UNDAV
Universidad Nacional de Avellaneda

Ingeniería Informática
Ciencias Básicas

Notas sobre Teoría de Colas

Primer Cuatrimestre de 2016

Gastón Andrés Freire

29 de marzo de 2016

Índice general

Capítulo 1

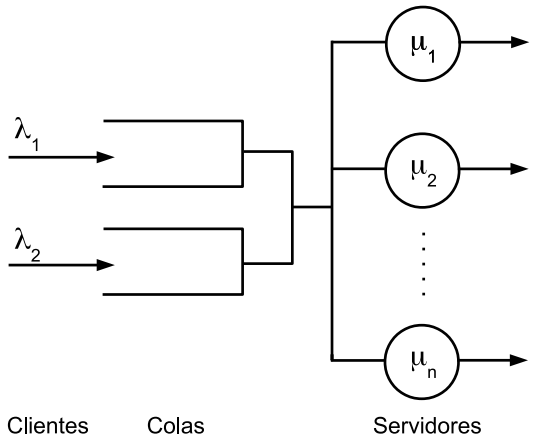
Introducción

Los primeros trabajos en Teoría de Colas se deben al célebre matemático e ingeniero danés A.K. Erlang, el cual publicó en 1909 "*La teoría de probabilidades y las conversaciones telefónicas*".

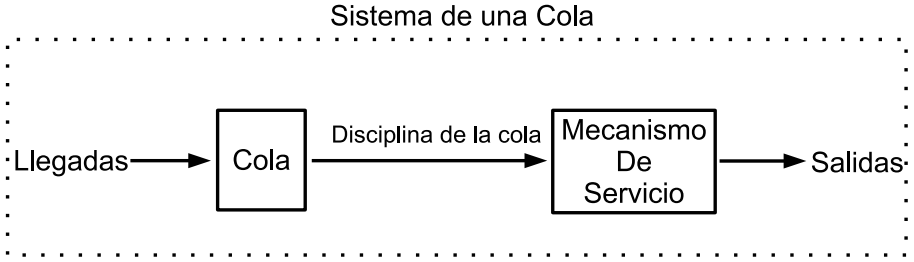
Continuando en esa línea de investigación en 1927 E.C. Molina publica "*Aplicación de la teoría de probabilidades a problemas de líneas telefónicas*" y a comienzos de 1930 F. Pollaczek publica sus trabajos sobre *llegadas poissonianas* y *servicios arbitrarios*.

Aunque hasta 1950 el desarrollo de esta teoría no tuvo mayores representantes que los mencionados anteriormente, con la explosión de las nuevas tecnologías la investigación en dicha área se vio disparada al punto de ser hoy una disciplina en sí misma, que permite abordar problemas de crucial importancia en las redes informáticas y muchos otros procesos donde hay servidores que atienden a un cierto número de clientes, cuya llegada se produce aleatoriamente siguiendo algún tipo de distribución, luego de lo cual deben distribuirse en colas que siguen una cierta disciplina, para ser atendidos mediante algún mecanismo de servicio cuya salida se produce también aleatoriamente siguiendo algún tipo de distribución.

Un Sistema de Colas puede ser representado de la siguiente forma:



Los elementos de una de dichas colas pueden ser representados de la siguiente forma:



De esta manera cuando un cliente entra en el sistema, debe posicionarse en la cola correspondiente mediante algún mecanismo que permita asignarlo a la misma. La asignación puede ser al azar o mediante alguna estrategia, por ejemplo elegir la cola menos sobrecargada. Una vez ubicado en la cola, la misma sigue una cierta disciplina, que puede ser:

FIFO: First in First Out.

LIFO: Last in First Out.

SIRO: Service in Random Order.

RR: Round Robin, que otorga un pequeño espacio de procesamiento equitativo a cada cliente a intervalos regulares, hasta que estos últimos completen su tarea.

Una vez determinada la disciplina de la cola, el cliente deberá esperar su turno según la misma, luego de lo cual será asignado a un servidor, el cual mediante un mecanismo de servicio atenderá al cliente, procesándolo y dejando al mismo listo para salir del sistema.

El objetivo de estas notas es presentar de forma ordenada y sistemática, los elementos más importantes en la Teoría de Colas, para que el estudiante pueda familiarizarse con la misma mediante un abordaje teórico completo, pero circunscripto a ciertos casos o ejemplos particulares, sin pretender resolver la generalidad de ellos.

Concretamente los casos que trataremos corresponden a los más fácilmente abordables teóricamente, suponiendo que la distribución del tiempo entre llegadas así como también la del tiempo de servicio es exponencial de parámetro λ. También supondremos que no hay limitación para la cantidad de clientes en una cola — *cosa que muchas veces puede hacerse* — y en este contexto analizaremos dos subcasos. El primero si contamos tan sólo con un servidor para atender a todos los clientes. El segundo si se dispone de s servidores.

Otro dato importante es que el análisis se concentrará en el estudio del sistema en *tiempo estacionario*, es decir cuando es posible suponer que el mismo se ha estabilizado y funciona con independencia del tiempo. En situaciones reales esto ocurre a partir de un cierto tiempo t_0 luego del cual el sistema comienza a comportarse de la misma forma con independencia del tiempo, cosa que no ocurre en momentos anteriores a t_0 pues es el lapso que le lleva al mismo estabilizarse.

Por ejemplo al momento de abrir un supermercado el mismo se encuentra vacío, las cajas están libres y a medida que los clientes van llegando el sistema comienza a evolucionar. Desde este momento hasta el instante t_0 podemos observar cambios marcados en la evolución del mismo, pero a partir de dicho instante comienza a funcionar más monótamente y se vuelve imposible percibir grandes cambios en el mismo. Para hacer alusión a esto es que suele hablarse de *tiempo estacionario*.

Aunque todo lo dicho anteriormente resulta *vago*, el objetivo de los capítulos que siguen es establecer las definiciones y teoremas necesarios para que toda esta incertidumbre se vaya discipando y comencemos a elaborar un lenguaje preciso que nos permita abordar estas cuestiones. A eso nos dedicaremos en las siguientes secciones.

Capítulo 2

Procesos Estocástico

Imaginemos un supermercado que funciona de lunes a viernes de 8 : 00 a 20 : 00 horas y posee una sola caja. Si la *v.a.* X cuenta el número de personas en dicha caja, en seguida nos damos cuenta que la distribución de la misma debe depender del tiempo t en la que observo la caja. Probablemente no dependa del día, si pensamos que el funcionamiento del supermercado no tiene por qué diferir de un día a otro, pero es indudable la dependencia del tiempo, por lo que deberíamos hablar de una *familia de variables aleatorias* del tipo:

$$(X_t)_{t\in T} \ / \ X_t : \Omega \longrightarrow \mathbb{R}$$

donde T es un cierto conjunto de índices, en este caso un momento del día, que podría ser un horario entre las 00 : 00 y 23 : 59. La dependencia del tiempo es obvia y basta con observar que aún cuando no conocemos nada sobre las distribuciones de X_t para t entre las 8 : 00 y las 20 : 00, es claro que si t no pertenece a dicho rango, entonces $X_t = 0$ sin lugar a dudas, ya que el supermercado se encuentra cerrado.

De esta manera definimos a un *proceso estocástico* como una familia de variables aleatorias:

$$\mathcal{F} = \left\{(X_t)_{t\in} \ / \ X_t : \Omega \longrightarrow \mathbb{R} \wedge R_{X_t} = E \, \forall \, t \in T\right\}$$

El espacio muestral Ω debe ser común a todas las variables, al conjunto de índices T se lo denomina *espacio de tiempos* y al rango E de estas variables se lo denomina *espacio de estados*.

Los procesos estocásticos suelen clasificarse según si los espacios de tiempo y estados son continuos o discretos. Por ejemplo si T es continuo pero E es discreto, el proceso estocástico se dirá *en tiempo continuo y con espacio de estados discreto*.

En el ejemplo el espacio de tiempos podría representarse mediante un intervalo $T = [0, 24)$, mientras que el espacio de estados podría ser $E = \mathbb{N}_0$.

Según las definiciones anteriores, fijar un tiempo t equivale a quedarse con una *v.a.* de la familia, a saber X_t. Pero otra cosa que puede hacerse es fijar un evento $A \subseteq \Omega$ y variar t para construir la función:

$$g_A(t) = X_t(A)$$

cosa que se denomina *trayectoria del evento A a lo largo del tiempo*, siendo común utilizar $A = \{\omega\}$ un suceso elemental, en cuyo caso simplemente hablamos de la *trayectoria de ω a lo largo del tiempo*.

Por ejemplo si el evento $A =$ "*Hay* 5 *personas en la caja*" entonces la función $g_A(t)$ es constante puesto que $X_t(A) = 5 \, \forall \, t$ para dicho evento, pudiendo interpretarse esto último como que el tiempo t sólo afecta a la probabilidad de que suceda el evento, no al resultado de la *v.a.* X_t cuando se la evalúa en dicho evento.

2.1. Características y propiedades básicas

Hay ciertas características y propiedades básicas que son de interés al estudiar *procesos estocásticos*, las cuales enumeraremos brevemente a continuación.

- La *función de medidas* asociada a un proceso estocástico es:

$$m(t) = E[X_t]$$

para todo $t \in T$ y para cada valor de t devuelve simplemente la esperanza de la *v.a.* asociada a dicho tiempo.

Por ejemplo en un determinado momento del día la media de clientes en la cola de un supermercado puede variar con respecto a otro momento, como suele ocurrir en las llamadas "*horas pico*".

- La *función de autocovarianzas* es una aplicación de dos variables:

$$c(s,t) = \text{Cov}(s,t) = E[X_s \cdot X_t] - E[X_s] \cdot E[X_t]$$

y sirve para evaluar el grado de dependencia temporal de la familia de variables. Es claro que en la situación ideal de independencia del tiempo la función de autocovarianzas $c(s,t) = 0$ para todos s, $t \in T$, ya que si X_s es independiente de X_t entonces $E[X_s \cdot X_t] = E[X_s \cdot X_t]$. Es inmediato que $c(s,t)$ es simétrica, es decir $c(s,t) = c(t,s)$. Por ejemplo si en nuestro supermercado la cantidad de clientes en la caja no dependiera del momento del día, no sólo sería constante la función de medias sino que también sería nula la función de autocovarianzas.

- Diremos que el proceso estocástico es *débilmente estacionario* si se cumplen las siguientes condiciones:

  1. Existe una constante $m \in \mathbb{R}$ tal que $m(t) = m$ para todo $t \in T$.
  2. Existe una función $\gamma(x)$ tal que $c(s,t) = \gamma(t-s)$ para todos s, $t \in T$. Esto significa que la función de autocovarianzas depende únicamente de la diferencia entre dos instantes de tiempo, y no de los instantes de tiempo propiamente dichos.

  En un proceso de este tipo no importa el tiempo t las *v.a.* X_t tienen la misma esperanza. Pero además tendrán que tener la misma varianza pues:

$$V[X_t] = c(t,t) = \gamma(t-t) = \gamma(0)$$

  Si además supiéramos por ejemplo que las X_t pertenecen a una misma familia de *v.a.* cuyas distribuciones quedan determinadas por sus respectivas esperanzas y varianzas, entonces la condición de ser débilmente estacionario automáticamente implicaría que todas la X_t tienen la misma distribución. Lo único que no queda garantizado es la independencia, por supuesto.

- Un proceso estocástico se dirá *fuertemente estacionario* si dados $t_1 < t_2 < \cdots < t_n \in T$ y $h > 0$:

$$(x_{t_1+h}, \cdots, x_{t_n+h}) \overset{\text{d}}{=} (x_{t_1}, \cdots, x_{t_n})$$

  Tomando $n = 1$ y $t = s + h$ con $h > 0$, entonces $X_s \overset{\text{d}}{=} X_t$ y por lo tanto $m(s) = m(t)$ para todo s, $t \in T$. Pero esto implica que $m(t) = m$ para todo $t \in T$.

  Por otro lado tomando $n = 2$ y dado $s < t$:

$$c(s,t) = \text{Cov}(X_s, X_t) = \text{Cov}\left(X_{0+s}, X_{(t-s)+s}\right) = \text{Cov}(X_0, X_{t-s}) = \gamma(t-s)$$

  De esta forma vemos que si un proceso estocástico es fuertemente estacionario entonces automáticamente debe ser débilmente estacionario.

- Finalmente un proceso estocástico se dirá *markoviano* si para $t_1 < t_2 < \cdots < t_{n-1} < t_n$ se tiene:

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

  Esto quiere decir que conociendo el estado del sistema en el presente, la distribución de probabilidad de posibles valores del sistema en el futuro sólo depende del estado del sistema en el presente, y no de los valores que haya tomado en el pasado.

2.2. Idea intuitiva de estacionariedad

En general cuando se habla de *estacionariedad* de un proceso estocástico se está pensando en un momento t_0 a partir del cual si $t \geq t_0$ se produce un equilibrio en las *v.a.* X_t de forma tal que la probabilidad de que X_t tome un cierto valor deja de depender del tiempo:

$$P\left(X_t = n\right) = P\left(X_{t_0} = n\right) \; \forall \, t \geq t_0$$

A continuación se ejemplificará una situación que permite comprender intuitivamente de qué se trata la idea de estacionariedad. Para ello consideremos un supermercado que cuenta con una línea de cajas extensa y personal necesario para ir habilitando nuevas cajas a medida que es necesario. Inicialmente hay una única caja abierta y cuando las habilitadas comienzan a tener todas colas mayores o iguales a 5 clientes se abre una nueva. Un día en particular al iniciar la jornada el flujo de clientes que ingresan a las cajas es constantemente de un cliente por minuto, y cada caja atiende en promedio un cliente cada cinco minutos. Se requiere modelar la situación, haciendo evolucionar el sistema, a los efectos de determinar el momento en el cual el sistema comienza a comportarse estacionariamente.

Para simplificar supondremos que el flujo de clientes es constante, llegando a la zona de cajas exactamente un cliente por minuto, el cual decidirá hacer fila en la primera caja que vea con menor cantidad de gente. También supondremos que el tiempo que lleva atender a una persona en cada caja es constantemente 5 minutos.

La aleatoriedad en t se obtendrá introduciendo incertidumbre respecto del momento en que entra el primer cliente. Supondremos que al abrir el supermercado, en tiempo $t = 0$, no sabemos en qué minuto entrará el primer cliente, pero una vez que esto ocurre el sistema evolucionará según lo descripto previamente.

Veamos cómo evolucione el sistema minuto a minuto en la siguiente tabla, donde por comodidad supondremos que el primer cliente entra en $t = 0$, aunque podría hacerlo en otro momento. Para cada tiempo y para cada caja se vuelca el número de clientes que se encuentran haciendo la fila.

En la tabla puede comprobarse que al sistema le lleva 40 minutos comenzar a entrar en tiempo estacionario. Al minuto 41 hay 5 cajas habilitadas con 4 personas cada una, y como llegan en promedio una persona cada minuto, eso le da el tiempo justo a las cajas para, en el lapso de 5 minutos, despachar 5 clientes, que es el número necesario para garantizar que no se saturen las cajas. Observamos además que la caja número 6 jamás necesitó ser habilitada, para un flujo de 5 clientes por minuto, lo cual es lógico si tenemos presente que cada caja atiende a razón de un cliente cada cinco minutos.

