Simulación

Ejemplo de simulación de eventos discretos: línea de espera

Jorge de la Vega Góngora

Departamento de Estadística
Instituto Tecnológico Autónomo de México

Laboratorio práctico

Modelo de línea de espera

Ejemplo 1: Línea de espera I

- Consideren una unidad de servicio con un sólo servidor (cajero, taquilla, una pista de aeropuerto, etc.)
- Los clientes se forman para recibir el servicio

Estación Tren en Beijing en hora pico

Se desea saber cuál es el tiempo promedio de permanencia en la fila, medido como el tiempo que transcurre desde que el cliente llega a la fila hasta que comienza a recibir el servicio. Las variables que determinan el estado del sistema son:

Ejemplo 1: Línea de espera II

- El estado del servidor (libre u ocupado)
- El número de clientes esperando en la fila $\{0, 1, 2, \ldots\}$
- El tiempo entre llegadas de clientes a la fila (tiempos de interarribo, se suponen iid $T_1, T_2, \ldots$)
- El tiempo de servicio a cada cliente (tiempos de servicio iid $S_1, S_2, \ldots$, se suponen independientes también de los tiempos de interarribo).

En este sistema hay dos tipos de eventos relevantes, que cambian el estado del sistema:

1. Llegada de un cliente (puede modificar el estado del servidor o el número de clientes esperando en fila)
2. Completar el servicio o partida de un cliente (puede modificar el estado del servidor, o el número de clientes esperando en fila, la generación de un tiempo de servicio o de llegada).

Estación metro Pantitlán, a cualquier hora

Otros supuestos del modelo, que podrían modificarse:

- Se supone una política de primeras entradas, primeras salidas (FIFO).
- Se considera la atención a un número fijo de clientes en el sistema, es decir, la simulación termina cuando se han atendido n clientes. Entonces el tiempo de simulación es aleatorio.
- Hay clientes que tienen una espera en fila de 0, que son los que se atienden inmediatamente.

Recordemos las definiciones que hemos visto antes:

$$t_i = \text{ Tiempo de llegada del } i\text{-ésimo cliente.}$$

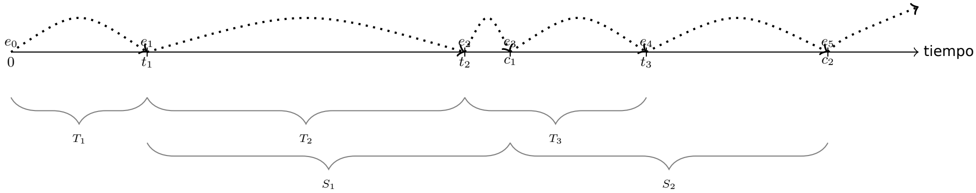
$$T_i = t_i - t_{i-1} = \text{ tiempo entre las llegadas del } i-1 \text{ e } i \text{ cliente (tiempo de interarribo).}$$

$$S_i = \text{ tiempo de servicio al cliente } i.$$

$$D_i = \text{ tiempo de espera en cola del cliente } i.$$

$$c_i = t_i + D_i + S_i = \text{ tiempo transcurrido entre la llegada del cliente } i \text{ y su salida del sistema.}$$

$$e_i = \text{ tiempo de ocurrencia del } i\text{-ésimo evento de cualquier tipo (reloj de simulación).}$$



Mediciones de interés I

- Además de estimar el tiempo promedio que un cliente pasa en la fila, puede haber otras cantidades que sea de interés medir, por ejemplo, desde el punto de vista del servidor. En este contexto es usual calcular tres mediciones:
  1. El *promedio esperado del tiempo de espera* en fila de los n clientes: $d(n)$. Si en una simulación particular se observan tiempos de espera $D_1, \ldots, D_n$, entonces tenemos *un* estimado de $d(n)$. Para calcular su valor esperado, necesitaremos obtener varias simulaciones.

     $$\widehat{d(n)} = \frac{\sum_{i=1}^{n} D_i}{n}$$

     Noten que estos tiempos no son independientes en una simulación en particular. Además, en una simulación obtenemos una muestra de tamaño 1 de $d(n)$.
  2. El *número promedio esperado de clientes en fila que no están siendo atendidos*, $q(n)$.

