | VENTAS DE CD | VENTAS DE AA |
|---|---|---|---|---|---|---|---|
| 1 | 30 | 40 | 13 | 13 | 38 | 79 | 36 |
| 2 | 3 | 47 | 7 | 14 | 30 | 82 | 21 |
| 3 | 30 | 50 | 23 | 15 | 35 | 80 | 47 |
| 4 | 39 | 49 | 32 | 16 | 30 | 85 | 81 |
| 5 | 33 | 56 | 58 | 17 | 34 | 94 | 112 |
| 6 | 33 | 53 | 60 | 18 | 40 | 89 | 139 |
| 7 | 34 | 55 | 90 | 19 | 36 | 96 | 230 |
| 8 | 38 | 63 | 93 | 20 | 32 | 100 | 201 |
| 9 | 36 | 68 | 63 | 21 | 40 | 100 | 122 |
| 10 | 39 | 65 | 39 | 22 | 36 | 105 | 84 |
| 11 | 30 | 72 | 37 | 23 | 40 | 108 | 74 |
| 12 | 36 | 69 | 29 | 24 | 34 | 110 | 62 |

Sean $x_1,x_2...,x_n$ valores observados de una serie temporal, donde $x_t$ es el valor de esa serie que se observa durante el periodo t. Uno de los métodos de predicción que más se usa es el de la media variable o móvil. Se define $f_{t,k}$ como el pronóstico para el periodo t + 1 que se hace después de observar $x_t$. Para el método de la media variable,

$f_{t,k}$ = media de las n observaciones últimas
= media de $x_{t-1},x_{t-2}...,x_{t-n-+1}$

donde n es un parámetro dado. ¡

Para ilustrar el uso del método de la media variable, seleccionamos n = 3 y usamos ese método para predecir las ventas de TV durante los primeros seis meses de los datos de la Tabla 1. Los cálculos necesarios se dan en la Tabla 2, Para los meses 1 a 3 no hemos observado todavía 3 meses de datos y, por lo tanto, para $j < 3$, no podemos dar un pronóstico de la media variable para las ventas de eso meses. Por ejemplo:

$f_{3,1}= (30+32+30)/3$

Tabla 2 Predicciones de la media variable (N = 3) para ventas de TV

| MES | VENTAS REALES | VENTAS PRONOSTICADAS |
|---|---|---|
| 1 | 30 | |
| 2 | 32 | |
| 3 | 30 | |
| 4 | 39 | 30.67 |
| 5 | 33 | 33.67 |
| 6 | 33 | 34.00 |

Nótese que de un periodo al siguiente, la predicción "progresa", reemplazando la observación más "antigua" en el promedio por la más reciente.

SELECCIÓN DE N

¿Cómo se debe seleccionar a N, el número de periodos que se usan para calcular el promedio progresivo? Para contestar esta pregunta, necesitamos definir una medida de la exactitud de predicción. Usamos la desviación absoluta media DAM, como medida. Antes de definirla, necesitamos definir el concepto de error de pronóstico. Dado un pronóstico para $x_j$, definimos a $e_j$ como el error en la predicción de $x_j$, como

$ej = | x_j - (\text{pronóstico de } x_j)|$

Según la Tabla 2, vemos que $e_4 =|39- 30.67| = 8.33$, $e_5 =| 33 - 33.67| = 0.67$, y $e_6 = |34 – 34| = 0$. La desviación absoluta media es tan sólo el promedio de los valores absolutos de todas las $e_j$. Así, para los periodos 1 a 6, la predicción de promedio progresivo da una desviación absoluta media DAM,

$DAM = ( e_4 + e_5 + e_6 )/3 = (8.33 + 0.67 + 0 )/3 = 3$

Así, en promedio, los pronósticos de ventas de TV tienen un error de 3 por mes.

Parece razonable seleccionar una N que reduzca la desviación absoluta media a un mínimo. Si aplicamos la técnica de la media variable a los 24 meses de ventas de la Tabla 1, encontramos las DAM de la Tabla 3. Así, un periodo de media variable de 3 ó 4 parece obtener las mejores predicciones de ventas de TV.

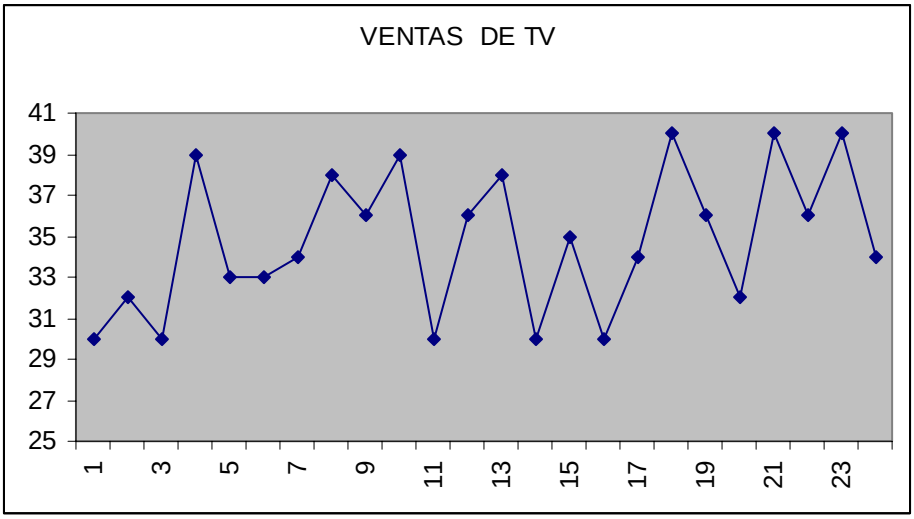
Tabla 3
desviaciones absolutas medias para pronósticos de ventas

| N | DAM |
|---|---|
| 2 | 3.21 |
| 3 | 2.78 |
| 4 | 2.79 |
| 5 | 2.99 |
| 6 | 3.27 |

Las predicciones con media variable trabajan bien para una serie temporal que varía alrededor de un nivel base constante. De la Fig. 1, parece que las ventas mensuales de TV fluctúan alrededor de un nivel base de 35. De manera formal, estos pronósticos trabajan bien si

(1) $x_t = b + \varepsilon_t$

Figura 1 Ventas de TV




En las Figuras 2 y 3 vemos que las ventas de discos compactos, CD, y de acondicionadores de aire no están bien representados en la Ec. (1). Vemos en la Fig. 2 que hay una tendencia ascendente en las ventas de CD y, por lo tanto, éstas no fluctúan con respecto a un nivel base. En la Fig. 3 vemos que las ventas de acondicionadores de aire presentan variaciones estacionales. Los máximos y los mínimos de la serie se repiten a intervalos regulares de 12 meses. En la Fig. 3 también se observa que las ventas de acondicionadores de aire presentan una tendencia ascendente. En los casos donde la tendencia, la variación estacional o ambas se manifiestan, en general el método de la media variable da malas predicciones. Para terminar esta sección tenga en cuenta que además de la tendencia y la variación estacional, una serie temporal puede presentar comportamiento cíclico. Por ejemplo, las ventas de automóviles siguen, con frecuencia, el ciclo de la economía nacional. El comportamiento cíclico es mucho más irregular que un patrón estacional y, frecuentemente es difícil de descubrir.