Un dato interesante es que a partir de $t \geq 40$ minutos, si X_t es la cantidad de clientes que hay esperando en las cajas, entonces por un lado $R_{X_t} = \{19, 20, 21\}$ y:

$$P\left(X_t = 19\right) = \frac{2}{5} \qquad P\left(X_t = 20\right) = \frac{2}{5} \qquad P\left(X_t = 21\right) = \frac{1}{5}$$

ya que el sistema entra en un estado cíclico de funcionamiento donde la cantidad de clientes sigue el patrón:

$$20,\, 21,\, 19,\, 20,\, 21,\, \mathbf{20},\, \mathbf{21},\, \mathbf{19},\, \mathbf{20},\, \mathbf{21},\, 20,\, 21,\, 19,\, 20,\, 21,\, \cdots$$

por lo que en un sentido práctico, la estacionariedad tiene que ver con este tipo de comportamiento regular del sistema, el cual encuentra un punto de equilibrio alrededor del cual se estabiliza.

Si el primer cliente hubiera ingresado en tiempo $t = t_0$ entonces el momento a partir del cual el sistema entraría en fase estacionaria sería $t \geq t_0 + 40$, y a partir de dicho momento la distribución de X_t sería la que hallamos previamente, con independencia de t.

Si el flujo de clientes fuera n por minuto y se tardara lo mismo en atenderlo, hubiera sido necesario habilitar n cajas hasta lograr que el sistema comience a entrar en tiempo estacionario, y la longitud de los bucles sería n en lugar de 5.

Al cerrar las puertas del supermercado ocurre el proceso inverso pues se van vaciando las cajas y la evolución del sistema se detiene.

A continuación se presenta la tabla que permite comprobar la validez de las afirmaciones anteriores:

Tiempo *(min)*	**N° Caja**						**Tiempo *(min)***	**N° Caja**					
	1	2	3	4	5	6		1	2	3	4	5	6
0	**1**						29	5	5	4	3		
1	2						30	**4**	**4**	5	4		
2	3						31	5	4	5	4		
3	4						32	5	5	5	4		
4	5	0					33	5	5	**4**	**4**		
5	**4**	**1**					34	5	5	5	4		
6	4	2					35	**4**	**4**	5	5		
7	4	3					36	5	4	5	5		
8	4	4					37	5	5	5	5	0	
9	4	5					38	5	5	**4**	**4**	**1**	
10	**4**	**4**					39	5	5	4	4	2	
11	5	4					40	**4**	**4**	4	4	3	
12	5	5	0				41	4	4	4	4	4	*
13	5	5	**1**				42	5	4	4	4	4	*
14	5	5	2				43	5	5	**3**	**3**	**3**	*
15	**4**	**4**	3				44	5	5	4	3	3	*
16	4	4	4				45	**4**	**4**	4	4	3	*
17	5	4	4				46	4	4	4	4	4	
18	5	5	**3**				47	5	4	4	4	4	
19	5	5	4				48	5	5	**3**	**3**	**3**	
20	**4**	**4**	5				49	5	5	4	3	3	
21	5	4	5				50	**4**	**4**	4	4	3	
22	5	5	5	0			51	4	4	4	4	4	*
23	5	5	**4**	**1**			52	5	4	4	4	4	*
24	5	5	4	2			53	5	5	**3**	**3**	**3**	*
25	**4**	**4**	4	3			54	5	5	4	3	3	*
26	4	4	4	4			55	**4**	**4**	4	4	3	*
27	5	4	4	4			⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	5	5	**3**	**3**			**Tiempo estacionario**						

Vemos entonces que hasta alcanzar el tiempo estacionario el sistema dinámicamente se va ajustando al flujo de clientes y ritmo de atención, por lo que debemos tener presentes estas cuestiones al momento de elaborar un modelo matemático para describirlo, cosa que se hará en la próxima sección.

Otra cosa a tener presente es que en la realidad los eventos no suelen suceder de forma determinística y aunque se conozca el promedio de clientes que llegan por minuto y el tiempo promedio de atención, esto no quiere decir que a cada minuto llegue exactamente ese número de clientes ni que exactamente se despachen cada cinco minutos, tiempos que en realidad constituyen fenómenos aleatorios.

2.3. Procesos de nacimiento y muerte

En lo que sigue definiremos como $o(h)$ a cualquier función de una variable verificando:

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

siendo inmediato que combinaciones lineales y productos de funciones $o(h)$ dan como resultado una función $o(h)$.

Consideremos un proceso estocástico $(N_t)_{t\geq 0}$ en tiempo continuo y con espacio de estados discreto $E = \mathbb{N}_0$. Diremos que el sistema se encuentra en estado E_n en el instante t cuando $N_t = n$ y denotemos por $P_n(t) = P(N_t = n)$. El proceso se dirá *de nacimiento y muerte* si y sólo si existen dos sucesiones de números reales no negativos $(\lambda_n)_{n\in\mathbb{N}}$ y $(\mu_n)_{n\in\mathbb{N}}$ tales que se cumplen las siguientes condiciones:

1. Los cambios de estado permitidos son desde E_0 hasta E_1 o desde E_n hasta E_{n-1} o E_{n+1} para todo $n \in \mathbb{N}$.

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

3. La probabilidad de que el sistema tenga más de un cambio de estado entre los instantes t y $t+h$ es $o(h)$.

2.3.1. Distribución de probabilidades de N_t

Para poder determinar la distribución de probabilidades de cada N_t, es decir $P(N_t = n) = P_n(t)$, no podemos hacerlo directamente debido a que se trata de un proceso diferencial que evoluciona a través del tiempo. Lo que sí podemos hacer es encontrar una relación entre lo que ocurre en el instante $t+h$ y lo que ocurre en el instante t, una suerte de cociente incremental que permita luego estudiar la variación de la probabilidad en un instante t.

Una vez conocida la variación entre los instantes t y $t+h$ es posible obtener el cociente incremental para $P_n(t)$, cosa que permite establecer una relación entre $P_n'(t)$, $P_n(t)$, $P_{n-1}(t)$ y $P_{n+1}(t)$.

Para empezar si utilizamos el teorema de la probabilidad total obtenemos:

$$\begin{aligned} P_n(t+h) &= P(N_{t+h} = n) \\ &= P(N_{t+h} = n \,|_{N_t=n-1}) \cdot P(N_t = n-1) + P(N_{t+h} = n \,|_{N_t=n}) \cdot P(N_t = n) \\ &+ P(N_{t+h} = n \,|_{N_t=n+1}) \cdot P(N_t = n+1) + \sum_{m \in \mathbb{N}_0 \setminus \{n-1,n,n+1\}} P(N_{t+h} = n \,|_{N_t=m}) \cdot P(N_t = m) \end{aligned}$$

El primer término contempla la probabilidad de que haya habido un nacimiento entre el instante t y $t+h$. El segundo de que el sistema no hubiera cambiado, el tercero contempla la posibilidad de una muerte, y por último, la sumatoria contempla que hubiera habido más de un cambio.

Si tenemos presente que:

$$\begin{aligned} P(N_{t+h} = n \,|_{N_t=n}) \cdot P(N_t = n) &= (1 - (\lambda_n h + o(h)) - (\mu_n h + o(h)) - o(h)) \cdot P(N_t = n) \\ &= P(N_t = n) - \left(\lambda_n + \mu_n + \frac{o(h)}{h}\right) h \cdot P(N_t = n) \end{aligned}$$

entonces:

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

Reescribiendo la igualdad anterior obtenemos el cociente incremental:

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

Tomando $\lim_{h \to 0}$ en la igualdad, del lado izquierdo surge $P_n'(t)$ y del lado derecho todos los términos $\frac{o(h)}{h}$ tienden a cero, obteniendo:

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

En el caso de que $n = 0$ el término λ_{n-1} no está presente y μ_0 tampoco porque no puede haber una muerte cuando aún no hay nacimientos:

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

Por lo tanto, el sistema de ecuaciones diferenciales que gobierna las distribuciones de N_t es:

$$P_n'(t) = \begin{cases} \lambda_{n-1} P_{n-1}(t) - (\lambda_n + \mu_n) P_n(t) + \mu_{n+1} P_{n+1}(t) & , \text{ si } n \geq 1 \\ -\lambda_0 P_0(t) + \mu_1 P_1(t) & , \text{ si } n = 0 \end{cases}$$

En general, a tiempo inicial $N_0 = 0$, por lo que las condiciones iniciales para este sistema de ecuaciones diferenciales es:

$$P_0(0) = 1 \qquad\qquad P_n(0) = 0 \ (\forall n \geq 1)$$

2.3.2. Un caso particular interesante

Supongamos que no hubiera muertes, es decir $\mu_n = 0$ para todo $n \geq 1$ y que la tasa de nacimientos fuera independiente de la población, es decir $\lambda_n = \lambda$ para todo $n \in \mathbb{N}_0$. En ese caso el sistema se reduce a:

$$P_n'(t) = \begin{cases} \lambda P_{n-1}(t) - \lambda P_n(t) & , \text{ si } n \geq 1 \\ -\lambda P_0(t) & , \text{ si } n = 0 \end{cases}$$

Si $n = 0$ entonces:

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

La solución general de esta ecuación diferencial se obtiene muy fácilmente y es:

$$P_0(t) = c \cdot e^{-\lambda t}$$

Además como $P_0(0) = 1$ entonces debe ser $c = 1$ por lo que:

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

Por inducción en n puede probarse que la solución para n en general es:

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

por lo que fijado un instante t, la *v.a.* N_t sigue una distribución $\mathcal{P}(\lambda t)$.

Esto quiere decir que los Procesos de Poisson pueden pensarse como procesos de nacimiento y muerte cuando las tasas de natalidad son independientes de la población y cuando las tasas de mortalidad son todas nulas.

2.3.3. Análisis del proceso estacionario

Ya se mencionó en la sección anterior que la fase más importante de un proceso de nacimiento y muerte suele ser la estacionaria, que es cuando el sistema se encuentra adaptado a las tasas de natalidad y mortalidad y su evolución se independiza del tiempo t.

Si asumimos que en el intervalo de tiempo $[a, b]$ $P_n(t_1) = P_n(t_2)$ para todo par $t_1, t_2 \in [a, b]$, entonces $\forall\, t \in [a, b]$ $P_n(t) = p_n$ y además $P_n'(t) = 0$ para todo $t \in (a, b)$. Las ecuaciones se reducen a:

$$\begin{aligned} (\lambda_n + \mu_n) \cdot p_n &= \lambda_{n-1} \cdot p_{n-1} + \mu_{n+1} \cdot p_{n+1} \\ p_1 &= \frac{\lambda_0}{\mu_1} \cdot p_0 = \end{aligned}$$

Para el caso $n = 1$ vemos que utilizando la ecuación para p_1 en la de más arriba:

$$\begin{aligned} (\lambda_1 + \mu_1) \cdot p_1 &= \lambda_0 \cdot p_0 + \mu_2 \cdot p_2 \\ \Leftrightarrow (\lambda_1 + \mu_1) \cdot \frac{\lambda_0}{\mu_1} p_0 &= \lambda_0 \cdot p_0 + \mu_2 \cdot p_2 \\ \Leftrightarrow \frac{\lambda_1 \lambda_0}{\mu_1} p_0 + \cancel{\lambda_0 p_0} &= \cancel{\lambda_0 p_0} + \mu_2 \cdot p_2 \end{aligned}$$

de donde podemos despejar p_2 como:

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

Vemos que salta a la vista la forma general de p_n y conjeturamos que para todo $n \geq 1$:

$$p_n = \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1} \cdot p_0$$

cosa que demostraremos por inducción en n.

El caso $n = 1$ ya está establecido. Supongamos ahora que la propiedad es válida para $k \leq n$ y demostremos que entonces vale para $n + 1$:

$$\begin{aligned}
(\lambda_n + \mu_n) \cdot p_n &= \lambda_{n-1} \cdot p_{n-1} + \mu_{n+1} \cdot p_{n+1} \\
\Rightarrow (\lambda_n + \mu_n) \cdot \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1} \cdot p_0 &= \lambda_{n-1} \cdot \frac{\lambda_{n-2} \cdots \lambda_0}{\mu_{n-1} \cdots \mu_1} \cdot p_0 + \mu_{n+1} \cdot p_{n+1} \\
\Rightarrow \frac{\lambda_n \cdots \lambda_0}{\mu_n \cdots \mu_1} \cdot p_0 + \cancel{\frac{\lambda_{n-1} \cdots \lambda_0}{\mu_{n-1} \cdots \mu_1} \cdot p_0} &= \cancel{\frac{\lambda_{n-1} \cdots \lambda_0}{\mu_{n-1} \cdots \mu_1} \cdot p_0} + \mu_{n+1} \cdot p_{n+1} \\
\Rightarrow \mu_{n+1} \cdot p_{n+1} &= \frac{\lambda_n \cdots \lambda_0}{\mu_n \cdots \mu_1} \cdot p_0 \\
\Rightarrow p_{n+1} &= \frac{\lambda_n \cdots \lambda_0}{\mu_{n+1} \cdots \mu_1} \cdot p_0
\end{aligned}$$

Vemos entonces que la propiedad vale también para $n + 1$ y eso demuestra que es válida para todo $n \in \mathbb{N}$.

Si para cada $n \in \mathbb{N}$ definimos las constantes:

$$c_n = \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1} \qquad \Rightarrow \qquad p_n = c_n \cdot p_0$$

Además como:

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

ya que $p_n = P(N = n)$ y la suma de todas ellas debe ser igual a 1, entonces:

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

Vemos entonces que:

$$p_0 = \frac{1}{1 + \sum\limits_{n=1}^{+\infty} c_n}$$

2.3.4. Condición necesaria y suficiente para la estacionariedad

En la sección anterior vimos que si un proceso de nacimiento y muerte con tasas de natalidad $(\lambda_n)_{n \in \mathbb{N}_0}$ y $(\mu_n)_{n \in \mathbb{N}}$ es estacionario y para cada $n \in \mathbb{N}$ definimos las constantes:

$$c_n = \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1} \qquad \Rightarrow \qquad p_n = c_n \cdot p_0$$

Además demostramos que:

$$p_0 = \frac{1}{1 + \sum\limits_{n=1}^{+\infty} c_n}$$

o equivalentemente:

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

Esto significa que si un proceso de nacimiento y muerte es estacionario entonces la serie:

$$\sum_{n=1}^{+\infty} \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1}$$

es convergente. Pero lo interesante es que también vale la reciproca: si la serie converge entonces el proceso de nacimiento y muerte es estacionario.

Por lo tanto la condición necesaria y suficiente para la estacionariedad de un proceso de nacimiento y muerte es la convergencia de dicha serie.

Capítulo 3

Nociones básicas sobre teoría de colas

Las *colas* o *líneas de espera* se producen cuando la demanda de un servicio supera la capacidad para atender dicho servicio. Algunos ejemplos de colas son los siguientes:

- Las colas que se producen en las cajas de cualquier establecimiento comercial.
- Las líneas de espera que se producen en los peajes cuando la capacidad para cobrar se ve superada por la cantidad de autos que llegan.
- Las colas de electrodomésticos que esperan ser reparados en un servicio técnico.
- Las colas de procesos que esperan ser atendidos por un servidor.
- La información que solicitamos a internet puede recibirse con demora cuando hay congestionamiento en la red, pues el servidor divide su tiempo para enviar equitativamente paquetes de datos a todos los clientes, y en caso de haber más de los que puede atender se satura la red.
- Las colas de autos que se producen en los cruces de calles cuando hay demasiada circulación de vehículos.

Y la lista podría continuar indefinidamente ya que uno de los fenómenos más comunes cuando se ofrece un servicio es la aparición de colas o líneas de espera como resultado de una demanda mayor a la esperada. Contar con un modelo matemático de las mismas ayuda a poder simular su comportamiento, generalmente con el objeto de encontrar la manera de optimizar los tiempos de atención de manera tal que a un costo mínimo se pueda atender toda la demanda de clientes.

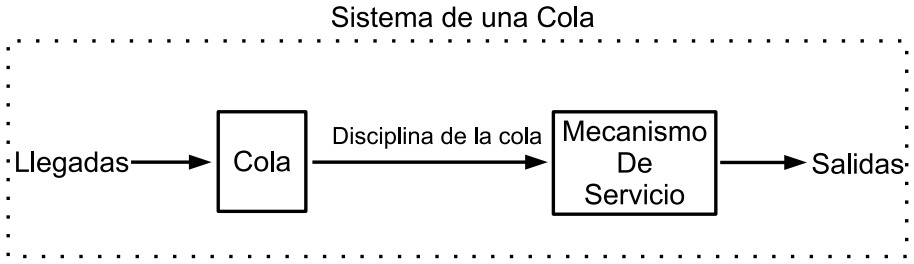
Interesa poder responder preguntas como:

- ¿Cuántas cajas debo habilitar para garantizar colas menores a tantos clientes?
- ¿Cuál es la demanda de clientes que satura mi sistema?
- ¿Cuánto se descongestiona mi sistema si habilito tres servidores más?
- ¿Cuántos servidores debo habilitar para reducir a la mitad el tamaño de las colas en horas pico?

Es evidente que antes de poder hablar de *sistemas de colas* tendríamos que definir precisamente qué es una cola. Eso haremos en la siguiente sección.

3.1. Sistema de una cola

El diagrama básico que representa un sistema de una cola es el siguiente:



Los elementos que lo componen son los que describiremos a continuación.

3.1.1. Fuente de entrada o población potencial

Es el espacio de todos los posibles clientes, que puede ser finito o infinito, constituido por toda la población de individuos que potencialmente pueden solicitar el servicio.

3.1.2. Clientes

Son los elementos del conjunto formado por la fuente de entrada o población potencial, es decir cada individuo que puede solicitar el servicio es un cliente. Lo que interesa en realidad es conocer el mecanismo mediante el cual los clientes van llegando. Si $t_1 < t_2 < t_3 < \cdots < t_n < \cdots$ son los tiempos en los que llegan los diferentes clientes, entonces las diferencias entre estos tiempos $t_2 - t_1$, $t_3 - t_2$, $\cdots$, $t_n - t_{n-1}$, $\cdots$ denotadas por $\tau_n = t_n - t_{n-1}$ con $n \in \mathbb{N}$ nos permitirán aproximar las tasas de llegada de los mismos por unidad de tiempo, que es lo que interesa.

Por convención cuando se trata de una población infinita suele suponerse que τ_n no depende del número de clientes que se estén procesando, mientras que cuando la población es finita sí depende de dicho número.

3.1.3. Capacidad de la cola

Es el número máximo de clientes que pueden estar haciendo cola simultáneamente. Para simplificar el análisis es común suponer que la capacidad de la cola es infinita, en los casos en que la misma sea suficientemente grande como para que sea plausible hacerlo.

3.1.4. Disciplina de la cola

La disciplina de una cola es de gran importancia en la teoría, pues determina la manera en la que se eligen los clientes que van llegando para ser procesados. Hay numerosos mecanismos de selección, siendo los más comunes los siguientes:

FIFO: Abreviación de *First In First Out*, en este disciplinamiento el primer cliente en llegar es el primero en salir. Algo similar ocurre en las cajas de los supermercados, en las colas de los bancos, etc...

LIFO: Abreviación de *Last In First Out*, significa que el último en llegar es el primero en salir. Por ejemplo en una pila, el último dato que se empuja en ella es el primero en salir, por lo que se trata de una estructura *LIFO*.

RSS o SIRO: Abreviación de *Random Selection of Service* o *Service In Random Order*, aquí los clientes se eligen al azar.

RR: Abreviación de *Round Robin*, según la cual se asignará un pequeño tiempo de procesamiento equitativamente a cada proceso de la cola, para asegurar que puedan ser atendidos simultáneamente. Algo similar ocurre en los procesos que corren en multitarea, pues el procesador asigna un breve lapso de tiempo a cada proceso para simular que todos están corriendo al mismo tiempo.

3.1.5. Mecanismo de servicio

Es el procedimiento mediante el cual se atiende a los clientes que solicitan servicio. En general lo que interesa es conocer para cada servidor la distribución de probabilidades que nos permita conocer el tiempo que le llevará atender o procesar un cliente, es decir la tasa de atención. No necesariamente todos los servidores deben tener la misma tasa de atención.

Por ejemplo en un supermercado con 5 cajas, todos los cajeros no tendrán la misma destreza para atender a los clientes, por lo que cada uno tendrá su propia tasa de atención, y podemos resumir entonces el mecanismo de servicio de cada uno indicando dichas tasas.

3.1.6. La cola

Es el conjunto de clientes o procesos que esperan por ser atendidos, entendiendo que estos últimos ya han solicitado el servicio pero que aún no pasaron por el mecanismo de servicio.

3.1.7. El sistema de la cola

Es el conjunto formado por la cola, la disciplina de la cola y el mecanismo de servicio, que nos permite comprender el funcionamiento de todo el sistema en conjunto.

3.1.8. Formalización matemática

Para representar las llegadas y salidas de los clientes utilizaremos procesos de nacimiento y muerte, por lo que es necesario acordar la notación a utilizar:

$N(t)$**:** Es el número de clientes presentes en el sistema en el instante t, con $t \geq 0$.

$N_q(t)$**:** Es el número de clientes presentes en la cola en el instante t.

$P_n(t)$**:** Es la probabilidad de que en el instante t haya n clientes en el sistema, es decir $P(N(t) = n)$.

λ_n**:** Es la tasa de llegada de clientes por unidad de tiempo cuando en el mismo hay n clientes.

μ_n**:** Es la tasa de salida del sistema por unidad de tiempo cuando en el mismo hay n clientes.

ρ**:** Vamos a definir la *constante de utilización del sistema* o *intensidad de tráfico* cuando hay n clientes en el mismo como:

$$\rho_n = \frac{\lambda_n}{\mu_n}$$

lo que nos da una idea sobre si en ese momento la cola tiende a crecer o disminuir. Si $\rho_n < 1$ se despachan más clientes de los que entran, por lo que la cola tenderá a bajar. Si $\rho_n = 1$ entonces la cola tiende a permanecer sin cambios. Y si $\rho_n > 1$ la misma tiende a crecer.

Es muy importante mantener a raya el valor de ρ_n porque no es para nada deseable que sea $\rho_n > 1$. Si esto ocurre el sistema podría colapsar debido a la acumulación de los procesos que no pueden ser atendidos por ser la demanda mayor que la capacidad de procesamiento.

Por ley general se debe mantener $\rho_n < 1$ y cuánto más cerca a 1 se encuentre dicho valor, menos tiempo libre habrá para los servidores, los cuales dejarán de tener tiempo libre en el caso de ser $\rho_n = 1$.

En las definiciones anteriores las únicas variables que dependen del tiempo t son $N(t)$, $N_q(t)$ y $P_n(t)$. Si nos interesaran procesos estacionarios entonces la dependencia del tiempo en dichas variables desaparecería y tendría sentido hacer las siguientes definiciones:

N**:** Número de clientes en el sistema, ahora independiente de t.

N_q**:** Número de clientes en la cola.

p_n**:** Probabilidad de que haya n clientes en el sistema.

L**:** Cantidad de clientes que se espera haya en el sistema, es decir $E[N]$.

L_q**:** Cantidad de clientes que se espera haya en la cola, o sea $E[N_q]$.

$\mathcal{W}$**:** Es la *v.a.* que mide la cantidad de tiempo que un cliente elegido al azar permanece en el sistema.

$\mathcal{W}_q$**:** Es la *v.a.* que mide la cantidad de tiempo que un cliente elegido al azar permanece en la cola.

W**:** Es el tiempo medio que un cliente permanece en el sistema, es decir $E[\mathcal{W}]$.

W_q**:** Es el tiempo medio que un cliente permanece en la cola, es decir $E[\mathcal{W}_q]$.

W_s**:** Es el tiempo medio de servicio, es decir el tiempo medio que tarda el mecanismo de servicios en procesar a un cliente.

3.1.9. Notación de Kendall

En 1953 Kendall introdujo su notación abreviada para designar a un sistema de una cola, el cual posteriormente evolucionó hasta llegar a la tradicional forma:

$$A/B/s/K/H/Z$$

donde cada una de estas letras describe un aspecto del sistema de una cola.

A**:** Es la distribución del tiempo entre llegadas de clientes al sistema, la cual puede ser:

- M**:** Exponencial de parámetro λ_n.
- D**:** Determinística.
- E_k**:** Erlang con segundo parámetro k, es decir que los λ_n se interpretan como si fueran los k_n.
- U**:** Uniforme.
- Γ**:** Gamma.
- G**:** Genérica. En este caso los parámetros se interpretan como el inverso del tiempo medio del servidor:

$$\lambda_n = \frac{1}{E[G_n]}$$

B**:** Es la distribución del tiempo de servicio, la cual se describe análogamente a A.

s**:** Cantidad de servidores, pudiendo ser ∞.

K**:** Capacidad de la cola, cuyo valor por defecto es ∞.

H**:** Tamaño de la población, cuyo valor por defecto es también ∞.

Z**:** Disciplina de la cola, cuyo valor por defecto es *FIFO*.

La nomenclatura se puede utilizar únicamente de las siguientes formas:

- $A/B/s$: Se asume que los tres parámetros restantes toman los valores por defecto.
- $A/B/s/K$: Se asume que los dos últimos parámetros toman valores por defecto.
- $A/B/s/K/H$: El último parámetro es por defecto.
- $A/B/s/K/H/Z$: Se especifican todos los parámetros.

Por ejemplo $M/M/3$ significa que la cola tiene entrada exponencial, salida exponencial y 3 servidores, con capacidad infinita de cola, tamaño de población infinito y disciplina de cola *FIFO*.

3.1.10. Fórmulas de Little

Si las tasas de llegada fueran constantes $\lambda_n = \lambda$ para todo $n \in \mathbb{N}$ entonces sería lógico escribir:

$$L = \lambda \cdot W$$

pues si cada cliente permanece en el sistema un tiempo medio dado por W e ingresan al mismo un promedio de λ clientes por unidad de tiempo, entonces la igualdad anterior es razonable. Por ejemplo si cada cliente permanece en el sistema un promedio de 2 minutos e ingresan al sistema 5 clientes por minuto, entonces en promedio habrá 10 clientes en el sistema:

$$\begin{array}{rccccccccc} & & +5 & +5 & +5 & +5 & +5 & +5 & \\ t: & 0 & 1 & 2 & 3 & 4 & 5 & 6 & \cdots \\ n: & 0 & 5 & 10 & 10 & 10 & 10 & 10 & \cdots \\ & & & & -5 & -5 & -5 & -5 & \end{array}$$

Análogamente es válido escribir:

$$L_q = \lambda \cdot W_q$$

En el caso de que λ no sea constante, la misma será una variable aleatoria con $R_\lambda = \{\lambda_n : n \in \mathbb{N}\}$. Es más como $P(\lambda = \lambda_n) = P(N = n) = p_n$ entonces:

$$E[\lambda] = \sum_{n=0}^{+\infty} \lambda_n \cdot P(\lambda = \lambda_n) = \sum_{n=0}^{+\infty} \lambda_n \cdot p_n = \overline{\lambda}$$

lo que quiere decir que el valor esperado de λ es $\overline{\lambda}$, que es la tasa media de llegada de clientes al sistema.

Si multiplicamos esta tasa media de llegada de clientes al sistema por el tiempo medio que permanece un cliente en el sistema, obtenemos la Primera Fórmula de Little:

$$L = E[N] = \overline{\lambda} \cdot W$$

que establece que la cantidad media de clientes en el sistema es la tasa media de llegada por el tiempo medio de permanencia de los mismos.

La segunda es similar a la primera y relaciona la cantidad media de clientes en la cola con la tasa media de llegada de clientes al sistema y el tiempo medio de permanencia en la cola de un cliente al azar:

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

Si ahora μ_n es la tasa media de clientes que puede atender un servidor cuando en el mismo hay n clientes entonces para calcular $\overline{\mu}$ hay que utilizar probabilidad condicional. En la sumatoria para calcular la esperanza, a cada valor μ_n habrá que multiplicarlo por la probabilidad de que $N = n$, pero sabiendo de antemano que el servidor está ocupado, cosa que equivale a condicionar con $n \geq 1$. Esto es muy importante porque caso contrario estaría mal calculada $\overline{\mu}$.

Procediendo según las indicaciones del párrafo anterior:

$$\begin{aligned} \overline{\mu} &= \sum_{n=1}^{+\infty} \mu_n \cdot P(N = n\,|_{n\geq 1}) \\ &= \sum_{n=1}^{+\infty} \mu_n \cdot \frac{p_n}{1 - p_0} \\ &= \frac{1}{1 - p_0} \cdot \sum_{n=1}^{+\infty} \mu_n \cdot p_n \end{aligned}$$

es la cantidad media de clientes que puede atender un servidor por unidad de tiempo, por lo que:

$$\frac{1}{\overline{\mu}}$$

es el tiempo medio que tarda el servidor en atender un cliente elegido al azar.

Esto permite obtener la tercera fórmula de Little afirmando que el tiempo medio de permanencia en el sistema de un cliente es el tiempo medio que permanece en la cola más el tiempo medio que tarda en ser atendido por el servidor, esto es:

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

Por lo tanto las Fórmulas de Little afirman que:

$$L = \overline{\lambda} \cdot W$$
$$L_q = \overline{\lambda} \cdot W_q$$
$$W = W_q + \frac{1}{\overline{\mu}}$$

3.2. Distribuciones Exponencial - Gamma - Poisson

3.2.1. Distribución Gamma

Diremos que $X \sim \Gamma(\alpha, \lambda)$ — *se lee* X *es* Gamma *de parámetros* α *y* λ — si y sólo si su función de densidad es:

$$f(x) = \frac{e^{-\lambda x} x^{\alpha-1} \lambda^{\alpha}}{\Gamma(\alpha)} \qquad \Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x}\, dx$$

Si $X \sim \Gamma(\alpha, \lambda)$ entonces:

$$E[X] = \frac{\alpha}{\lambda} \qquad V(X) = \frac{\alpha}{\lambda^2}$$

Una de las propiedades importantes de esta distribución es que si $X \sim \Gamma(\alpha, \lambda)$ entonces $a \cdot X \sim \Gamma\left(\alpha, \frac{\lambda}{a}\right)$. En el caso $\lambda = 1$ la distribución se llama Gamma Standard, y utilizando la propiedad anterior es fácil comprobar que $\lambda \cdot X \sim \Gamma(\alpha, 1)$ por lo que esto permite calcular probabilidades de una *v.a.* Gamma a partir de la estandarizada:

$$P(X \leq a) = P(\lambda X \leq \lambda a) = P(Y \leq \lambda a)$$

donde $Y \sim \Gamma(\alpha, 1)$.