Figura 2 Ventas de CD

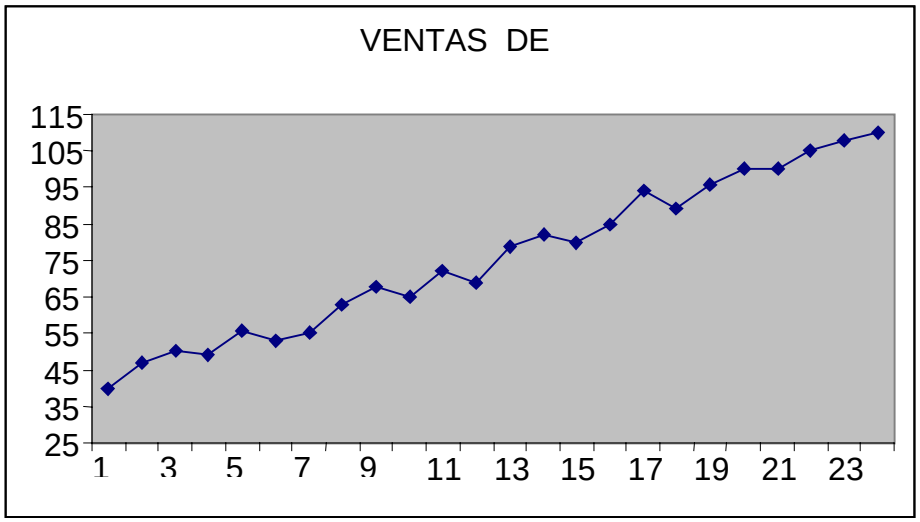
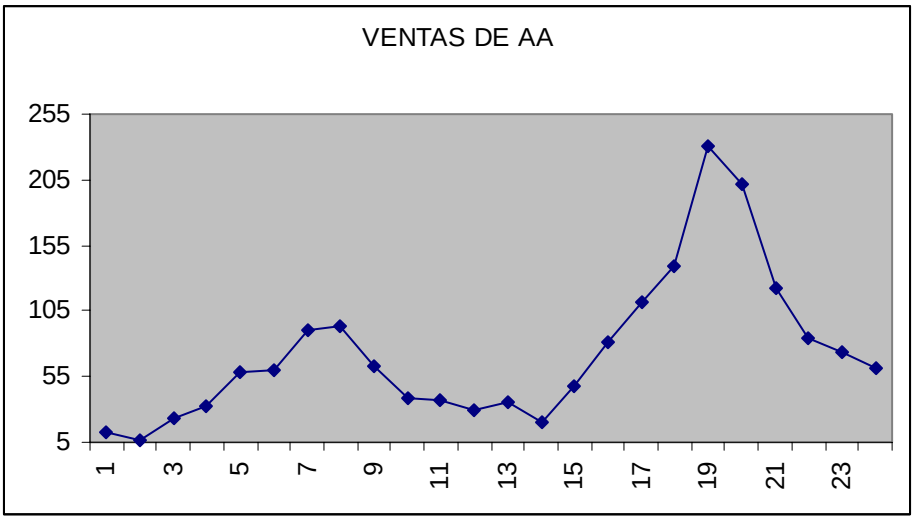



Figura 3. Ventas de Aires acondicionados




## 4.1 ATENUACIÓN EXPONENCIAL SIMPLE

Si una serie temporal fluctúa con respecto a un nivel base, se puede utilizar una atenuación exponencial simple para obtener buenos pronósticos de valores futuros de la serie. Para representar esta técnica, sea $A_t$ = promedio atenuado de una serie temporal después de observar $x_t$. $A_t$ es el pronóstico del valor de la serie temporal durante cualquier periodo futuro después de observar $x_t$. La ecuación clave de la atenuación exponencial simple es

$$A_t = \alpha x_t + (1-\alpha)A_{t-1} \quad (2)$$

En la Ecuación (2) $\alpha$ es la constante de atenuación que satisface $0 < \alpha < 1$. Para inicializar el procedimiento de predicción, antes de observar x1, debemos tener un valor de $A_0$. En general, hacemos que $A_0$ sea el valor observado del periodo inmediatamente anterior al periodo 1. Como en los pronósticos de media variable, sea $f_{t,k}$ el pronóstico para $x_{t+k}$ que se hizo al final del periodo t. Entonces

$$A_t = f_{t,k} \quad (3)$$

Si se supone que tratamos de predecir un periodo más adelante, el error de predicción de $x_t$ representado de nuevo por $e_t$, está dado por

$$e_t = | x_t - (\text{pronóstico de } x_t)| = | x_t - A_t | \quad (4)$$

Mostraremos la atenuación exponencial simple, con $\alpha = 0.1$, para los seis primeros meses de ventas de TV. Los resultados se dan en la Tabla 4. Suponemos que se vendieron 32 TV el último mes, de modo que empezamos el procedimiento con $A_0 = 32$.

Para los meses 1 a 6, la desviación absoluta media de nuestro pronóstico está dada por

DAM = (|-2|+|0.2|+ | -1.82|+|7.36| + |0.63+|1.56| ) /6 = 2.26

Tabla 4 Atenuación exponencial simple para ventas de TV

| MES | VENTAS REALES | PREDICCIÓN | ERROR |
|---|---|---|---|
| 1 | 30 | 32.00 | 2.00 |
| 2 | 32 | 31.80 | 0.20 |
| 3 | 30 | 31.82 | 1.82 |
| 4 | 39 | 31.64 | 7.36 |
| 5 | 33 | 32.37 | 0.63 |
| 6 | 33 | 32.44 | 0.56 |

Tabla 5 Desviación absoluta media para ventas de TV

| $\alpha$ | DAM |
|---|---|
| 0.05 | 3.20 |
| 0.10 | 3.04 |
| 0.15 | 2.94 |
| 0.20 | 2.89 |
| 0.25 | 2.88 |
| 0.30 | 2.90 |
| 0.35 | 2.94 |
| 0.40 | 2.98 |
| 0.45 | 3.05 |
| 0.50 | 3.13 |

Para todo el periodo de 24 meses podemos determinar, con hojas de cálculo, el valor de $\alpha$ que produzca la menor desviación absoluta media. En la Tabla 5 se presentan los resultados. El valor mínimo se alcanza para $\alpha = 0.222$.

# OBSERVACIONES

1. Como $\alpha < 1$, la atenuación exponencial "empareja" las variaciones en una serie temporal al no dar un peso total a la última observación.

2. Si $\alpha = 2/ n+1$ la atenuación exponencial simple, con parámetro $\alpha$ de atenuación, y un pronóstico de media variable de n periodos dan los mismos resultados. Por ejemplo, $\alpha = 1/3$ equivale aproximadamente a una media variable de cinco periodos.