Otra propiedad importante de la distribución gamma es que es *reproductiva con respecto a su parámetro* α. Esto quiere decir que si $X_i \sim \Gamma(\alpha_i, \lambda)$ para $1 \leq i \leq n$ entonces:

$$X = \sum_{i=1}^{n} X_i \sim \Gamma(\alpha, \lambda) \qquad \alpha = \sum_{i=1}^{n} \alpha_i$$

3.2.2. Distribución Exponencial

La distribución exponencial es un caso particular de la Gamma, pues $X \sim \varepsilon(\lambda)$ si y sólo si $X \sim \Gamma(1, \lambda)$, por lo que su función de densidad es:

$$f(x) = \lambda e^{-\lambda x} \qquad F(x) = \left[1 - e^{-\lambda x}\right] \cdot I_{(0,+\infty)}(x)$$
$$E[X] = \frac{1}{\lambda} \qquad V[X] = \frac{1}{\lambda^2}$$

La distribución exponencial tiene varias propiedades que son muy importantes. La primera es la de monotonía, pues la probabilidad de que X tome valores en un intervalo de longitud fija Δt decrece a medida que el extremo izquierdo del intervalo aumenta.

La segunda es la falta de memoria, pues la probabilidad de que $X > t + \Delta t$ sabiendo que $X > t$ equivale a la probabilidad de que $X > \Delta t$.

La tercera tiene que ver con la distribución del mínimo de n *v.a.* exponenciales, que resulta ser una nueva exponencial pero de parámetro el promedio de todos ellos.

Y por último la cuarta propiedad tiene que ver con la suma de n exponenciales independientes de un mismo parámetro λ, que resulta ser una $\Gamma(n, \lambda)$.

A continuación se analizarán más en detalle cada una de estas propiedades.

3.2.2.1. Propiedad de monotonía

Debemos demostrar que $P(X \in [a,b]) > P(X \in [a+\Delta t, b+\Delta t])$:

$$\begin{aligned} P(X \in [a,b]) &= P(a \leq X \leq b) = F(b) - F(a) \\ &= \left(1 - e^{-\lambda b}\right) - \left(1 - e^{-\lambda a}\right) \\ &= e^{-\lambda a} - e^{-\lambda b} \\ &> e^{-\lambda \Delta t} \cdot \left(e^{-\lambda a} - e^{-\lambda b}\right) \\ &= e^{-\lambda(a+\Delta t)} - e^{-\lambda(b+\Delta t)} \\ &= P(a + \Delta t < X < b + \Delta t) \\ &= P(X \in [a+\Delta t, b+\Delta t]) \end{aligned}$$

3.2.2.2. Propiedad de falta de memoria

Como anticipamos antes, esta propiedad afirma que:

$$P(X > t + \Delta t \,|_{X>t}) = P(X > \Delta t)$$

Para demostrarlo basta utilizar la definición de probabilidad condicional y la expresión de la acumulada $F(x)$:

$$\begin{aligned} P(X > t + \Delta t \,|_{X>t}) &= \frac{P(X > t + \Delta t \wedge X > t)}{P(X > t)} \\ &= \frac{P(X > t + \Delta t)}{P(X > t)} = \frac{e^{-\lambda(t+\Delta t)}}{e^{-\lambda t}} \\ &= e^{-\lambda \Delta t} = P(X > \Delta t) \end{aligned}$$

Si por ejemplo el tiempo que falta para completar una tarea es $T \sim \varepsilon(\lambda)$ y ya ha pasado un tiempo t_0 sin que hayamos completado la misma, la falta de memoria indica que $P(T > t_0 + a \,|_{T>t_0}) = P(T > a)$. Esto quiere decir que la probabilidad de que tengamos que esperar un tiempo adicional mayor que a desde t_0 es la misma que la de tener que esperar un tiempo mayor que a desde $t = 0$.

3.2.2.3. Propiedad del mínimo

Si $X_1, \cdots, X_n$ son *v.a.* independientes con distribución $X_i \sim \varepsilon(\lambda_i)$ para $1 \leq i \leq n$ y consideramos la *v.a.*:

$$X = \min_{1 \leq i \leq n} X_i$$

entonces $X \sim \varepsilon(\lambda)$ con:

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

Consideremos X como en el enunciado y sea $Y \sim \varepsilon(\lambda)$ con λ el definido previamente. Entonces:

$$\begin{aligned} P(X \leq a) &= 1 - P(X > a) \\ &= 1 - P(X_1 > a \wedge \cdots \wedge X_n > a) \\ &= 1 - \prod_{i=1}^{n} P(X_i > a) \\ &= 1 - e^{-\lambda_1 \cdot a} \cdot e^{-\lambda_2 \cdot a} \cdots e^{-\lambda_n \cdot a} \\ &= 1 - e^{-(\lambda_1 + \cdots + \lambda_n) \cdot a} \\ &= 1 - e^{-\lambda \cdot a} \\ &= 1 - P(Y > a) \\ &= P(Y \leq a) \end{aligned}$$

Pero entonces X tiene la misma distribución que Y, por lo que:

$$X \sim \varepsilon(\lambda)$$

tal como queríamos demostrar.

3.2.2.4. Suma de exponenciales independientes de un mismo parámetro

Consideremos $X_1, \cdots, X_n$ *v.a.i.* todas exponenciales de parámetro λ y hagamos $X = X_1 + \cdots + X_n$. Entonces:

$$X \sim \Gamma(n, \lambda)$$

La demostración es inmediata utilizando que Γ es reproductiva con respecto al parámetro α.

3.2.3. Distribución de Poisson

Diremos que X es una *v.a.* Poisson de parámetro λ si y sólo si su función de densidad puntual es:

$$p_X(k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \qquad E[X] = \lambda \qquad V[X] = \lambda$$

El parámetro λ se interpreta como el número de ocurrencias promedio de un determinado fenómeno en un lapso determinado de tiempo. La propiedad más importante de las *v.a.* de Poisson es su aditividad en el parámetro λ. Si $X_i \sim \mathcal{P}(\lambda_i)$ para $1 \leq i \leq n$ son independientes y se define $X = X_1 + \cdots + X_n$, entonces $X \sim \mathcal{P}(\lambda)$ con $\lambda = \lambda_1 + \cdots + \lambda_n$.

3.3. Aplicaciones a la teoría de colas

Las propiedades de las distribuciones estudiadas en la sección anterior tienen una aplicación directa a la teoría que estamos estudiando. A continuación mencionaremos algunos ejemplos importantes.

3.3.1. Falta de memoria de la exponencial

Si por ejemplo el tiempo que falta para que llegue el próximo cliente es $T \sim \varepsilon(\lambda)$ y ya ha pasado un tiempo t_0 sin que llegue el próximo cliente, la falta de memoria indica que $P(T > t_0 + a \,|_{T>t_0}) = P(T > a)$. Esto quiere decir que la probabilidad de que falte esperar un tiempo mayor que a desde t_0 es la misma que la probabilidad de tener que esperar un tiempo mayor que a desde $t = 0$.

Análogamente si el tiempo que falta para que salga del sistema un cliente es $T \sim \varepsilon(\lambda)$ y ya ha pasado un tiempo t_0 sin que haya salido, la probabilidad de que haya que esperar un tiempo adicional mayor que a es la misma que la de que haya que esperar un tiempo mayor que a desde $t = 0$.

3.3.2. Tiempo hasta que lleguen los próximos n clientes

Supongamos que X_i para $1 \leq i \leq n$ es el tiempo transcurrido entre que llega el $i-$ésimo cliente y el $i+1-$ésimo cliente, y que todas ellas son independientes $X_i \sim \varepsilon(\lambda)$. Entonces el tiempo transcurrido entre que llega un cliente y los n próximos sería:

$$X = X_1 + \cdots + X_n \sim \Gamma(n, \lambda)$$

3.3.3. Tiempo hasta que salga del sistema el próximo cliente

Imaginemos una cola de tipo $M/M/s$ donde cada uno de los s servidores tiene una distribución exponencial de parámetro μ. Si X es el tiempo hasta que salga el próximo cliente y X_i para $1 \leq i \leq n$ es el tiempo hasta que salga el cliente atendido por el $i-$ésimo servidor, entonces es claro que:

$$X = \min_{1 \leq i \leq n} X_i$$

por lo que según la propiedad del mínimo para exponenciales, debe ser $X \sim \varepsilon(s\mu)$.

3.3.4. Tiempo hasta que es atendido un cliente

Bajo las mismas hipótesis que el ejemplo anterior supongamos que hay n clientes en el sistema y llega uno más a la misma:

$$\begin{array}{ccccccc} & & & & & & 1 \\ \boldsymbol{n+1} & n & \cdots & s+2 & s+1 & \vdots \\ & & & & & & s \end{array}$$

Para que el cliente $n+1$ comience a ser atendido deben salir $n-s+1$ clientes. Ya vimos que el tiempo que tarda en salir del sistema el cliente $i-$ésimo es $X_i \sim \varepsilon(s\mu)$, por lo que el tiempo hasta que sea atendido el cliente $n+1$ será:

$$X = X_1 + X_2 + \cdots + X_{n-s+1} \sim \Gamma(n+s-1, s\mu)$$

3.4. Ejercicios

1. Mediante las fórmulas de Little calcular el número esperado de servidores ocupados, para una cola $M/M/s$.

2. En un modelo $M/M/2$ donde los servidores tienen la misma tasa de atención $\mu = 2$ clientes por minuto, llega un cliente justo cuando hay otros dos haciendo cola para entrar en el mecanismo de servicio. ¿Qué tiempo medio transcurrirá hasta que el cliente recién llegado salga del sistema?

3. En una sastrería hay una sección de arreglo y reforma de la ropa vendida a sus clientes, que es atendida por un sastre. El número de clientes que requieren arreglos arriban a dicha sección con una distribución Poisson con una media de 24 clientes por hora. Debido a que el servicio es gratuito, todos los clientes están dispuestos a esperar el tiempo que sea necesario para poder utilizarlo. El tiempo de atención es en promedio de 2 minutos por cliente, siendo exponencial la distribución de los tiempos de servicio. Calcular:

   a) ¿Cuál es en promedio, el número de clientes en la sección?
   b) ¿Cuánto tiempo permanece, en promedio, un cliente en la sección?
   c) ¿Cuál es la probabilidad de que el sastre esté desocupado?
   d) ¿Cuál es en promedio, el número de clientes que están esperando recibir el servicio?

4. Un establecimiento de reparaciones, atendido por un solo operario, recibe un promedio de cuatro clientes por hora, los cuales traen pequeños aparatos a reparar. El mecánico los inspecciona para encontrar los defectos y muy a menudo puede arreglarlos de inmediato, o de otro modo emitir un diagnóstico. En promedio, todo le toma 6 minutos por aparato. Los arribos tienen una distribución Poisson y el tiempo de servicio tiene una distribución exponencial. Calcular:

   a) La probabilidad de que el taller esté vacío.
   b) La probabilidad de que tres clientes estén en el taller.
   c) La probabilidad de encontrar por lo menos un cliente en el taller.
   d) El número promedio de clientes en el taller.
   e) El tiempo promedio que un cliente debe permanecer en el taller.
   f) El número promedio de clientes que esperan ser atendidos.
   g) El tiempo promedio que un cliente debe esperar para ser atendido.

5. Teniendo en cuenta el ejercicio 4, considerar todas las suposiciones anteriores, excepto que si hay tres clientes en el taller, cualquier otro cliente que llegue se retirará.

   Determinar entonces:

   a) La probabilidad de que el taller esté vacío.

b) La probabilidad de que tres clientes estén en el taller.

c) La probabilidad de encontrar por lo menos un cliente en el taller.

d) El número promedio de clientes en el taller.

e) El tiempo promedio que un cliente debe permanecer en el taller.

f) El número promedio de clientes que esperan ser atendidos.

g) El tiempo promedio que un cliente debe esperar para ser atendido.

h) La cantidad promedio de clientes que se retiran sin ser atendidos.

Capítulo 4

Modelos del tipo $M/M/\cdots$

En este capítulo se estudiarán algunos de los modelos más importantes en teoría de colas, que son aquellos que suponen al tiempo de llegada entre dos clientes, así como también al tiempo de servicio, variables aleatorias de de tipo exponencial.

Los modelos más comunes son $M/M/1$, $M/M/s$, $M/M/1/K$, $M/M/s/K$, $M/M/1/\infty/H$, $M/M/s/\infty/H$ con y sin repuestos, y por último $M/M/\infty$.

En cada uno de ellos se supondrá que la distribución del tiempo de llegada entre dos clientes es exponencial de parámetro $(\lambda_n)_{n\in\mathbb{N}_0}$ y que la distribución del tiempo de servicio es $(\mu_n)_{n\in\mathbb{N}}$. También se supondrá que la disciplina de la cola es $FIFO$ y se realizará el análisis buscando establecer el comportamiento del modelo en tiempo estacionario.

Recordemos las expresiones matemáticas más importantes que se utilizarán en el desarrollo:

1. Condición necesaria y suficiente para la estacionariedad:

$$c_n = \frac{\lambda_{n-1}\cdots\lambda_0}{\mu_n\cdots\mu_1} \qquad\qquad \sum_{n=1}^{+\infty} c_n < +\infty$$

2. Distribución de probabilidades a tiempo estacionario:

$$p_n = c_n \cdot p_0 \qquad\qquad p_0 = \frac{1}{1+\sum\limits_{n=1}^{+\infty} c_n}$$

3. Las Fórmulas de Little:

$$\left.\begin{aligned}\overline{\lambda} &= \sum_{n=0}^{+\infty} p_n\cdot\lambda_n \\ \overline{\mu} &= \frac{1}{1-p_0}\cdot\sum_{n=1}^{+\infty}\mu_n\cdot p_n\end{aligned}\right\} \Rightarrow \left.\begin{aligned}L &= \overline{\lambda}\cdot W \\ L_q &= \overline{\lambda}\cdot W_q \\ W &= W_q + \frac{1}{\overline{\mu}}\end{aligned}\right.$$

4.1. El modelo $M/M/1$

En este modelo se supone que $\lambda_n = \lambda\,\forall\, n \in \mathbb{N}_0$ y $\mu_n = \mu\,\forall\, n \in \mathbb{N}$. Además que se dispone de un único servidor para atender a los clientes, que la capacidad de la cola es infinita y la población potencial también.

Para empezar:

$$c_n = \frac{\lambda_{n-1}\cdots\lambda_0}{\mu_n\cdots\mu_1} = \frac{\lambda^n}{\mu^n} = \left(\frac{\lambda}{\mu}\right)^n \qquad\qquad \rho = \frac{\lambda}{\mu} \qquad\qquad \sum_{n=1}^{+\infty} c_n = \sum_{n=1}^{+\infty}\rho^n$$

por lo que el proceso será estacionario si y sólo si $|\rho| < 1$ y como $\rho > 0$ esto equivale a que $0 < \rho < 1$.