3. Para ver porqué se le llama atenuación exponencial, veamos la Ec. (2) para t -1:

$$A_{t-1} = \alpha x_{t-1} + (1-\alpha)A_{t-2} \qquad (5)$$

Al sustituir la Ec. (5) en la Ec. (2), se obtiene

$$A_t = \alpha x_t + (1-\alpha)A_{t-1} = \alpha x_t + (1-\alpha)\alpha x_{t-1} + (1-\alpha)^2 A_{t-2} \qquad (6)$$

Observe que

$$A_{t-2} = \alpha x_{t-2} + (1-\alpha)A_{t-3} \qquad (7)$$

Al sustituir la Ec. (7) en la Ec. (6) se obtiene

$$A_t = \alpha x_t + (1-\alpha)A_{t-1} = \alpha x_t + (1-\alpha)\alpha x_{t-1} + (1-\alpha)^2\left[\alpha x_{t-2} + (1-\alpha)A_{t-3}\right]$$

Si se repite este proceso se obtiene

$$A_t = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \ldots + \alpha(1-\alpha)^k x_{t-k} + \ldots \qquad (8)$$

Como $\alpha + \alpha(1-\alpha) + \alpha(1-\alpha)^2 + \ldots = 1$, la Ec. (8) muestra que si regresamos un número "infinito" de periodos, el promedio atenuado actual es un promedio ponderado de todas las observaciones del pasado. El peso que se da a la observación de k periodos en el pasado disminuye en forma exponencial, por un factor (1 - $\alpha$). Mientras mayor sea el valor de $\alpha$ se da más peso a las observaciones más recientes. Por ejemplo, para $\alpha = 0.2$, las tres observaciones más recientes tienen 49% del peso (20%, 16% y 13%), en tanto que para $\alpha = 0.5$, las tres observaciones más recientes tienen 88% del peso (50%, 25% y 13%).

4. En la práctica se escoge $\alpha$ en general como 0.10, 0.30 o 0.50. Si el valor que minimiza la desviación absoluta media es mayor que 0.5, entonces es probable que haya tendencia, variaciones estacionales o comportamiento cíclico. y el método de atenuación exponencial simple no se recomienda como técnica de predicción. En esos casos se pueden tener, probablemente, mejores pronósticos con el método de Holt (atenuación exponencial con tendencia, que se analiza en la

siguiente sección, o el de Winter (atenuación exponencial con tendencia y variación estacional, que se estudia posteriormente.

5. Aun cuando no fluctúe una serie temporal con respecto a un nivel base constante, la atenuación exponencial simple puede dar buenos pronósticos. Si $x_t = m_t + \varepsilon_t$, y $m_t = m_{t-1} + \delta_t$,. donde $\varepsilon$ y $\delta$ son términos independientes de error, cada uno con media 0, entonces la atenuación exponencial simple dará buenos pronósticos. Esto significa que si la demanda media $m_t$ de un producto está variando al azar con respecto al tiempo, la atenuación exponencial sencilla puede dar buenas predicciones de la demanda de un producto.

## 4.2 MÉTODO DE HOLT: ATENUACIÓN EXPONENCIAL CON TENDENCIA

Si creemos que una serie temporal presenta una tendencia lineal, sin variación estacional, el método de Holt con frecuencia da buenos pronósticos. Al final del t-ésimo periodo el método de Holl genera una estimación del nivel base, $L_t$, y de la tendencia por periodo, $T_t$ de la serie. Por ejemplo, suponga que $L_{20} = 20$ y que $T_{20} = 2$. Esto quiere decir que después de observar a $x_{20}$ creemos que el nivel base de la serie es 20 y que aumenta dos unidades por período. Así, dentro de cinco periodos calculamos que el nivel base de la serie será 30.

Después de observar a $x_t$, se usan las Ecuaciones. (9) y (10) para actualizar las estimaciones de la base y la tendencia. $\alpha$ y $\beta$ son constantes de atenuación, cada una con valores entre 0 y 1.

$$L_t = \alpha x_t + (1-\alpha)(L_{t-1} + T_{t-1}) \qquad (9)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \qquad (10)$$

Como antes, definimos $f_{t,k}$ como el pronóstico de $x_t$ hecho al final del periodo t.

Entonces

$$f_{t,k} = L_t + kT_t \qquad (11)$$

Para empezar con el método de Holt, necesitamos una estimación inicial, $L_0$ de la base y la otra estimación, $T_0$, de la tendencia. Haríamos a $T_0$ al aumento promedio mensual en la serie temporal durante el año anterior, y que $L_0$ fuera igual a la observación del ultimo mes.

En la Fig. 2 se ve que las ventas de CD (discos compactos) presentan una tendencia ascendente, pero no se. observa variación estacional obvia. Por lo tanto, el método de Holt debe dar una buena predicción. Supongamos que las ventas de CD durante cada uno de los últimos doce meses son 4, 6, 8, 10, 14, 18, 20, 22, 24, 28, 31 y 34. Entonces,

$T_0 = [(6-4) + (8-6) + ... + (34-31)] /11 = (34-4)/11 = 2.73$

Y, a continuación, calculamos $L_0 = 34$.

Aplicando el método de Holt a los primeros seis meses de ventas de discos compactos, con $\alpha = 0.30$ y $\beta = 0.10$, obtenemos los resultados que se presentan en la Tabla 6.

Para los primeros seis meses de ventas de discos compactos, calculamos la desviación absoluta media:

DAM = (3.27 + 6.47 + 4.51 + 1.00 + 3.18 + 4.00)/6 = 3.74

Para el periodo completo de 24 meses vemos que DAM = 2.85.

Tabla 6 Método de Holt para ventas de discos compactos

| MES | VENTAS REALES | Lt | Tt | PREDICCIÓN | ERROR |
|---|---|---|---|---|---|
| | | 34.00 | 2.73 | | |
| 1 | 40 | 37.71 | 2.83 | 36.73 | 3.27 |
| 2 | 47 | 42.47 | 3.02 | 40.53 | 6.47 |
| 3 | 50 | 46.85 | 3.15 | 45.49 | 4.51 |
| 4 | 49 | 49.70 | 3.12 | 50.00 | 1.00 |
| 5 | 56 | 53.78 | 3.22 | 52.82 | 3.18 |
| 6 | 53 | 55.80 | 3.10 | 57.00 | 4.00 |

Si probamos varias combinaciones de $\alpha$ y $\beta$ podríamos determinar los valores de tales parámetros que minimizar la desviación absoluta media. Si tanto el valor de $\alpha$ como de $\beta$ no son menores de 0.5, entonces es probable que haya comportamiento cíclico o variación estacional, y se debería aplicar otro método de predicción.

En resumen, el método de Holt da buenos pronósticos para una serie con tendencia lineal. Esa serie se puede modelar como $x_t = a + bt + \varepsilon_t$, siendo a = nivel base al iniciar el periodo 1, b = tendencia por periodo y $\varepsilon_t$ el término de error para el periodo t.

# 4.3 EXPONENCIAL CON VARIACIÓN ESTACIONAL

Lo que se llama, correctamente, método de Winter se usa para predecir series temporales en las que se encuentren presentes tendencia y variación estacional. Como se mencionó antes, en la Fig. 3 se ve que las ventas de acondicionadores de aire presentan una tendencia ascendente y, al mismo tiempo, variación estacional. Por lo tanto, el método de Winter es candidato lógico para pronosticar esas ventas.