Vemos entonces que la estacionariedad se produce siempre que $\lambda < \mu$ lo cual es lógico si pensamos que para poder evitar el crecimiento indiscriminado de la cola, es necesario poder atender más clientes de los que llegan. Por otra parte, como:

$$\sum_{n=1}^{+\infty} \rho^n = \frac{\rho}{1-\rho}$$

entonces:

$$p_0 = \frac{1}{1+\frac{\rho}{1-\rho}} = 1-\rho$$
$$p_n = \rho^n \cdot (1-\rho)$$

Como los λ y μ son resulta fácil deducir:

$$\overline{\lambda} = \sum_{n=0}^{+\infty} \lambda_n \cdot p_n = \sum_{n=0}^{+\infty} \lambda \cdot p_n = \lambda \sum_{n=0}^{+\infty} p_n = \lambda$$
$$\overline{\mu} = \frac{1}{1-p_0} \sum_{n=1}^{+\infty} \mu_n \cdot p_n = \frac{\mu}{\frac{\lambda}{\mu}} \sum_{n=1}^{+\infty} p_n = \frac{\mu^2}{\lambda} \cdot (1-p_0) = \frac{\mu^2}{\lambda} \cdot \frac{\lambda}{\mu} = \mu$$

Además:

$$L = E[N] = \sum_{n=0}^{+\infty} n \cdot p_n = \sum_{n=1}^{+\infty} n \cdot \rho^n (1-\rho)$$
$$= (1-\rho) \sum_{n=1}^{+\infty} n \cdot \rho^n = (1-\rho) \cdot \rho \sum_{n=1}^{+\infty} n \cdot \rho^{n-1}$$

La última serie se puede resolver por integración ya que:

$$\int s(\rho)\, d\rho = \sum_{n=1}^{+\infty} \rho^n = \frac{\rho}{1-\rho}$$

por lo que recuperamos $s(\rho)$ derivando ambos términos:

$$\sum_{n=1}^{+\infty} n\rho^{n-1} = s(\rho) = \frac{1}{(1-\rho)^2}$$

Finalmente:

$$L = \cancel{(1-\rho)} \cdot \rho \frac{1}{\cancel{(1-\rho)}^2} = \frac{\rho}{1-\rho} = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} = \frac{\lambda}{\mu-\lambda}$$

Podemos ahora deducir W a partir de la primera fórmula de little:

$$\frac{\lambda}{\mu-\lambda} = L = \lambda \cdot W \Rightarrow W = \frac{1}{\mu-\lambda}$$

Por la tercer fórmula de little:

$$W_q = W - \frac{1}{\overline{\mu}} = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\mu-(\mu-\lambda)}{\mu(\mu-\lambda)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

y de esto último se deduce:

$$L_q = \lambda \cdot W_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Ahora determinaremos la distribución $\mathcal{W}$ del tiempo que permanece un cliente elegido al azar en el sistema. Es claro que el tiempo depende de cuántos clientes haya en el sistema al momento de entrar nuestro cliente elegido al azar. Por ejemplo si n clientes cuando ingresa el $n+1$, el mismo tendrá que esperar a que sirvan a estos n primero clientes antes de ingresar al mecanismo de servicio, luego de lo cual tendrá que esperar que lo atiendan a él mismo. Por lo tanto deberá esperar $n+1$ veces el tiempo de un cliente que llega cuando no hay nadie en el sistema.

Si llamamos $(\mathcal{W}_n)_{n\in\mathbb{N}_0}$ al tiempo que permanecería un cliente al azar cuando ingresa al sistema y en el mismo ya había otros n clientes, y denotamos por $(W_i)_{1\leq i\leq n+1}$ al tiempo que tarda en ser atendido por el servidor el $i-$ésimo cliente, entonces cada $W_i \sim \varepsilon(\mu)$ y $\mathcal{W}_n = W_1 + \cdots + W_{n+1}$ es suma de $n+1$ exponenciales independientes de parámetro λ. Por la propiedad 3.2.2.4 se tiene que $\mathcal{W}_n \sim \Gamma(n+1,\mu)$ para todo $n \in \mathbb{N}_0$.

Juntando lo anterior con el teorema de la probabilidad total, podemos obtener la distribución de $\mathcal{W}$ como sigue:

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

Y como:

$$\sum_{n=0}^{+\infty}\frac{(\lambda x)^n}{n!} = e^{\lambda x}$$

entonces:

$$\begin{aligned} P(\mathcal{W}\leq t) &= \int_0^t (\mu-\lambda)\,e^{-(\mu-\lambda)x}\,dx \\ &= -e^{-(\mu-\lambda)x}\Big|_0^t \\ &= 1-e^{-(\mu-\lambda)t} \end{aligned}$$

Pero entonces:

$$\mathcal{W} \sim \varepsilon(\mu-\lambda)$$

Para el tiempo $\mathcal{W}_q$ que un cliente permanece en la cola, hay que discriminar si el cliente entra cuando en el sistema no hay nadie o si lo hace cuando ya hay otros clientes, pues en el primer caso el tiempo en la cola será 0 y en el otro caso tendrá que esperar a que el servidor despache a los clientes que están por delante de él. Esto es:

$$P(\mathcal{W}_q \leq t) = P(\mathcal{W}_q = 0) + P(0 < W_q \leq t)$$

El primer término de la suma es sencillo pues:

$$P(\mathcal{W}_q = 0) = p_0 = 1-\frac{\lambda}{\mu}$$

ya que un cliente que entra no espera nada sólo cuando en el sistema no hay clientes.

El segundo término podemos descomponerlo de esta forma:

$$P(0<W_q\leq t) = P(0<\mathcal{W}_q\leq t\,|_{N=0})\cdot P(N=0) + P(0<\mathcal{W}_q\leq t\,|_{N\geq 1})\cdot P(N\geq 1)$$

Si $N = 0$ entonces la probabilidad de que $\mathcal{W}_q > 0$ es nula, por lo que el primer término vuela, resultando:

$$P\left(0 < W_q \leq t\right) = P\left(0 < \mathcal{W}_q \leq t\,|_{N\geq 1}\right) \cdot P\left(N \geq 1\right)$$

Ahora bien, si sabemos que $N \geq 1$ entonces el evento $\mathcal{W}_q > 0 \wedge \mathcal{W}_q \leq t$ coincide con el evento $\mathcal{W}_q \leq t$ pues $\mathcal{W}_q > 0$ es verdadero si $N \geq 1$. Pero entonces:

$$P\left(0 < \mathcal{W}_q \leq t\,|_{N\geq 1}\right) = P\left(\mathcal{W}_q \leq t\,|_{N\geq 1}\right)$$

Por otra parte el tiempo que un cliente permanece en la cola, sabiendo que $N \geq 1$ es el mismo tiempo que un cliente permanecería en el sistema, sabiendo que $N \geq 0$. Por lo tanto:

$$P\left(\mathcal{W}_q \leq t\,|_{N\geq 1}\right) = P\left(\mathcal{W} \leq t\,|_{N\geq 0}\right)$$

Pero saber que $N \geq 0$ es lo mismo que no saber nada, pues siempre ocurre que $N \geq =$, de donde:

$$P\left(\mathcal{W} \leq t\,|_{N\geq 0}\right) = P\left(\mathcal{W} \leq t\right)$$

Por lo expuesto anteriormente:

$$\begin{aligned} P\left(0 < W_q \leq t\right) &= P\left(0 < \mathcal{W}_q \leq t\,|_{N\geq 1}\right) \cdot P\left(N \geq 1\right) \\ &= P\left(\mathcal{W} \leq t\right) \cdot \left(1 - p_0\right) \\ &= \left(1 - e^{-(\mu-\lambda)t}\right) \cdot \frac{\lambda}{\mu} \\ &= \frac{\lambda}{\mu} - \frac{\lambda}{\mu} e^{-(\mu-\lambda)t} \end{aligned}$$

Juntando todo resulta:

$$P\left(\mathcal{W}_q \leq t\right) = 1 - \frac{\cancel{\lambda}}{\cancel{\mu}} + \frac{\cancel{\lambda}}{\cancel{\mu}} - \frac{\lambda}{\mu} e^{-(\mu-\lambda)t} = 1 - \frac{\lambda}{\mu} e^{-(\mu-\lambda)t}$$

Un dato interesante es que $\mathcal{W}_q$ no es continua ni discreta, sino una mezcla de ambas. Discreta es claro que no puede ser, y continua tampoco pues:

$$P\left(\mathcal{W}_q = 0\right) = 1 - \frac{\lambda}{\mu}$$

y en las continuas las probabilidades puntuales son todas cero.

4.1.1. Resumen de resultados

Resumiendo, en una cola $M/M/1$:

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \cdot \left(1 - \frac{\lambda}{\mu}\right) \qquad L = \frac{\lambda}{\mu - \lambda} \qquad L_q = \frac{\lambda^2}{\mu\left(\mu - \lambda\right)} \qquad W = \frac{1}{\mu - \lambda} \qquad W_q = \frac{\lambda}{\mu\left(\mu - \lambda\right)}$$

El tiempo que un cliente permanece en el sistema es $\mathcal{W} \sim \varepsilon\left(\mu - \lambda\right)$ y su distribución es:

$$P\left(\mathcal{W} \leq t\right) = 1 - e^{-(\mu-\lambda)t}$$

El tiempo que un cliente permanece en la cola es $\mathcal{W}_q$, es una variable mixta con distribución:

$$P\left(\mathcal{W}_q \leq t\right) = 1 - \frac{\lambda}{\mu} e^{-(\mu-\lambda)t}$$

4.1.2. Ejemplo de aplicación

En un supermercado con una única caja cuyo tiempo de atención a cada cliente se distribuye exponencialmente con una media de 5 minutos por cliente, se sabe que llegan a la caja un promedio de 9 clientes por hora.

1. ¿Qué probabilidad hay de que haya más de dos personas esperando en la cola?
2. ¿Qué tiempo medio permanece en el sistema un cliente? ¿Qué relación hay con el tiempo que tarda en ser atendido una vez que entra en la caja?
3. ¿Cuál es la probabilidad de que un cliente permanezca en el sistema más de 7 minutos?
4. ¿Cuál es el tamaño promedio de la cola?
5. Repetir los puntos anteriores si en lugar de 9 clientes por hora llegaran a la caja 11 clientes por hora. ¿Qué ocurre con la intensidad de tráfico? ¿Hay indicios para suponer que el sistema se encuentra funcionando al límite de sus posibilidades? ¿Qué ocurriría si llegara a ingresar un cliente por hora más?

Solución:

Conviene medir las tasas en horas, pues de esta forma el mecanismo de servicios despacharía una media de 12 clientes por hora, y estarían llegando a razón de 9 clientes por hora, por lo que:

$$\lambda = 9 \qquad\qquad \mu = 12$$

La intensidad de tráfico es:

$$\rho = \frac{9}{12} = \frac{3}{4} = 0{,}75 < 1$$

por lo que el proceso es estacionario y se pueden utilizar los resultados analizados para este modelo $M/M/1$.

Primero se nos pide:

$$\begin{aligned} P(N \geq 4) &= 1 - P(N \leq 3) \\ &= 1 - \sum_{n=0}^{3} p_n = 1 - \sum_{n=0}^{3} \left(\frac{\lambda}{\mu}\right)^n \cdot \left(1 - \frac{\lambda}{\mu}\right) \\ &= 1 - \frac{1}{4} \sum_{n=0}^{3} \left(\frac{3}{4}\right)^n = 1 - \frac{1}{4} \frac{1 - \left(\frac{3}{4}\right)^4}{1 - \frac{3}{4}} \\ &= \left(\frac{3}{4}\right)^4 = \frac{81}{256} \approx 0{,}3164 \end{aligned}$$

El tiempo medio que permanece en el sistema un cliente es:

$$W = \frac{1}{\mu - \lambda} = \frac{1}{12 - 9} = \frac{1}{3}$$

medido en horas, es decir 20 minutos. El tiempo medio de servicio es:

$$W_s = \frac{1}{\overline{\mu}} = \frac{1}{\mu} = \frac{1}{12}$$

medido en horas, es decir 5 minutos, por lo que la relación pedida es:

$$\frac{W}{W_s} = \frac{20\,min}{5\,min} = 4$$

cosa que indica que el tiempo medio que pasa desde que un cliente llega a la fila hasta que se retira del establecimiento es 4 veces superior al tiempo que tarda en ser atendido en la caja.

Lo tercero que se pide es:

$$\begin{aligned} P\left(\mathcal{W} > \frac{7}{60}\right) &= 1 - P\left(\mathcal{W} \leq \frac{7}{60}\right) \\ &= 1 - \left[1 - e^{-(12-9)\cdot\frac{7}{60}}\right] \\ &= e^{-\frac{7}{20}} \approx 0{,}7047 \end{aligned}$$

El tamaño promedio de la cola es:

$$L_q = \frac{\lambda^2}{\mu\,(\mu - \lambda)} = \frac{81}{12 \cdot 3} = \frac{9}{4} = 2{,}25$$

es decir que se espera haya aproximadamente 2 personas en la cola.

El punto 5) se deja como ejercicio para fijar ideas.

4.2. El modelo $M/M/s$

En este modelo se supone que $\lambda_n = \lambda \,\forall\, n \in \mathbb{N}_0$ y $\mu_n = \mu \,\forall\, n \in \mathbb{N}$. Además que se dispone de s servidores para atender a los clientes, que la capacidad de la cola es infinita y la población potencial también.

Para empezar:

$$\lambda_n = \lambda \,(\forall\, n \in \mathbb{N}_{\not\vdash})$$

Pero no ocurre lo mismo para u_n, pues debemos tener presente lo siguiente:

1. Si hubiera un sólo cliente en el sistema, $n = 1$, entonces aunque haya s servidores la tasa de atención sería $\mu_1 = \mu$, ya que los demás servidores están desocupados.

2. Para $n = 2$ clientes, ambos están siendo atendidos por lo que la tasa de atención se duplica a $\mu_2 = 2\mu$.

3. Si hubieran $n = s - 1$ clientes $\mu_{s-1} = (s-1) \cdot \mu$.

4. Para $n \geq s$ clientes todos los servidores estarán ocupados y por lo tanto $\mu_n = s \cdot \mu$.

Resumiendo:

$$\mu_n = \begin{cases} n\mu & \text{, si } 1 \leq n \leq s \\ s\mu & \text{, si } n \geq s+1 \end{cases}$$

Esto impacta en los coeficientes c_n:

$$\begin{aligned} c_1 &= \frac{\lambda_0}{\mu_1} = \frac{\lambda}{\mu} \\ c_2 &= \frac{\lambda_1\lambda_0}{\mu_2\mu_1} = \frac{\lambda^2}{2\mu^2} = \frac{1}{2\cdot 1} \cdot \left(\frac{\lambda}{\mu}\right)^2 = \frac{1}{2!}\left(\frac{\lambda}{\mu}\right)^2 \\ c_3 &= \frac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1} = \frac{\lambda^3}{3\cdot 2\cdot 1\cdot \mu^3} = \frac{1}{3!}\left(\frac{\lambda}{\mu}\right)^3 \\ &\vdots \\ c_s &= \frac{\lambda_{s-1}\cdots\lambda_0}{\mu_s\cdots\mu_1} = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \end{aligned}$$

por lo que para $1 \leq n \leq s$:

$$c_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n$$

Supongamos que $n = s + k$ con $k \geq 1$, entonces:

$$\begin{aligned} c_n &= \frac{\lambda_{s+k-1} \cdots \lambda_s \cdots \lambda_{s-1} \cdots \lambda_0}{\mu_{s+k} \cdots \mu_{s+1} \cdots \mu_s \cdots \mu_1} \\ &= \frac{\lambda_{s-1} \cdots \lambda_0}{\mu_s \cdots \mu_1} \cdot \frac{\lambda_{s+k-1} \cdots \lambda_s}{\mu_{s+k} \cdots \mu_{s+1}} \\ &= \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \cdot \frac{1}{s^k} \cdot \left(\frac{\lambda}{\mu}\right)^k \\ &= \frac{1}{s!} \cdot \frac{1}{s^k} \cdot \left(\frac{\lambda}{\mu}\right)^{s+k} \\ &= \frac{1}{s! s^{n-s}} \cdot \left(\frac{\lambda}{\mu}\right)^n \end{aligned}$$

por lo que si $n \geq s + 1$:

$$c_n = \frac{1}{s! s^{n-s}} \cdot \left(\frac{\lambda}{\mu}\right)^n$$

De esta forma, podemos resumir una expresión para c_n como sigue:

$$c_n = \begin{cases} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n & \text{, si } 1 \leq n \leq s \\ \frac{1}{s! s^{n-s}} \cdot \left(\frac{\lambda}{\mu}\right)^n & \text{, si } n \geq s + 1 \end{cases}$$

Esto basta para conocer:

$$p_n = c_n \cdot p_0$$

pero aún falta determinar:

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

por lo que será necesario obtener el valor de la serie:

$$\begin{aligned} \sum_{n=1}^{+\infty} c_n &= \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s+1}^{+\infty} \frac{1}{s! s^{n-s}} \cdot \left(\frac{\lambda}{\mu}\right)^n \\ &= \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s+1}^{+\infty} \frac{1}{s! s^{n-s}} \cdot \left(\frac{\lambda}{\mu}\right)^n \\ &= \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{s^s}{s!} \sum_{n=s+1}^{+\infty} \left(\frac{\lambda}{s\mu}\right)^n \\ &= \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{s^s}{s!} \sum_{n=0}^{+\infty} \left(\frac{\lambda}{s\mu}\right)^{n+s+1} \\ &= \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{s^s}{s!} \left(\frac{\lambda}{s\mu}\right)^{s+1} \sum_{n=0}^{+\infty} \left(\frac{\lambda}{s\mu}\right)^n \\ &= \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{s^s}{s!} \left(\frac{\lambda}{s\mu}\right)^{s+1} \cdot \frac{1}{1 - \frac{\lambda}{s\mu}} \end{aligned}$$

Trabajando con la última expresión resulta:

$$\sum_{n=1}^{+\infty} c_n = \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{\lambda^{s+1}}{s! \mu^s (s\mu - \lambda)}$$

Esto permite determinar:

$$p_0 = \frac{1}{1 + \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{\lambda^{s+1}}{s! \mu^s (s\mu - \lambda)}}$$

Resulta más sencillo determinar L_q antes que L pues:

$$\begin{aligned} L_q &= 0 \cdot (p_0 + p_1 + \cdots + p_s) + \sum_{n=s+1}^{+\infty} (n-s) \cdot p_n \\ &= \sum_{n=s+1}^{+\infty} (n-s) \cdot \frac{1}{s! s^{n-s}} \cdot \left(\frac{\lambda}{\mu}\right)^n p_0 \\ &= \frac{1}{s!} \cdot \left(\frac{\lambda}{\mu}\right)^s p_0 \sum_{n=s+1}^{+\infty} (n-s) \cdot \left(\frac{\lambda}{s\mu}\right)^{n-s} \\ &= \frac{1}{s!} \cdot \left(\frac{\lambda}{\mu}\right)^s \frac{\lambda}{s\mu} p_0 \sum_{n=1}^{+\infty} n \cdot \left(\frac{\lambda}{s\mu}\right)^{n-1} \end{aligned}$$

Si llamamos:

$$\rho = \frac{\lambda}{s\mu} \Rightarrow \sum_{n=1}^{+\infty} n \cdot \left(\frac{\lambda}{s\mu}\right)^{n-1} = \sum_{n=1}^{+\infty} n \cdot \rho^{n-1}$$

y como previamente se demostró que:

$$\sum_{n=1}^{+\infty} n\rho^{n-1} = \frac{1}{(1-\rho)^2} = \frac{1}{\left(1 - \frac{\lambda}{s\mu}\right)^2} = \frac{(s\mu)^2}{(s\mu - \lambda)^2}$$

entonces:

$$L_q = \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \frac{\lambda}{s\mu} \frac{(s\mu)^2}{(s\mu - \lambda)^2} p_0 = \frac{\lambda^{s+1}}{(s-1)! u^{s-1} (s\mu - \lambda)^2} \cdot p_0$$

También es simple la obtención de W_q a partir de la segunda fórmula de little:

$$W_q = \frac{L_q}{\lambda} = \frac{\lambda^s}{(s-1)! u^{s-1} (s\mu - \lambda)^2} \cdot p_0$$

Y las demás surgen muy fácilmente a partir de las dos anteriores:

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

Es posible obtener las distribuciones de $\mathcal{W}$ y $\mathcal{W}_q$ pero el cálculo resulta bastante más complicado que el correspondiente al modelo $M/M/1$, por lo que incluiremos los mismos en el resumen sin deducirlos previamente aquí.

4.2.1. Resumen de resultados

Resumiendo, en una cola $M/M/s$:

$$c_n = \begin{cases} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n & \text{, si } 1 \le n \le s \\ \frac{1}{s! s^{n-s}} \cdot \left(\frac{\lambda}{\mu}\right)^n & \text{, si } n \ge s+1 \end{cases}$$

$$\sum_{n=1}^{+\infty} c_n = \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{\lambda^{s+1}}{s! \mu^s (s\mu - \lambda)}$$

$$p_n = c_n \cdot p_0$$

$$p_0 = \frac{1}{1 + \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{\lambda^{s+1}}{s! \mu^s (s\mu - \lambda)}}$$

$$L_q = \frac{\lambda^{s+1}}{(s-1)! u^{s-1} (s\mu - \lambda)^2} \cdot p_0$$

$$W_q = \frac{\lambda^s}{(s-1)! u^{s-1} (s\mu - \lambda)^2} \cdot p_0$$

$$W = \frac{\lambda^s}{(s-1)! u^{s-1} (s\mu - \lambda)^2} \cdot p_0 + \frac{1}{\mu}$$

$$L = \frac{\lambda^{s+1}}{(s-1)! u^{s-1} (s\mu - \lambda)^2} \cdot p_0 + \frac{\lambda}{\mu}$$

La expresión para la distribución de $\mathcal{W}_q$ es:

$$P\left(\mathcal{W}_q \leq t\right) = 1 - \frac{\lambda^s \cdot p_0}{(s-1)!\mu^{s-1}\left(s\mu - \lambda\right)} \cdot e^{-(s\mu-\lambda)t} \cdot I_{[0,+\infty)}\left(t\right)$$

- La expresión para la distribución de $\mathcal{W}$ es:
  - Si $\frac{\lambda}{\mu} \neq s-1$:

$$P\left(\mathcal{W} \leq t\right) = \left[1 + \frac{\lambda - s\mu + \mu P\left(\mathcal{W}_q = 0\right)}{s\mu - \lambda - \mu} e^{-\mu t} + \frac{\lambda^s \cdot p_0}{(s-1)!\mu^{s-2}\left(s\mu - \lambda\right)\left(s\mu - \lambda - \mu\right)} \cdot e^{-(s\mu-\lambda)t}\right] \cdot I_{[0,+\infty)}\left(t\right)$$

  - Si $\frac{\lambda}{\mu} = s-1$:

$$P\left(\mathcal{W} \leq t\right) = \left[1 - \left(1 + \frac{\lambda^s t \cdot p_0}{(s-1)!\mu^{s-2}\left(s\mu - \lambda\right)}\right) e^{-\mu t}\right] \cdot I_{[0,+\infty)}\left(t\right)$$

4.2.2. Ejemplo de aplicación

Supongamos que un supermercado dispone de 3 cajas para atender a sus clientes, los cuales a medida que van llegando se ubican en una fila común, y van ingresando a los cajeros a medida que se van desocupando. La tasa de llegada de clientes es de 1 cliente cada 2 minutos y cada cajero tarda en atender a un cliente, un promedio de 5 minutos.

Según la información anterior se deprende que $\lambda = 0{,}5$ clientes por minuto y $\mu = 0{,}2$ clientes por minuto para cada cajero. Y se deduce que:

$$c_1 = \frac{0{,}5}{0{,}2} = 2{,}5 \qquad c_2 = \frac{1}{2} \cdot (2{,}5)^2 = 3{,}125 \qquad c_3 = \frac{1}{6} \cdot (2{,}5)^3 = \frac{125}{48} \approx 2{,}6042$$

y para $n \geq 4$ se obtiene:

$$c_n = \frac{2{,}5^n}{6 \cdot 3^{n-3}} = \frac{27}{6} \cdot \left(\frac{5}{6}\right)^n = 4{,}5 \cdot \left(\frac{5}{6}\right)^n$$

Antes de poder establecer las variables del sistema debemos hallar p_0:

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

por lo que primero debemos establecer el valor de la serie:

$$\begin{aligned}
\sum_{n=1}^{+\infty} c_n &= \sum_{n=1}^{s} \frac{1}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n + \frac{\lambda^{s+1}}{s!\mu^s\left(s\mu - \lambda\right)} \\
&= \frac{5}{2} + \frac{25}{8} + \frac{125}{48} + \frac{\left(\frac{1}{2}\right)^4}{6 \cdot \left(\frac{1}{5}\right)^3 \cdot \left(\frac{3}{5} - \frac{1}{2}\right)} \\
&= \frac{120}{48} + \frac{150}{48} + \frac{125}{48} + \frac{625}{48} \\
&= \frac{85}{4}
\end{aligned}$$

por lo que:

$$p_0 = \frac{1}{1 + \frac{85}{4}} = \frac{4}{89}$$

Así:

$$p_0 = \frac{4}{89} \qquad p_1 = \frac{10}{89} \qquad p_2 = \frac{25}{178} \qquad p_3 = \frac{125}{1068}$$

y para $n \geq 4$:

$$p_n = \frac{9}{2} \cdot \left(\frac{5}{6}\right)^n$$

Y las demás variables del sistema resultan:

$$L_q = \frac{\left(\frac{1}{2}\right)^4}{2\left(\frac{1}{5}\right)^2\left(\frac{3}{5}-\frac{1}{2}\right)^2} \cdot \frac{4}{89} = \frac{1250}{16} \cdot \frac{4}{89} = \frac{625}{178} \approx 3{,}51$$

$$W_q = \frac{\left(\frac{1}{2}\right)^3}{2\left(\frac{1}{5}\right)^2\left(\frac{3}{5}-\frac{1}{2}\right)^2} \cdot \frac{4}{89} = \frac{1250}{8} \cdot \frac{4}{89} = \frac{625}{89} \approx 7{,}022$$

$$W = \frac{\left(\frac{1}{2}\right)^3}{2\left(\frac{1}{5}\right)^2\left(\frac{3}{5}-\frac{1}{2}\right)^2} \cdot \frac{4}{89} + \frac{1}{\frac{1}{5}} = \frac{625}{89} + 5 \approx 12{,}022$$

$$L = \frac{\left(\frac{1}{2}\right)^4}{2\left(\frac{1}{5}\right)^2\left(\frac{3}{5}-\frac{1}{2}\right)^2} \cdot \frac{4}{89} + \frac{\frac{1}{2}}{\frac{1}{5}} = \frac{625}{178} + \frac{5}{2} = \frac{535}{89} \approx 6{,}011$$

Vemos entonces que se espera haya 6 clientes en el sistema cuando el mismo se encuentra en proceso estacionario, un promedio de 3 a 4 personas en la cola, y el tiempo de espera en la cola de aproximadamente 7 minutos, con un promedio de tiempo en el sistema de 12 minutos, para un cliente elegido al azar.

4.3. Problemas y Ejercicios

1. Calcular la intensidad de tráfico correspondiente al ejemplo de aplicación 4.2.2. Repita el ejemplo de aplicación pero agregando una caja adicional. Vuelva a calcular la intensidad de tráfico y compare los nuevos valores obtenidos para las variables del sistema. ¿Considera que se ha descongestionado el mismo?

2. Supongamos que al sistema ingresan a razón de 12 clientes por hora y son atendidos por dos servidores, cada uno con una capacidad de atención de 7 clientes por hora. Se desea comparar el rendimiento de este sistema con otro similar, pero disponiendo de un sólo servidor con capacidad de atender a 14 clientes por hora. Para ello, en cada uno de los modelos planteados se pide hallar:

   a) El número esperado de clientes en el sistema.
   b) El número esperado de clientes en la cola.
   c) El tiempo promedio de permanencia en la cola.
   d) El tiempo promedio de permanencia en el sistema.
   e) La intensidad de tráfico.

   Según los valores obtenidos: ¿Es cierta la hipótesis afirmando que da lo mismo tener dos cajas que una sola pero el doble de rápida?

   En caso de haber respondido afirmativamente la pregunta anterior, argumentar por qué. En caso de haber optado por la negativa, explicar en qué se diferencian ambos modelos a la luz de los datos obtenidos.

3. Para sacar el registro hay que completar una serie de trámites sucesivamente, según la descripción que se hará a continuación. Primero se realiza un apto físico en la oficina de un médico que tarda aproximadamente 15 minutos por paciente, y donde llegan a razón de 3 pacientes por hora. Luego hay que realizar un psicotécnico en otra oficina donde hay dos psicólogos atendiendo, cada uno a razón de un paciente cada 20 minutos. La tasa de llegada de pacientes por hora a esta oficina es la misma que la tasa de salida que hay en la oficina del médico. Los pacientes que salen del psicotécnico van a realizar el examen teórico donde hay tres personas atendiendo, y cada uno de ellos tarda unos 25 minutos en tomar examen a cada aspirante. Finalmente,

todos los que aprobaron el examen teórico pasan a dar la prueba de manejo, donde hay dos autos disponibles y se tarda 15 minutos por persona. Para analizar el flujo máximo de gente por el sistema se supondrá que no hay reprobados en ninguna de las instancias, por lo que todos los que terminan de salir del examen de manejo pasan a una ventanilla que es la que expide finalmente el registro. Es una única ventanilla con capacidad de atender 10 personas por hora. Para cada una de las instancias, entendidas como un sistema de una cola, se pide calcular:

a) El número esperado de clientes en el sistema.

b) El número esperado de clientes en la cola.

c) El tiempo promedio de permanencia en la cola.

d) El tiempo promedio de permanencia en el sistema.

e) La intensidad de tráfico.

A partir de los resultados obtenidos calcular, para el sistema global del trámite completo:

a) El número esperado de personas realizando el trámite de sacar el registro.

b) El tiempo promedio que un aspirante elegido al azar se tarda en sacar el registro.

c) El tiempo promedio que un aspirante elegido al azar se encuentra haciendo filas.

d) La intensidad de tráfico global.

Según el análisis realizado: ¿Está sobrecargado el sistema o funciona fluidamente?

4. Un banco está desarrollando la prestación de un nuevo servicio, para lo cual ha habilitado una ventanilla. Como el desarrollo del mismo está basado en una campaña publicitaria que hace mención al mínimo tiempo de espera que se requiere, el gerente de la sucursal ha decidido encarar el estudio científico del problema a fin de no exponerse a un fracaso. Hasta ahora se cuenta con los siguientes datos:

   - *Lapso medio entre arribo de usuarios*: 8 minutos, distribución exponencial.
   - *Tiempo medio de atención en ventanilla*: 8 minutos, distribución exponencial.

   Determinar:

   a) La probabilidad de esperar.

   b) La longitud promedio de la cola.

   c) La velocidad promedio de arribos que haría que el tiempo de espera en la cola supera los 4 minutos.