Para explicar el método de Winter necesitamos dos definiciones. Sea c = número de periodos en la duración del comportamiento estacional (c == 4 para datos trimestrales y c = 12 para datos mensuales). Sea $s_t$ una estimación de un factor estacional multiplicativo para el mes t, que se obtiene después de observar a $x_t$. Por ejemplo, suponga que el mes 7 es julio y que $s_7 = 2$. Entonces, después de observar las ventas de acondicionadores de aire del mes 7, creemos que las ventas de acondicionadores de aire en julio serán, en igualdad de circunstancias, iguales al doble de las ventas esperadas durante un mes promedio. Si el mes 24 es diciembre, y $s_{12} = 0{,}4$, entonces, después de observar las ventas del mes 24, podemos predecir que las ventas de acondicionadores de aire en diciembre serán el 40% de las ventas esperadas durante un mes promedio. En los cálculos siguientes, $L_t$ y $T_t$ tienen el mismo significado que en el método de Holt. Cada periodo, $L_t$, $T_t$ y $s_t$ se actualizan, en ese orden, mediante las Ecuaciones (12) a (14). Nuevamente $\alpha$, $\beta$ y $\gamma$ son constantes de atenuación, cada una de las cuales está entre 0 y 1:

$$L_t = \alpha \frac{x_t}{s_{t-c}} + (1-\alpha)(L_{t-1} + T_{t-1}) \qquad (12)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \qquad (13)$$

$$s_t = \gamma \frac{x_t}{L_t} + (1-\gamma)s_{t-c} \qquad (14)$$

La ecuación (12) actualiza la estimación de la base de la serie calculando el promedio ponderado de las dos cantidades siguientes:

1. $(L_{t-1} + T_{t-1})$ que es nuestra estimación de nivel base antes de observar a $x_t$.

2. La observación $\frac{x_t}{s_{t-c}}$ sin variación estacional, que es una estimación de la base obtenida partir del periodo actual.

La Ecuación (13) es idéntica a la (10) para $T_t$ que se usó para actualizar la tendencia en el método de Holt.

La Ecuación (14) actualiza la estimación de la variación estacional del mes t, calculando un promedio ponderado de las dos cantidades siguientes:

1. La estimación más reciente de la variación estacional $s_{t-c}$ del mes t.

y

2. $\frac{x_t}{L_t}$ que es una estimación de la variación estacional, calculada para el mes actual. '

Al final del periodo t, el pronóstico $f_{t,k}$ para el mes k es

$$f_{t,k} = (L_t + kT_t)s_{t+k-c} \qquad (15)$$

Así, para pronosticar el valor de la serie durante el periodo t + k, multiplicamos la estimación de la base de ese periodo ($L_t + kT_t$) por la estimación más reciente del factor de variación estacional $s_{t+k-c}$ del mes (t + k).

## 4.4 INTRODUCCIÓN AL MÉTODO DE WINTER

Para obtener buenos pronósticos con el método de Winter debemos contar con buenas estimaciones iniciales de base, tendencia y factores estacionales. Sean

$L_0$ = estimación de la base al inicio del mes 1
$T_0$ = estimación de la tendencia al inicio del mes 1
$s_{-11}$ = estimación del factor estacional de enero al inicio del mes 1 (16)
$s_{-10}$ = estimación del factor estacional de febrero al inicio del mes 1
.
.
.
$s_0$ = estimación del factor estacional de diciembre al inicio del mes 1

Hay diversos métodos para estimar los parámetros en (16), Escogeremos uno sencillo que necesita dos años de datos. Suponga que los dos años últimos de ventas mensuales fueron los siguientes:

Año -2: 4,3,10,14,25,26,38,40, 28,17,16,13
Año -1; 9,6,18,27,48,50,75,77,52,33,31,24

Ventas totales durante el año -2 = 234
Ventas totales durante el año -1 = 450

Estimamos $T_0$ mediante la siguiente ecuación:

$$\frac{(\text{Ventas mensuales promedio durante el año -1}) - (\text{Ventas mensuales promedio durante el año - 2})]}{12}$$ o sea :

$$\frac{\frac{450}{12} - \frac{234}{12}}{12} = 1.5$$

Para estimar a L0 determinamos primero la demanda mensual promedio durante el año - 1, que es 450/12. Con ello se estima la base a mitad del año - 1, o sea, el mes 6.5 de ese año. Para llevar esa estimación al final del mes 12 del año - 1, sumamos $(12\text{-}6.5)T_0 = 5.5T_0$. Así, la estimación de $L_0$ es $37.5 + 5.5(1.5) = 45.75$.

Para estimar el factor de variación estacional para un mes dado (por ejemplo, en enero $s_{-11}$ ), calculamos una estimación de la variación estacional para enero en el año - 2 y en el año - 1, y los promediamos. En el ano - 2 la demanda mensual promedio fue 234/12 = 19.5; en enero de ese año se vendieron 4 acondicionadores de aire. Por lo tanto

Estimación de la variación estacional para enero del año -2 = 4/19.5 = 0.205

Igualmente,

Estimación de la variación estacional para enero del año -1 = 9/37.5 = 0.240

Por último, calculamos que $s_{-11}$= (.205+.240)/2 = 0.22. De igual modo obtenemos el resto de valores:

Tabla 7. Calculo de los factores estacionales

| Referencia | Año -2 | Año -1 | s -2 | s-1 | s |
|---|---|---|---|---|---|
| -11 | 4 | 9 | 0.2051 | 0.2400 | 0.222564 |
| -10 | 3 | 6 | 0.1538 | 0.1600 | 0.156923 |
| -9 | 10 | 18 | 0.5128 | 0.4800 | 0.496410 |
| -8 | 14 | 27 | 0.7179 | 0.7200 | 0.718974 |
| -7 | 25 | 48 | 1.2821 | 1.2800 | 1.281026 |
| -6 | 26 | 50 | 1.3333 | 1.3333 | 1.333333 |
| -5 | 38 | 75 | 1.9487 | 2.0000 | 1.974359 |
| -4 | 40 | 77 | 2.0513 | 2.0533 | 2.052308 |
| -3 | 28 | 52 | 1.4359 | 1.3867 | 1.411282 |
| -2 | 17 | 33 | 0.8718 | 0.8800 | 0.875897 |
| -1 | 16 | 31 | 0.8205 | 0.8267 | 0.823590 |
| 0 | 13 | 24 | 0.6667 | 0.6400 | 0.653333 |
| TOTAL | 234 | 450 | | | |
| PROMEDIO | 19.5 | 37.5 | | | |

En la siguiente tabla se muestra los resultados de la aplicación del método de Winter, para $\alpha = 0.5$, $\beta=0.4$ y $\gamma =0.6$

Tabla 8 Método de Winter para ventas acondicionadores de aire

| | | $L_t$ | $T_t$ | $s_t$ | $f_{t-1,1}$ | Error |
|---|---|---|---|---|---|---|
| MES | Ventas | 45.75 | 1.5 | | | |
| 1 | 13 | 52.8301 | 3.7320 | 0.2367 | 10.5162 | 2.4838 |
| 2 | 7 | 50.5850 | 1.3412 | 0.1458 | 8.8759 | 1.8759 |
| 3 | 23 | 49.1294 | 0.2225 | 0.4795 | 25.7767 | 2.7767 |
| 4 | 32 | 46.9299 | (0.7463) | 0.6967 | 35.4827 | 3.4827 |
| 5 | 58 | 45.7299 | (0.9278) | 1.2734 | 59.1623 | 1.1623 |
| 6 | 60 | 44.9010 | (0.8882) | 1.3351 | 59.7361 | 0.2639 |
| 7 | 90 | 44.7986 | (0.5739) | 1.9951 | 86.8971 | 3.1029 |
| 8 | 93 | 44.7698 | (0.3559) | 2.0673 | 90.7628 | 2.2372 |
| 9 | 63 | 44.5271 | (0.3106) | 1.4134 | 62.6806 | 0.3194 |
| 10 | 39 | 44.3711 | (0.2487) | 0.8777 | 38.7291 | 0.2709 |
| 11 | 37 | 44.5238 | (0.0882) | 0.8280 | 36.3387 | 0.6613 |
| 12 | 29 | 44.4117 | (0.0977) | 0.6531 | 29.0313 | 0.0313 |