5. Una empresa tiene cuatro máquinas cortadoras de césped. Las mismas se rompen o necesitan mantenimiento cada 15 días —*distribución exponencial*. Para su atención y mantenimiento tiene un empleado que en promedio tarda 7 días con cada máquina. En promedio, por cada día de trabajo, las máquinas reportan un ingreso de $50. Se desea saber:

   a) El número promedio de máquinas funcionando.

   b) El porcentaje de tiempo que el empleado se encuentra inactivo.

   c) Cuánto tiempo, en promedio, estará en funcionamiento una máquina.

   d) Existe la posibilidad de contratar una persona más, que cobra $700 por mes y que tarda lo mismo que el empleado. Considerando meses de 24 días, ¿conviene contratar al nuevo empleado?

Capítulo 5

Simulación

5.1. Introducción

Uno de los elementos fundamentales en la simulación de colas es la capacidad de poder emular el comportamiento de una variable aleatoria, y poder generar una muestra de valores de la misma.

Por ejemplo, supongamos que el tiempo T_n entre que llega un cliente y el próximo — *cuando en el sistema hay n clientes* — es una *v.a.* exponencial de parámetro λ_n, tal como sucede en los procesos de nacimiento y muerte. Para fijar ideas pensemos que $\lambda_n = 6$ clientes por minuto, de modo que $T_n \sim \varepsilon(\lambda_n)$ con y $E[T_n] = \frac{1}{6}$. Esto significa que el tiempo promedio que tarda en ingresar al sistema un cliente cuando en el sistema hay n clientes es aproximadamente un sexto de minuto, es decir 10 segundos.

Pero en una simulación sería incorrecto esperar exactamente 10 segundos hasta que llegue el próximo cliente, pues aunque se sepa que el tiempo promedio es 10 segundos, rara vez ocurra que llegue exactamente cuando se espera en promedio que lo haga. Tranquilamente podría pasar que llegue en 12, 15, 8 o 25 segundos. Y la única manera de simular esto decentemente es disparar una variable aleatoria que se comporte como T_n, para pedirle un valor que sea interpretado como el tiempo que hay que esperar en segundos hasta la llegada del próximo cliente.

Si disparamos T_n y la misma proporciona un valor $T_n = 13$, entonces habrá que esperar 13 segundos. Una vez llegado el cliente próximo, habrán pasado 13 segundos, por lo que es probable que ya los servidores hayan despachado algunos clientes. Esto significa que al momento de llegar este cliente en el sistema va a haber un número m de clientes, no necesariamente $m = n+1$. Si queremos emular el tiempo de llegada del próximo cliente habrá que disparar ahora una T_m variable aleatoria exponencial de parámetro λ_m.

Vemos entonces que la capacidad de saber generar una muestra aleatoria de variables con una determinada distribución es algo muy importante para el proceso de simulación. Casi cualquier cosa que se pretenda simular será modelada disparando una variable aleatoria, por eso se necesita disponer de algún método para generar *muestras aleatorias* siguiendo una determinada distribución.

Precisamente esto analizaremos en las siguientes secciones.

5.2. Generación de una muestra aleatoria $U([0,1])$

Para generar una muestra aleatoria uniforme del intervalo $[0,1]$, se puede hacer sencillamente a partir de cualquier función generadora de números aleatorios naturales comprendidos en un determinado rango. Supongamos que disponemos de una función $rand()$ que devuelve un número natural entre 1 y N con N par. Si generamos una muestra aleatoria:

$$r_1 := rand() \qquad r_2 = rand() \qquad \cdots \qquad r_m = rand() \qquad \cdots$$

entonces cada uno de los r_i es un número entero entre 1 y N.

Podemos ahora definir al $i-$ésimo dígito del número d como:

$$d_i = \begin{cases} 0 & \text{, si } 1 \leq r_i \leq \frac{N}{2} \\ 1 & \text{, si } \frac{N}{2} + 1 \leq r_i \leq N \end{cases}$$

y el número:

$$d = \sum_{i=1}^{+\infty} d_i \cdot 2^{-i}$$

sería el primer número de la muestra.

En la práctica es imposible determinar d a partir de la serie infinita, por lo que debemos cortar en algún $m \in \mathbb{N}$. En la siguiente sección se analizarán posibles criterios para determinar un valor de m adecuado.

Repitiendo el procedimiento descripto tantas veces como sea necesario se obtiene la muestra de números aleatorios de una $X \sim U([0,1])$.

5.2.1. Cantidad de dígitos binarios del desarrollo

En el proceso descripto anteriormente no se ha analizado cómo ha de elegirse el número m donde cortar la serie para aproximar:

$$d = \sum_{i=1}^{+\infty} d_i \cdot 2^{-i} \approx \sum_{i=1}^{m} d_i \cdot 2^{-i} = D_m$$

Indudablemente un criterio razonable sería asegurar que el número hallado D_m sea indistinguible de D_{m+k} en nuestro sistema de representación numérico. Por ejemplo si utilizamos un sistema de representación de punto flotante con una mantisa de 32 bits, bastaría indudablemente con $m = 32$, pues aquellos dígitos por encima del número 32 serían pasados inadvertidos.

Por ejemplo imaginemos que el sistema de representación numérico es de punto flotante con una mantisa de 5 dígitos decimales y consideremos al $1{,}2345 \times 10^{-1}$. Si quisiéramos representar un número entre $1{,}2345 \times 10^{-1}$ y $1{,}2346 \times 10^{-1}$, como ser el $1{,}23457 \times 10^{-1}$, entonces el sistema redondearía y mostraría el número más cercano a éste último representable en el sistema, que es el $1{,}2346 \times 10^{-1}$. ¿Pero cuál es la distancia entre $1{,}2345 \times 10^{-1}$ y $1{,}23457 \times 10^{-1}$? La misma es:

$$\left|1{,}23457 \times 10^{-1} - 1{,}2345 \times 10^{-1}\right| = 0{,}00007 = 7 \times 10^{-5}$$

En realidad lo que está pasando es que si $|D - D_m| \leq 10^{-5}$ entonces para el sistema de representación sería D_m indistinguible de D_{m+k}. Sería apropiado entonces definir:

$$m = \text{mín}\left\{n \in \mathbb{N} : \ |D - D_n| \leq 10^{-5}\right\}$$

donde en este caso 5 es la cantidad de dígitos decimales de la mantisa.

Ahora bien:

$$\begin{aligned} |D - D_m| = \sum_{i=m+1}^{+\infty} d_i \cdot 2^{-i} \leq \sum_{i=m+1}^{+\infty} 2^{-i} = \sum_{i=0}^{+\infty} \left(\frac{1}{2}\right)^{i+m+1} \\ = \left(\frac{1}{2}\right)^{m+1} \sum_{i=0}^{+\infty} \left(\frac{1}{2}\right)^{i} = \left(\frac{1}{2}\right)^{m+1} \cdot 2 = \frac{1}{2^m} \end{aligned}$$

Por lo tanto, la condición para elegir m, conocida la cantidad M de dígitos decimales de la mantisa es:

$$|D - D_m| \leq 10^{-M} \Leftrightarrow \frac{1}{2^m} \leq \frac{1}{10^M} \Leftrightarrow 10^M \leq 2^m \Leftrightarrow M \cdot \log(10) \leq m \cdot \log(2)$$

Teniendo presente que $\log(10) = 1$ entonces la condición es:

$$m \geq \frac{M}{\log(2)} \approx 3{,}322 \cdot M$$

por lo que bastaría tomar:

$$m = [3{,}322 \cdot M] + 1$$

5.3. Método de Montecarlo *(Inversión)*

Es un método para simular muestras de variables aleatorias continuas, cuando su distribución acumulada es estrictamente creciente. Como método es muy noble y ampliamente utilizado debido a su sencillez de implementación. A continuación se explicarán las bases de su funcionamiento.

Supongamos que se dispone de un método para generar muestras aleatorias de una variable $U \sim \mathcal{U}([0,1])$. Sea ahora una *v.a.* X con función de distribución acumulada F, es decir:

$$P(x \leq a) = F(a)$$

Si definimos la variable aleatoria:

$$Y = F^{-1}(U)$$

entonces Y sigue la distribución F. Es decir:

$$Y = F^{-1}(U) \overset{\mathrm{d}}{=} X$$

Para demostrar esta propiedad basta hacer:

$$P(Y \leq a) = P\left(F^{-1}(U) \leq a\right) = P(U \leq F(a)) = F(a) = P(X \leq a)$$

por lo que:

$$P(Y \leq a) = P(X \leq a) \Leftrightarrow \boxed{X \overset{\mathrm{d}}{=} Y}$$

5.3.1. Pasos para generar una muestra aleatoria

Sea X una variable aleatoria continua con función de distribución estrictamente creciente. Los pasos para generar una muestra aleatoria $(x_i)_{1 \leq i \leq n}$ son:

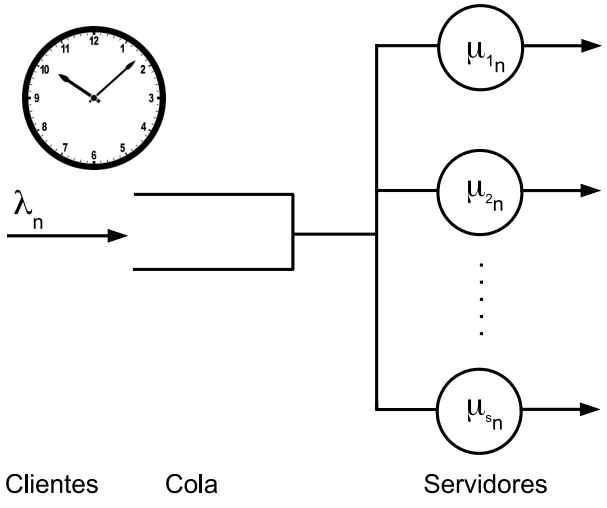
1. Generar muestras $\mu_1, \cdots, \mu_n$ de una variable $U \sim U([0,1])$.
2. Determinar analítica o numéricamente los valores $F^{-1}(\mu_i)$ para $i = 1, \cdots, n$.
3. La muestra aleatoria de X buscada será $x_1, \cdots, x_n$.

5.4. Simulación de un sistema de una cola

Los pasos para simular un sistema de una cola se basan en poder simular cada uno de sus componentes, simulando las variables aleatorias que rigen el funcionamiento de cada uno de ellos. A continuación se diseñará un algoritmo para simular un $M/M/s$.

5.4.1. Estructura del programa

El programa se diseñará orientado a objetos, y cada uno de los componentes de un sistema de una cola serán objetos con sus variables y métodos, que se unirán para formar un objeto mayor, que constituirá el sistema de una cola. El diagrama a seguir es el siguiente:



Hay un objeto que es el *reloj* o cronómetro del sistema, que contará la cantidad de instantes o ciclos de reloj transcurridos desde el inicio de la simulación. Debe haber también un objeto llamado *cola* que debe saber cuándo llega un cliente nuevo y además debe tener la capacidad de ir despachando clientes hacia el área de los servidores, en la medida que estos últimos lo soliciten. Las tasas de llegada de clientes serán los λ_n y dependen de la cantidad de clientes que haya en el sistema.

Cada uno de los s servidores será un objeto independiente, con su propia lista de μ_{i_n} para cada $1 \leq i \leq s$ y para cada $n \in \mathbb{N}$. La dependencia de n es nuevamente porque las tasas de atención dependen de la cantidad de clientes presentes en el sistema. Cada servidor debe tener la habilidad de saber cuánto tiempo le llevará atender al cliente que está procesando y deberá interactuar con la cola para solicitar el envío de un nuevo cliente.

Por último habrá un objeto llamado *sistema de una cola* que contendrá adentro un reloj, una cola y s servidores.

El programa principal será entonces muy pero muy sencillo, constando únicamente de tres partes principales:

1. Se crearán tantos objetos *sistema de una cola* como escenarios paralelos se pretendan simular simultáneamente.

2. Se inicializarán cada uno de los objetos.

3. Un bucle que llamará a los métodos iteradores de cada uno de los objetos mientras dure la simulación y controle además la interacción con el usuario.

5.4.2. Objetos que debe incluir el programa

5.4.2.1. Cronómetro del sistema

El cronómetro del sistema es un objeto que lleva la cuenta del tiempo transcurrido desde que empieza a correr la simulación. En cada paso del bucle del programa principal se incrementa una unidad de tiempo. Suelen utilizarse "*delays*" para ralentizar el proceso y que sea perceptible, ya que de lo contrario el sistema ingresaría en tiempo estacionario que sería imposible para el ojo humano visualizar la parte más interesante, que es todo lo que ocurre hasta llegar a la estacionariedad.

En este sentido, si elegimos *segundos* para la unidad de tiempo, podemos controlar el tiempo del proceso agregando al cronómetro un delay. Cada vez que el programa principal llame al cronómetro, el mismo deberá incrementar el tiempo de simulación a la vez que hace un *delay* de tantos milisegundos como se requiera. Por ejemplo un delay de 1000 milisegundos equivale a que la simulación se haga en *tiempo real*. Si hacemos un delay de 100 milisegundos, por cada segundo real transcurrirían 10 segundos en la simulación. Si se hace un delay de 10 milisegundos, cada segundo real equivaldrían a 100 segundos.

Los métodos del cronómetro deben incluir mínimamente:

Inicialización: Método que inicializa al cronómetro para que empiece a funcionar. Simplemente pone la variable tiempo de simulación a 0. También se debe llamar al inicializador de la estadística.

Tiempo de simulación: Es el tiempo acumulado desde que empezó la simulación. Se inicializa a 0. Se actualiza cada vez que se instancia el método iterador.

Método iterador: Avanza hasta el próximo instante de tiempo y llama al método display para que actualice la vista. También debe llamar al método que actualiza la estadística.

Solicitud de tiempo: Devuelve el tiempo total transcurrido de simulación.

Método display: Se encarga de hacer lo necesario para mostrar por pantalla o por el medio de entrada/salida que corresponda los datos del cronómetro.

Estadística inicializar: Es un método que inicializa el motor de estadística que llevará la cuenta sobre el relevamiento estadístico de datos a medida que transcurre la simulación.

Estadística actualizar: Es un método que se encarga de mantener actualizada la información estadística.

Estadística obtener: Es un método que devuelve la información estadística relevada hasta el momento para que el usuario pueda valerse de ella para analizar la simulación.

5.4.2.2. La cola

Es un objeto que debe incluir una serie de variables y métodos para setear las mismas. Como en general las "*colas*" son objetos pertenecientes a un objeto más grande llamado "*sistema de colas*", debe tener acceso a ciertas variables globales del sistema, como ser $N = \#$*clientes en el sistema*, etc.

N_q: Es la cantidad de clientes en la cola, inicializada a 0.

Tasas de tiempo: Son las λ_n para $n \in \mathbb{N}_0$, que indican las tasas de llegada de clientes, a razón de λ_n clientes por unidad de tiempo, cuando en el sistema se hallan n clientes. Esto significa que el tiempo transcurrido hasta la llegada del próximo cliente, cuando en el sistema hay n clientes, es una variable aleatoria $T_n \sim \varepsilon(\lambda_n)$.

Inicialización: Inicializa todas las variables y también llama al inicializador de la estadística.

Cantidad clientes: Es un método que devuelve la cantidad de clientes que hay en la cola a quien lo solicite.

Tiempo llegada último cliente: Es el tiempo que indicaba el cronómetro del sistema al llegar el último cliente. Cada vez que llegue un cliente esta variable debe ser actualizada.

Tiempo próximo cliente: Es el tiempo que hay que esperar hasta que llegue el próximo cliente. Se inicializa a una instanciación de $T_0 \sim \varepsilon(\lambda_0)$. Al llegar un cliente al sistema se debe configurar esta variable a una instanciación de $T_n \sim \varepsilon(\lambda_n)$, donde n es la cantidad de clientes presentes en el sistema contando al que acaba de llegar.