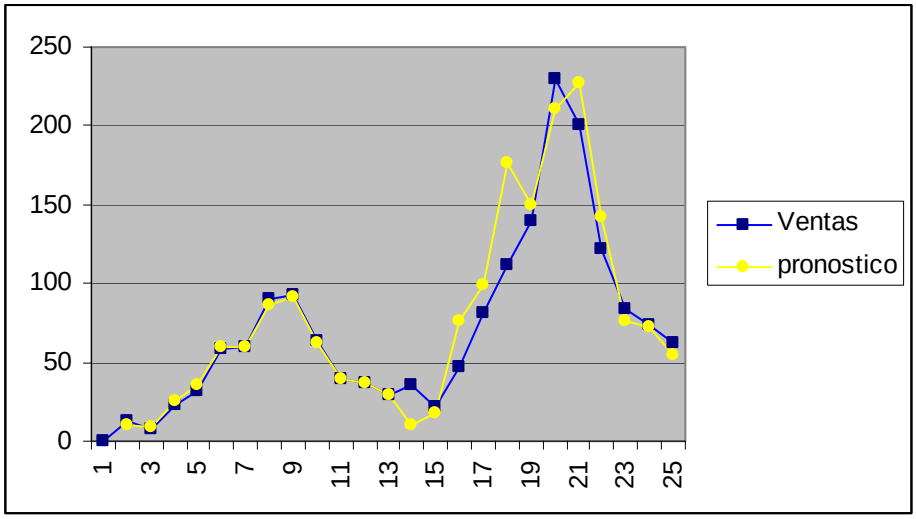
# OBSERVACIONES

1. Como en el método de Winter se usan tres constantes de atenuación, es difícil encontrar la combinación de valores que den como resultado la desviación absoluta media mínima, sin embargo se puede busca un método de optimización para un conjunto de valores observados.

2. Aunque los valores de $\alpha$ y $\beta$ que minimizan la DAM no deberían ser mayores que 0.5, como en el método de Holt, no es raro que el mejor valor de y sea mayor que 0.5. Esto se debe a que para datos mensuales, cada factor estacional mensual se actualiza durante sólo un doceavo de los periodos. Como los factores de variación-estacional se actualizan sin tanta frecuencia, quizá necesitemos dar mayor peso a cada observación, así que $\gamma > 0.5$ se puede admitir.

3. En la Fig. 4 se muestra qué tan similares son los pronósticos de ventas de acondicionadores de aire (para $\alpha = 0.5$, $\beta=0.4$ y $\gamma =0.6$) con las ventas reales. La concordancia entre valores predichos y ventas reales es bastante buena, excepto para los meses 15 y 17. Durante esos meses nuestros pronósticos son demasiado altos. Quizá se hayan contratados vendedores nuevos en esos meses, haciendo que las ventas fueran menores que las que se predijeron.

Figura 4
Pronósticos de ventas de acondicionadores de aire




# EXACTITUD DE PRONÓSTICOS

Para cualquier modelo de pronóstico en el que los errores estén distribuidos normalmente, podemos aplicar la desviación absoluta media para estimar Se = desviación estándar de nuestros errores de pronóstico. La relación entre la DAM y Se está dada por

$$s_e = 1.25 \ \text{DAM}$$

Si se supone que los errores se distribuyen normalmente, sabemos que un 68% de las predicciones debe quedar dentro de $s_e$ del valor real, y que aproximadamente el 95% de las predicciones deben estar a menor distancia que $2s_e$ del valor real. Así, para las ventas de acondicionadores de aire, vemos que $s_e$ = 1.25(10.48) = 13.10. Por lo tanto, debemos esperar que durante 0.68(24) = 16 meses, de los 24, nuestras predicciones de ventas tendrán un error máximo de 13.10 acondicionadores, y durante 0.95(24) = 23 o 24 meses de los 24, los pronósticos tengan un error máximo de 2(13.10) = 26.2 acondicionadores. En realidad, nuestras predicciones de ventas tienen una exactitud dentro de 13.10 durante 17 meses, y una exactitud dentro de 26.2 durante 22 meses.

Notemos que en la mayor parte de los casos en los que se requiere un pronóstico, el conocimiento acerca de la exactitud probable del pronóstico es casi tan importante como el pronóstico mismo. Por lo tanto, esta corta subsección es muy importante.

# 4.5 REGRESIÓN LINEAL SIMPLE

Con frecuencia, tratamos de predecir el valor de una variable, la variable dependiente, a partir del valor de otra variable, la variable independiente. A continuación presentamos algunos ejemplos:

| **Variable dependiente** | **Variable independiente** |
|---|---|
| Ventas de un producto | Precio del producto |
| Ventas de automóviles | Tasa de interés |
| Costo total de producción | Unidades producidas |

Si la variable dependiente y la independiente se relacionan en forma lineal, se puede aplicar la regresión lineal para estimar esa relación.

Para dar un ejemplo de regresión lineal simple, suponga que Giapetto produce trenes y soldados; el costo depende de la cantidad producida. Para formular este problema necesitamos calcular el costo de producción de un soldado y el de un tren. Suponga que deseamos determinar el costo de producción de un tren. Para estimarlo, hemos observado durante diez semanas el número de trenes producidos cada semana y el costo total incurrido en producirlos. Esta información se presenta en la Tabla 10.

Tabla 9 Datos semanales costos de trenes

| SEMANA | FRENES PRODUCIDOS | COSTO DE PRODUCCIÓN DE TRENES (dólares) |
|---|---|---|
| 1 | 10 | 257.40 |
| 2 | 20 | 601.60 |
| 3 | 30 | 782.00 |
| 4 | 40 | 765.40 |
| 5 | 45 | 895.50 |
| 6 | 50 | 1,133.00 |
| 7 | 60 | 1,152.80 |
| 8 | 55 | 1,132.70 |
| 9 | 70 | 1,459.20 |
| 10 | 40 | 970.10 |

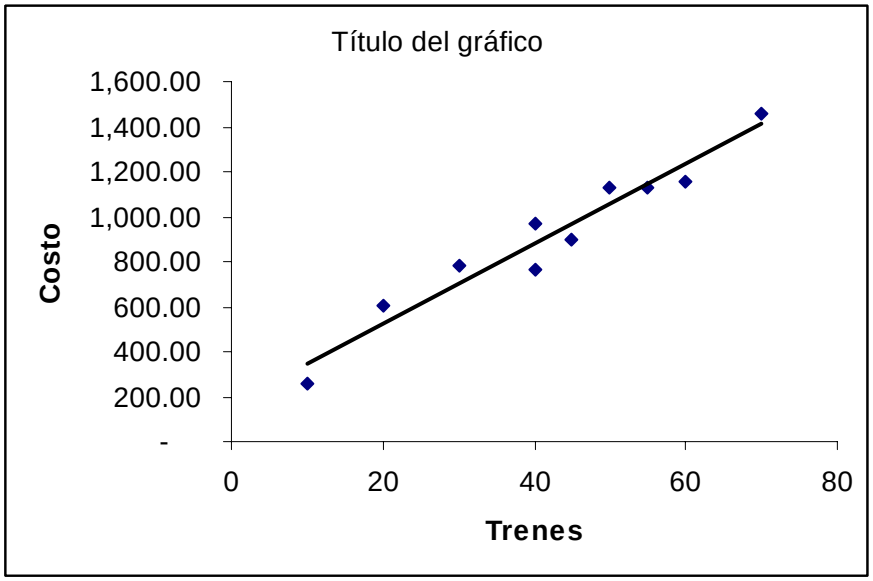
Los datos de la Tabla 9 aparecen graficados en la Fig. 5. Observe que parece haber una marcada relación lineal entre $x_i$, número de trenes producidos durante la semana i y $y_i$, costo de producirlos, también durante la semana i. La recta que se gráfica en la Fig. 5 parece acercarse de una manera que explicamos luego, a la representación de la relación lineal entre unidades producidas y costos de producción. Pronto veremos cómo se escogió esa línea.

Para empezar, modelamos la relación lineal entre $x_i$ y $y_i$ mediante la siguiente ecuación:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (17)$$

en la cual $\varepsilon_i$ es un término de error que representa el hecho que durante una semana en la que se producen $x_i$ trenes, el costo de producción podría no siempre ser igual a $\beta_0 + \beta_1 x_i$.

Figura 5 Diagrama de dispersión del costo de producción de trenes




Se desconocen los valores verdaderos de $\beta_0$ y $\beta_1$. Para encontrar estos valores, primero definimos a $e_i$ como error o residuo para el punto dato i: $y_i - (\beta_0 + \beta_1 x_i)$ de esta forma $\beta_0$ y $\beta_1$ más adecuados serán aquellos que minimicen $\sum e_i^2$, de esta forma podemos definir $f(\beta_0, \beta_1) = \sum e_i^2 = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$

Para encontrar el mínimo $\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} = 0$, lo cual arroja como resultado que los valores $\beta_0$ y $\beta_1$ deben ser

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (18)$$

Donde $\bar{x}$ es el promedio de las $x_i$ y $\bar{y}$ es el promedio de las $y_i$

A $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ se le llama línea de regresión de los cuadrados mínimos. En esencia, si esta línea se ajusta bien a los puntos (recta mejor ajuste).

Tabla 10 calculo de $\beta_0$ y $\beta_1$ para el ejemplo de los trenes

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 10 | 257.40 | (32) | (658) | 21042.24 | 1,024 |
| 20 | 601.60 | (22) | (313) | 6894.14 | 484 |
| 30 | 782.00 | (12) | (133) | 1595.64 | 144 |
| 40 | 765.40 | (2) | (150) | 299.14 | 4 |
| 45 | 895.50 | 3 | (19) | -58.41 | 9 |
| 50 | 1,133.00 | 8 | 218 | 1744.24 | 64 |
| 60 | 1,152.80 | 18 | 238 | 4280.94 | 324 |
| 55 | 1,132.70 | 13 | 218 | 2830.49 | 169 |
| 70 | 1,459.20 | 28 | 544 | 15238.44 | 784 |
| 40 | 970.10 | (2) | 55 | -110.26 | 4 |

Con $\bar{x} = 42$ y $\bar{y} = 914.97$

$$\hat{\beta}_1 = \frac{53{,}756.60}{3010} = 17.86, \qquad \hat{\beta}_0 = 914.97 - (17.86 * 42) = 164.88$$

Tabla 11 Cálculo de errores

| $x_i$ | $y_i$ | $\bar{y}_i$ | $e_i$ |
|---|---|---|---|
| 10 | 257.40 | 343.47 | (86.07) |
| 20 | 601.60 | 522.06 | 79.54 |
| 30 | 782.00 | 700.66 | 81.34 |
| 40 | 765.40 | 879.25 | (113.85) |
| 45 | 895.50 | 968.55 | (73.05) |
| 50 | 1,133.00 | 1,057.84 | 75.16 |
| 60 | 1,152.80 | 1,236.44 | (83.64) |
| 55 | 1,132.70 | 1,147.14 | (14.44) |
| 70 | 1,459.20 | 1,415.03 | 44.17 |
| 40 | 970.10 | 879.25 | 90.85 |

Toda línea de cuadrados mínimos tiene dos propiedades:

1. Pasa por el punto $(\bar{x}, \bar{y})$. Así, durante una semana en la que Giapetto produce 42 trenes, pronosticaríamos que esos trenes tendrían un costo de producción de 914.17 dólares.

2. La línea de cuadrados mínimos "divide" a los puntos, en sentido de que la suma de las distancias verticales desde los puntos arriba de la línea a la recta de los cuadrados mínimos es igual a la suma de las distancia verticales desde los puntos abajo de la recta de los cuadrados mínimos a dicha recta; ambas distancias se miden a partir de la recta misma.

¿QUÉ TAN BUENO ES EL AJUSTE?

¿Cómo determinamos lo bien o mal que ajusta la recta de cuadrados mínimos a los puntos de los datos? Para contestar esa pregunta necesitamos describir tres componentes de la variación: la suma de cuadrados totales (SST), la suma de cuadrados de error es (SSE) y la suma de regresión de cuadrados (SSR). La suma cuadrados totales es SST = $\sum (y_i - \bar{y})^2$ . Mide la variación total de y, con respecto al promedio. La suma de cuadrados de errores está dada por SSE = $\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$ . Si la recta de cuadrados mínimos pasara por todos los puntos de datos, SSE = 0. Así, si la SSE es pequeña indica que la línea de cuadrados mínimos se ajusta bien los datos. Definimos a la suma regresión de cuadrados como SSR = $\sum (\hat{y}_i - \bar{y})^2$ se puede demostrar que

$$SST = SSR + SSE \qquad (19)$$

Nótese que SST es función sólo de los valores de y. Para un buen ajuste, SSE debe ser pequeño y entonces, la Ec. (19) muestra que SSR será grande para tener un buen ajuste. De manera formal, podemos definir el coeficiente de determinación $R^2$ de y mediante

$R^2 = \frac{SSR}{SST}$ = porcentaje de la variación de y que queda explicado por x

También, la Ecuación (19) permite escribir

$1 - R^2 = \frac{SSE}{SST}$ = porcentaje de variación de y que no explica la variación de x

Para el ejemplo desarrollado se tiene que SST = 1,021,762 y SSE = 61,705. Entonces la Ec. 19 resulta SSR = SST - SSE = 960,057. Así vemos que $R^2$ = 0.94. Esto quiere decir que el número de trenes producidos durante un semana explica el 94% de la variación del costo semanal de producción. Todos 1os demás factores combinados pueden explicar cuando mucho el 6% de la variación y entonces podemos estar bastante seguros de que la relación lineal entre *x* y *y* es fuerte.