Método iterador: Cada vez que se ejecute lo que hace es chequear si ya transcurrió el tiempo necesario para que llegue el próximo cliente, en cuyo caso llama al método que se encarga de recibir a un cliente, para que se actualicen las variables correspondientes. También debe llamar al método display para que actualice la vista. Se debe llamar además al método que actualiza la estadística.

Despachar cliente: Es un método al que puede llamar cada uno de los servidores del sistema, para solicitar que le envíen un cliente. Lo que debe hacer la cola es fijarse si hay clientes para enviar o no. En caso de haberlos enviar uno, caso contrario informa que no es posible. Si hay clientes para enviar, deben actualizarse las variables correspondientes.

Recibir cliente: Es un método que suele llamar el método iterador cuando detecta que ya transcurrió el tiempo hasta la llegada del próximo cliente. Deben actualizarse N_q y el resto de las variables, como ser el tiempo de llegada del último cliente, el tiempo hasta que llegue el próximo cliente, etc...

Método display: Se encarga de hacer lo necesario para mostrar por pantalla o por el medio de entrada/salida que corresponda los datos del cronómetro.

Estadística inicializar: Es un método que inicializa el motor de estadística que llevará la cuenta sobre el relevamiento estadístico de datos a medida que transcurre la simulación.

Estadística actualizar: Es un método que se encarga de mantener actualizada la información estadística.

Estadística obtener: Es un método que devuelve la información estadística relevada hasta el momento para que el usuario pueda valerse de ella para analizar la simulación.

El objeto de la cola debe respetar la disciplina de la misma y eso debe notarse al llevar adelante la simulación. Por ejemplo si se elige una interfaz gráfica debería poder observarse que el primer cliente que llega es el primero que pasa a ser servido, ya que la disciplina es *FIFO*.

Es interesante incluir métodos que permitan "*tocar*" las variables del objeto al margen del transcurso de la simulación. Por ejemplo si le indicamos a la cola que configure $N_q = 100$ aunque en realidad $N_q = 10$, estaríamos experimentando qué ocurriría si de repente aparecieran 90 clientes de un tirón. También se puede experimentar cambiando las tasas λ_n. Los métodos para permitir estos cambios deben validar los datos ingresados. Por ejemplo no se debería permitir forzar un cambio de N_q al valor -20 pues es un valor que no tiene sentido y descontrolará con seguridad la simulación.

5.4.2.3. Cada uno de los s servidores

Los servidores son objetos con dos estados posibles: *libre* u *ocupado*. Los métodos y variables que habrían que definir son:

Tasas de tiempo: Son las μ_n para $n \in \mathbb{N}_1$, que indican las tasas de atención de clientes por el servidor, a razón de μ_n clientes por unidad de tiempo, cuando en el sistema se hallan n clientes. Esto significa que el tiempo transcurrido hasta que se logra procesar al cliente, cuando en el sistema hay n clientes, es una variable aleatoria $T_n \sim \varepsilon(\mu_n)$.

Estado: Es una variable que indica si el servidor está libre u ocupado. Se inicializa a 0. Se cambia a 1 cada vez que ingresa un cliente. Al despachar un cliente se debe solicitar otro a la cola. Si la cola puede enviar otro cliente, el servidor debe mantener en 1 su estado. Si la cola informa que no hay nadie más para enviar, el estado cambia a 0.

Inicializacion: Inicializa todas las variables y llama al inicializador de la estadística.

Solicitud estado: Es un método que informa del estado del servidor. Devuelve 1 si ocupado o 0 si libre.

Tiempo llegada último cliente: El el horario que indicaba el cronómetro del sistema al momento de recibir al cliente que se está atendiendo.

Tiempo de atención último cliente: Es el tiempo que le llevará al servidor atender al cliente que está procesando, contado desde el momento en que llega para ser atendido.

Solicitar cliente: Suponemos al estado *libre*. En tal caso el método debe llamar a la cola para solicitar le entregue un cliente. Si la cola informa que no pudo entregarlo debido a que no hay nadie en ella, se deja el estado en *libre*. Si la cola envía un cliente, se debe setear el tiempo de llegada del último cliente al estado que indique el cronómetro del sistema. Además se debe disponer de la cantidad de clientes en el sistema para disparar una instanciación de $T_n \sim \varepsilon(\mu_n)$. Con el resultado de la misma se debe setear el tiempo de atención del último cliente, para que se sepa cuánto demorará en ser atendido.

Despachar cliente: Suponemos al estado *ocupado*. En tal caso se configura el estado en *libre* y se llama al método *solicitar cliente*.

Método iterador: Cada vez que se ejecuta este método, si el estado es *libre*, debe llamar al método *solicitar cliente*. Si el sistema está ocupado se debe chequear si el cronómetro del sistema indica que ha pasado el tiempo necesario para procesar al cliente que está siendo atendido. Si el tiempo ha transcurrido se debe llamar al método *despachar cliente*. Se debe además llamar al método display para actualizar la vista del sistema. También se debe llamar al método que actualiza la estadística de la cola.

Método display: Se encarga de hacer lo necesario para mostrar por pantalla o por el medio de entrada/salida que corresponda los datos del cronómetro.

Al igual que en la cola, es interesante incluir métodos que permitan modificar dinámicamente los valores de los μ_n, siempre con el cuidado de validar la información ingresada por el usuario para impedir que se ingresen valores que puedan estropear la simulación.

Estadística inicializar: Es un método que inicializa el motor de estadística que llevará la cuenta sobre el relevamiento estadístico de datos a medida que transcurre la simulación.

Estadística actualizar: Es un método que se encarga de mantener actualizada la información estadística.

Estadística obtener: Es un método que devuelve la información estadística relevada hasta el momento para que el usuario pueda valerse de ella para analizar la simulación.

5.4.2.4. Sistema de una cola

Es un objeto que contiene un objeto de tipo *cronómetro*, un objeto de tipo *cola* y a s objetos de tipo *servidor*. Los métodos que habría que definir son:

Inicialización: Inicializa el cronómetro, la cola y los servidores según los parámetros iniciales.

N**:** Cantidad de clientes en el sistema. Normalmente se inicializa a 0.

Método iterador: Debe llamar al método iterador del cronómetro, la cola y cada uno de los servidores. Luego debe actualizar el valor de N según:

$$N = N_q + \sum_{i=1}^{s} E_i$$

donde E_i es el estado del $i-$ésimo servidor para $1 \leq i \leq s$. En general si $N_q \geq 1$ entonces seguro todos los servidores están ocupados y por lo tanto en este caso diréctamente se puede hacer $N = N_q + s$. Si por el contrario $N_q = 0$ entonces no se puede evitar calcular la sumatoria. También debe llamar al método display para que actualice la vista al usuario y al método actualizador de la estadística.

Método display: Se encarga de hacer lo necesario para mostrar por pantalla o por el medio de entrada/salida que corresponda los datos del cronómetro. Se debe llamar a los métodos display de cada uno de los objetos por separado.

Obtener estado: Es un método que devuelve el estado actual del sistema. Eso debe incluir a la cantidad de clientes en el sistema, en la cola, el estado de los servidores con sus tasas de atención, etc...

Setear estado: Es un método que permite setear todos los parámetros de los diversos objetos en el sistema de una cola, a otros valores definidos por el usuario. Esto permitiría por ejemplo cambiar la cantidad de personas en la cola para ver qué ocurre. O modificar los λ_i o los μ_i, etc...

Estadística inicializar: Es un método que inicializa el motor de estadística que llevará la cuenta sobre el relevamiento estadístico de datos a medida que transcurre la simulación. Se debe llamar al método inicializador de estadística de cada uno de los objetos que integran el sistema de una cola.

Estadística actualizar: Es un método que se encarga de mantener actualizada la información estadística. Se debe llamar al actualizador de la estadística de cada uno de los objetos que integran el sistema de una cola.

Estadística obtener: Es un método que devuelve la información estadística relevada hasta el momento para que el usuario pueda valerse de ella para analizar la simulación. Generalmente se llama a cada uno de los métodos *estadística obtener* de los objetos que integran el sistema de la cola y se devuelve la información conjunta.

5.4.3. Estadística en tiempo de simulación

A continuación se explicará una manera de generar estadística sobre las variables del sistema, que servirá como ejemplo para desarrollar los métodos encargados de llevar adelante la estadística que haga falta.

Supongamos que se desea establecer los p_n de manera empírica, llegando a su distribución de probabilidades como función de los datos obtenidos en la simulación, siendo N la variable aleatoria que cuenta el número de clientes en el sistema, por lo que en realidad N depende del tiempo t de simulación. Lo primero que se debe hacer es un vector como el siguiente:

N	0	1	2	$\cdots$	n	$\cdots$
$N[i]$	0	0	0	$\cdots$	0	$\cdots$

La tabla representa un vector llamado N con tantas posiciones como clientes pueda haber en el sistema. Como eso se desconoce a priori, se deberán arbitrar los medios para que el tamaño de N pueda crecer en la medida que se requiera. El valor $N[i]$ representa cuántas veces hubo en el sistema la cantidad i de clientes. Cada vez que el cronómetro del sistema avanza una unidad de tiempo habrá que fijarse la cantidad de clientes que hay en el sistema. Supongamos que fuera i, entonces habrá que incrementar una unidad el $i-$ésimo elemento del vector N, es decir habrá que hacer $N[i] := N[i]+1$. De esta forma, en la $i-$ésima posición del vector N se guarda permanentemente la cantidad de veces desde que se inició el cronómetro que hubo i clientes en el sistema.

Como el tiempo total de simulación registra cuántos instantes en total transcurrieron desde el inicio, entonces puede aproximarse:

$$p_n \approx \frac{N[n]}{t} \, \forall \, n \in \mathbb{N}_0$$

donde t es el tiempo de simulación que indica el cronómetro del sistema. En realidad lo que vale es:

$$p_n = \lim_{t \to +\infty} \frac{N[n]}{t} \, \forall \, n \in \mathbb{N}_0$$

por lo que habrá que esperar una considerable cantidad de instantes para que la aproximación sea fiable.

Puede aproximarse la distribución de probabilidades de cualquier otra magnitud observable en el sistema a partir del método descripto anteriormente.

5.4.4. El programa principal

Como se ha anticipado, el programa principal consta de un bucle infinito que cada vez que se ejecuta debe realizar una serie de acciones. Distinguiremos dos casos importantes.

5.4.4.1. Si se desea simular un único sistema de una cola

1. Al momento de inicialización crea un objeto de tipo "*Sistema de una cola*" y lo inicializa pasándole todos los parámetros necesarios.

2. El cuerpo principal del programa es un bucle que lo único que hace es llamar al método iterador del objeto creado en el paso anterior.

3. Se debe poder pausar la simulación para congelar un estado de la misma y reanudarla según los deseos del usuario.

4. Se deben incluir casilleros que muestren el estado de las variables principales de los diferentes objetos del sistema que puedan ser cambiadas por el usuario dinámicamente. Si el usuario realiza algún cambio se debe llamar al método "Setear estado" y pasarle los nuevos valores. La simulación debería continuar a partir del nuevo estado. Esto se hace para simular por ejemplo la llegada de un grupo inesperado de clientes, o un aumento o disminución de las tasas de llegada y/o atención. De esta forma el usuario puede pausar la simulación, cambiar el valor de N_q y observar qué efectos produce en el sistema. O bien duplicar los λ para observar el efecto producido al duplicar la tasa de llegada de clientes al sistema. O bien cambiar los μ de los servidores para ver qué efecto produce contratar a un cajero más eficiente, o viceversa.

5. Resetear la simulación.

6. Salir de la simulación.

5.4.4.2. Si se desean simular n sistemas de una cola en paralelo

1. Al momento de inicialización se deben crear n objetos de tipo "*Sistema de una cola*" e inicializarlos pasándoles a cada uno de ellos todos los parámetros necesarios.

2. El cuerpo principal del programa es un bucle que lo único que hace es llamar al método iterador de cada uno de los n objetos.

3. Los puntos que siguen son análogos al caso anterior.

5.5. Propuesta de Trabajo Práctico

1. Según las indicaciones de la sección anterior simular un sistema de una cola $M/M/s$, donde el usuario pueda elegir el valor de s. Utilizar un mismo λ con independencia de la cantidad de clientes en el sistema. Utilizar para cada uno de los servidores un μ fijo e independiente de la cantidad de clientes en el sistema.

2. Simular en paralelo s sistemas de una cola $M/M/1$ utilizando el mismo μ que en el punto anterior pero reemplazando λ por $\frac{\lambda}{s}$.

3. Simular un sistema de una cola $M/M/1$ con λ igual que en los puntos anteriores pero reemplazando μ por $s\mu$.

A continuación se indicarán las razones por las que se sugiere la propuesta de trabajo:

1. En el primer punto hay una única cola, pero varios servidores, de modo tal que el cliente no puede elegir en qué servidor ubicarse, sino que lo hace el sistema. Esto es muy parecido al sistema elegido por los bancos o ciertos supermercados, donde forman una única fila y los clientes van entrando a medida que los cajeros se van liberando.

2. Aquí se pretende simular un sistema convencional, donde hay s cajeros pero se van formando colas en cada uno de ellos y son los clientes quienes deciden en qué cajero ubicarse, según su propio criterio. En este caso a cada cola llega una tasa de $\frac{\lambda}{s}$ clientes. Pero como hay s colas, la tasa total de llegada es igual que en el caso anterior: $s \cdot \frac{\lambda}{s} = \lambda$.

3. Aquí si bien hay un sólo cajero la tasa de atención, al ser $s\mu$ es s veces más rápida que en los casos anteriores.

Como puede verse, en los tres casos la capacidad de atención del sistema es equivalente así como también la tasa neta de llegada de clientes. En los dos primeros es obvio porque ambos tienen s servidores y aunque en cada cola del segundo ingresan a razón de $\frac{\lambda}{s}$ clientes por unidad de tiempo, en la sumatoria de todas las s colas finalmente terminan llegando a razón λ clientes por unidad de tiempo. En el tercero, aunque hay un único servidor, tiene capacidad $s\mu$ por lo que equivale a s de los otros.

La idea es evaluar si alguna de las opciones aparece como más favorable, o si hay razones para suponer que un sistema es mejor que el otro. También podría ocurrir que los tres se comporten de manera equivalente y se estabilicen con una cantidad similar de clientes en el sistema y en la cola. Precisamente eso se pretende evaluar en la propuesta de trabajo.

También se debe llevar una cuenta de las siguientes variables, para cada uno de los casos:

- Las probabilidades $(p_n)_{n\in\mathbb{N}_0}$, calculadas en forma empírica.

- L, es decir la cantidad esperada de clientes en el sistema.

- L_q, es decir la cantidad esperada de clientes en la cola.

- W, o sea el tiempo promedio que un cliente elegido al azar permanece en el sistema.
- W_q, el tiempo promedio que un cliente elegido al azar permanece en la cola.

Todas las variables descriptas anteriormente deben ser calculadas por el estudiante en tiempo de ejecución y mantener las mismas actualizadas conforme avanza la simulación. No está permitido utilizar las relaciones que establecen sobre ellas las Fórmulas de Little. Al estabilizarse el sistema cuando ingresa y permanece un tiempo en estado estacionario, se debe estudiar y analizar la validez de las Fórmulas de Little sobre las mismas.