Una medida de la relación lineal entre x y y es la correlación lineal de muestra $r_{xy}$. Una correlación de muestra cercana a +1 indica una relación lineal positiva fuerte entre $_x$ y $_y$; una correlación de muestra cercana a -1 indica relación lineal negativa, y una correlación cercana a 0 indica una relación lineal débil entre x y y.

**EXACTITUD DE PREDICCIÓN**

Una medida de la exactitud de las predicciones obtenidas con regresión es el error estándar de la estimación ($s_e$). Si hacemos que n = número de observaciones, $s_e$ es

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

Para nuestro ejemplo,

$$s_e = \sqrt{\frac{61{,}705}{10-2}} = 87.8$$

En general es cierto que un 68% de los valores de y quedarán dentro de una distancia $s_e$ del valor predicho y, y que el 95% de los valores quedaran dentro de los márgenes $2s_e$ del valor predicho $\hat{y}$ [1]. En este ejemplo, esperamos que el 68% de las estimaciones de costo queden dentro de 87.80 dólares del costo real y 95%, dentro de 175.60 dólares del costo real. En realidad, para el 80% de los datos, el costo real está a menos de $s_e$ del costo pronosticado y para el 100% de los datos, el costo real está a menos de $2s_e$ del costo predicho.

Cualquier observación para la cual y no quede dentro de $2s_e$ de $\hat{y}$ se llama externa. Los puntos externos representan datos extraños que se deben revisar con cuidado. Naturalmente, si un punto externo es el resultado de un error al anotar los datos, se debe corregir. Si un dato externo es, hasta cierto punto, no característico de los demás puntos de datos, será mejor omitirlo y recalcular la recta de cuadrados mínimos. Como todos los errores son menores que $2s_e$ en valor absoluto, en el ejemplo de costo no hay puntos externos.

Se puede utilizar una t-student para crear un intervalo de confianza 1-$\alpha$, para un valor x:

$$t(\alpha/2, n-2) \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})}{\sum (x_i - \bar{x})^2}} \qquad (20)$$

donde $t(\alpha/2, n-2)$, es el valor de una t-student con extremos $\alpha/2$ y con n-2 grados de libertad.

[1] En realidad, 68% de los puntos debe quedar dentro de $s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})}{\sum (x_i - \bar{x})^2}}$ de $\hat{y}$ y un 95% dentro de $2s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})}{\sum (x_i - \bar{x})^2}}$

# 4.6 REGRESIÓN LINEAL MÚLTIPLE

En muchos casos podría ser útil usar más de una variable independiente para predecir el valor de una variable dependiente. En estos casos aplicamos la regresión múltiple. Por ejemplo, si tratamos de predecir las ventas mensuales de una cadena nacional de restaurantes de bocadillos de pollo, tendríamos que tener en cuenta el uso de las siguientes variables independientes: ingreso nacional, precio del pollo, dólares gastados en publicidad durante el mes y dólares gastados en publicidad durante el mes anterior.

Suponga que usamos k variables independientes para predecir la variable dependiente *y* y que tenemos n puntos de datos de la forma $(y_i, x_{1i}, x_{2i}, \cdots, x_{ki})$ donde $x_{ji}$ = valor de la j-ésima variable independiente para el í-ésimo punto dato y $y_i$ = valor de la variable dependiente para el i-ésimo punto de dato. En la regresión múltiple lineal, modelamos la relación mediante

$$y_i, = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad (21)$$

en la que $\varepsilon_i$ es un término de error con media 0 que representa el hecho de que el valor real de $y_i$ puede no ser igual a $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$.

El calculo de $\beta_j$ es similar al caso de una sola variable se definimos a $e_i$ como error o residuo para el punto dato i: $y_i - \hat{y}_i)$ de esta forma $\beta_0, \beta_1, \cdots, \beta_k$ más adecuados serán aquellos que minimicen $\sum e_i^2$, de esta forma podemos definir $f(\beta_0, \beta_1, \cdots, \beta_k) = \sum e_i^2$ luego procedemos a obtener el mínimo de forma análoga:

$\frac{\partial f(\beta_0, \cdots, \beta_k)}{\partial \beta_0} = \cdots = \frac{\partial f(\beta_0, \cdots, \beta_k)}{\partial \beta_k} = 0$, luego se obtiene que:

$$\begin{aligned} 2\sum \left(y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}\right) \cdot -1 = 0 \\ 2\sum \left(y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}\right) \cdot -x_{1i} = 0 \\ \vdots \\ 2\sum \left(y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}\right) \cdot -x_{ki} = 0 \end{aligned} \qquad (21)$$

Esas ecuaciones multiplicadas por ½ y agrupadas termino por termino nos brinda el siguiente sistema de k+1 ecuaciones lineales:

$$\begin{aligned}
&\beta_0 n + \beta_1 \sum x_{1i} + \cdots \beta_k \sum x_{ki} = \sum y_i \\
&\beta_0 \sum x_{1i} + \beta_1 \sum (x_{1i})^2 + \cdots \beta_k \sum x_{ki} x_{1i} = \sum y_i x_{1i} \\
&\qquad\qquad \vdots \\
&\beta_0 \sum x_{ki} + \beta_1 \sum x_{1i} x_{ki} + \cdots \beta_k \sum (x_{ki})^2 = \sum y_i x_{ki}
\end{aligned} \qquad (22)$$

Al despejar el sistema se encuentran los valores apropiados de los $\beta_0, \beta_1, \cdots, \beta_k$.

Al igual que el caso de una variable se definen de la misma forma la suma de cuadrados totales (SST), la suma de cuadrados de error es (SSE), la suma de regresión de cuadrados (SSR), $R^2 = \frac{SSR}{SST} =$ porcentaje de la variación de y que queda explicado por la k variables y se define el error estándar de la estimación ($s_e$):

$$s_e = \sqrt{\frac{SSE}{n-k-1}} \quad (23)$$

y se puede utilizar una t-student para crear un intervalo de confianza 1-α, para un valor x:

$$t(\alpha/2, n-k-1) \cdot s_e \sqrt{1 + \tfrac{1}{n} + \tfrac{(x-\bar{x})}{\sum (x_i - \bar{x})^2}} \qquad (24)$$

donde $t(\alpha/2, n-k-1)$, es el valor de una t-student con extremos α/2 y con n-k-1 grados de libertad.

# ANEXO A: PRUEBAS DE ALGUNOS RESULTADOS

**Sección A**

Se define $\rho_{xy} = P_x(T_y < \infty)$, que es la probabilidad de alcanzar el estado y, comenzando del estado x, para un tiempo $T_y$ que depende del estado y.

Se define la variable indicadora $1_y(z)$, para z perteneciente al conjunto de estados como:

$$1_y(z) = \begin{cases} 1, z = y \\ 0, z \neq y \end{cases}$$

Sea N(y) en número de veces n ≥ 1, que la cadena está en el estado y. A partir que $1_y(X_n)=1$ si la cadena se encuentra en el estado y en el momento n y $1_y(X_n)= 0$, de otro forma, se puede ver que

$$\text{(a1)} \quad N(y) = \sum_{n=1}^{\infty} 1_y(X_n)$$

El evento {N(y) ≥ 1 } es el mismo evento que $\{T_y < \infty\}$, de esta forma tenemos que

$$P[\{N(y) \geq 1\}] = P_x[\{T_y < \infty\}] = \rho_{xy}$$

Sean n,m dos enteros positivos. La probabilidad que x primero visite y en el instante m y luego visite y n pasos o unidades de tiempo posterior esta dado por $P_x(T_y=m)\ P_y(T_y=n)$, así

$$P_x(N(y) \geq 2) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} P_x(T_y = m)P_y(T_y = n) = \sum_{m=1}^{\infty} P_x(T_y = m)\sum_{n=1}^{\infty} P_y(T_y = n) = \rho_{xy}\rho_{yy}$$

De forma similar tenemos que (a2) $P_x(N(y) \geq m) = \rho_{xy}\ \rho_{yy}^{m-1} \quad m \geq 1$

Luego $P_x(N(y) = m) = P_x(N(y) \geq m) - P_x(N(y) \geq m+1)$ y por (a2) se tiene que

$$\text{(a3)} \quad P_x(N(y) = m) = \rho_{xy}\rho_{yy}^{m-1}(1-\rho_{yy}) \quad m \geq 1$$

y de esta forma:

$$\text{(a4)} \quad P_x(N(y) = 0) = 1 - P_x(N(y) \geq 1) = 1 - \rho_{xy}$$

Se define $E_x()$ como el valor esperado de la cadena de Markov comenzando en x, así tenemos que

$$\text{(a5)} \quad E_x(1_y(X_n)) = P_x(X_n = y) = P_{xy}(n)$$

Empleando (a1) y (a5) tenemos que $E_x(N(y)) = E_x(\sum_{n=1}^{\infty} 1_y(X_n)) = \sum_{n=1}^{\infty} E_x(1_y(X_n)) = \sum_{n=1}^{\infty} P_{xy}(n)$, de esta forma definiremos G(x,y) = $E_x(N(y))$, como el número de veces que se visita a y, comenzando de x.

Teorema 1

i) Sea y un estado transitorio, luego $P_x(N(y) < \infty) = 1$ y además:

(a6) $G(x,y) = \frac{\rho_{xy}}{1-\rho_{xy}}$, lo cual es finito para todo estado x

ii) Sea y un estado recurrente $P_x(N(y) = \infty) = 1$ y $G(x,y) = \infty$.

También,(a7) $P_x(N(y) = \infty) = P_x(T_y < \infty) = \rho_{xy}$.

Si $\rho_{xy} = 0$, luego G(x,y)=0, pero si $\rho_{xy} > 0$, luego $G(x,y) = \infty$.

Este teorema muestra de forma las propiedades de un estado transitorio a uno recurrente. En primer caso, el número de visitas que se realice es finito, independientemente de donde se comience. Mientras en el segundo caso el número de visitas es infinito, independientemente de donde se comience.

Prueba:

Sea y un estado transitorio. A partir del hecho que $0 \le \rho_{xy} \le 1$, se tiene que:

$$P_x(N(y) = \infty) = \lim_{m\to\infty} P_x(N(y) \ge m) = \lim_{m\to\infty} \rho_{xy}\ \rho_{xy}^{m-1} = 0$$

por (a3) tenemos G(x,y) = $E_x(N(y))$ =

$$\sum_{m=1}^{\infty} m P_x(N(y) = m) = \sum_{n=1}^{\infty} m\ \rho_{xy}\ \rho_{yy}^{m-1}\ (1-\rho_{yy})$$

tomando t=$\rho_{yy}$, la igualdad se transforma en

$$\sum_{n=1}^{\infty} m\ \rho_{xy}\ t^{m-1}(1-t) = \rho_{xy} * (1-t) * /(1-t)^2 = \frac{\rho_{xy}}{(1-t)} = \frac{\rho_{xy}}{(1-\rho_{yy})}$$

Sea y, recurrente. Si $\rho_{yy} = 1$, luego por (a2) se tiene

$$P_x(N(y) = \infty) = \lim_{m\to\infty} P_x(N(y) \geq= m) = \lim_{m\to\infty} \rho_{xy} = \rho_{xy}$$

En particular $P_y(N(y) = \infty) = 1$. Si una variable aleatoria no negativa tiene probabilidad positiva de ser infinita, su esperanza es infinita y por lo tanto

$$G(y,y) = E_y(N(y)) = \infty$$

Si $\rho_{xy}$=0, luego $P_x(T_y = m) = 0$ para todo entero m, lo que también es valido para $P_{xy}(n)=0$, y de esta forma $G(x,y)=0$. Si $P_{xy} > 0$ luego $P_x(N(y)=\infty)=\rho_{xy}>0$ y así,

$G(x,y)= E_x(N(y)) = \infty$, lo cual completa la prueba.

Si y es un estado transitorio $\sum_{n=1}^{\infty} P_{xy}^n = G(x,y) < \infty$ para todo x, es claro que $\lim_{n\to\infty} P_{xy}^n = 0$, para todo x.

Teorema 2. Sea x un estado recurrente y suponga que x alcanza y. Luego y es recurrente y $\rho_{xy} = \rho_{yx} = 1$.

Prueba.

Supongamos que $x \neq y$ y sea $n_0$ tal que

(a9) $n_0=\min(n\geq 1: P_x(T_y=n)>0)$

es claro que a partir de (a9)

(a10) $P_{xy}(m) = 0$, para $m<n_0$

Como $P_{xy}(n_0)> 0$, se pueden encontrar estados $y_1 \ldots y_{n_0-1}$ talque

$$P_x(X_1 = y_1, \ldots X_{n_0-1} = y_{n_0-1}, X_{n_0} = y_{n_0}) = P_{xy_1} \ldots P_{y_{n_0-1}y} > 0$$

Supongamos que $\rho_{yx} < 1$, entonces existe una probabilidad positiva 1-$\rho_{yx}$ de nunca alcanzar x, comenzando y. Esa probabilidad sería, comenzando en x:

$P_{xy_1} \ldots P_{y_{n_0-1}y}(1-\rho_{yx})$, pero esto implica que se visita primero los estados $y_1 \ldots y_{n_0-1}$ y que luego nunca se retorna al estado x, lo cual contradice el hecho de que x es un estado recurrente.

Como $\rho_{yx}$=1, existe n1 entero positivo, talque $P_{xy}(n1)>0$. Luego

$$P_{yy}(n_1+n+n_0) = P_y(X_{n_1+n+n_0} = y) \geq P_y(X_{n_1} = x, X_{n_1+n} = x, X_{n_1+n+n_0} = y) =$$
$$P_{yx}(n_1)P_{xx}(n)P_{xy}(n_0)$$

y esta forma

$$G(y,y) \geq \sum_{n=n_1+1+n_0} P_{yy}(n) = \sum_{n=1} P_{yy}(n_1+n+n_0) \geq P_{yx}(n_1)P_{xy}(n_0)\sum_{n=1}^{\infty} P_{xx}(n) =$$
$$P_{yx}(n_1)P_{xy}(n_0)G(x,x) = +\infty$$

lo cual implica que y también es un estado recurrente.

Corolario 1.

Sea C un conjunto cerrado irreducible de estados recurrentes. Luego $\rho_{xy}$ =1, $P_x(N(y)=\infty)$=1 y G(x,y)=$\infty$ para todas las escogencias de x,y en el conjunto C.

**Sección B**