     Para tomar en cuenta ese tiempo aleatorio, consideremos $Q(t)$ como el número de clientes en la cola al tiempo t, $Q(t) \in \{0, 1, 2, \ldots\}$.

     Sea T_n el tiempo total que se requiere observar el sistema para observar las n variables $\{D_1, \ldots, D_n\}$. Entonces

     $$q(n) = \sum_{i=0}^{\infty} i \frac{T_i}{T(n)}$$

donde T_i es el tiempo del sistema que la cola tiene longitud i (notar que $T(n) = \sum_{i=0}^{\infty} T_i$). Entonces $\sum_{i=0}^{\infty} iT_i = \int_0^{T(n)} Q(t)\, dt$ y:

$$\widehat{q(n)} = \frac{\sum_{i=0}^{\infty} iT_i}{T(n)} = \frac{\int_0^{T(n)} Q(t)\, dt}{T(n)}$$

La primera parte de la fórmula nos ayudará a calcular el estimador $\widehat{q(n)}$ como la suma de áreas de rectángulos.

3. La *utilización esperada del servidor*, $u(n)$. Es la proporción del tiempo de simulación que el servidor estuvo ocupado. Si definimos $B(t) = I($servicio al tiempo $t)$, entonces:

$$\widehat{u(n)} = \frac{\int_0^{T(n)} B(t)\, dt}{T(n)}$$

- La simulación nos tiene que dar los estimadores $\widehat{d(n)}$, $\widehat{q(n)}$ y $\widehat{u(n)}$, y de ser posible, debería permitirnos visualizar el comportamiento de las funciones $Q(t)$ y $B(t)$.

- Supongamos que se observan los siguientes tiempos de interarribo y de servicio para $n = 6$ clientes:

$$T_1 = 0.4, T_2 = 1.2, T_3 = 0.5, T_4 = 1.7, T_5 = 0.2, T_6 = 1.6$$

$$S_1 = 2.0, S_2 = 0.7, S_3 = 0.2, S_4 = 1.1, S_5 = 3.7, S_6 = 0.6$$

¿Cómo podemos simular manualmente este sistema?

Ejemplo

Podemos organizar la información para entender mejor el flujo de los eventos. Completar la siguiente tabla:

Cliente	T_i	t_i	Inicio Servicio	Duración S_i	Término Servicio c_i
		0			
1	0.4	0.4		2.0	
2	1.2	1.6		0.7	
3	0.5	2.1		0.2	
4	1.7	3.8		1.1	
5	0.2	4.0		3.7	
6	1.6	5.6		0.6	

Componentes de un modelo de simulación de eventos discretos I

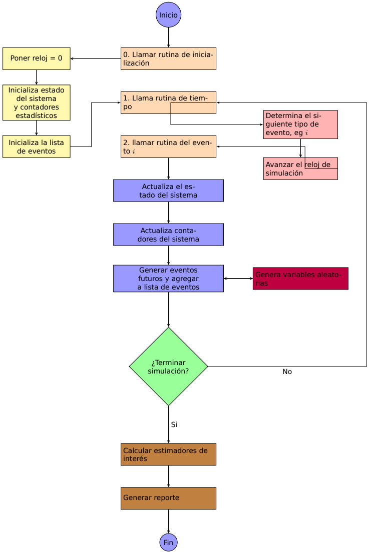
Los modelos de simulación discreta tienen componentes comunes, que facilitan su implementación computacional. Típicamente constan de:

- *Estado del sistema*: Es la colección de variables de estado necesarias para describir el sistema en particular.
- *Reloj de simulación*: la variable e que lleva el tiempo simulado.
- *Lista de eventos*: Una lista con los tiempos en los que ocurrirá cada tipo de evento.
- *Contadores estadísticos*: Variables que almacenan la información sobre el desempeño del sistema.
- *Rutina de inicialización*: Un subprograma que inicializa el modelo de simulación (sus variables, listas, etc.) en el tiempo 0.
- *Rutina de tiempo*: Un subprograma que determina el siguiente evento de la lista de eventos y avanza el reloj de simulación al tiempo en que ocurre ese evento.
- *Rutina de evento*: Un subprograma que actualiza el estado del sistema cuando ocurre cierto tipo de evento (puede haber una rutina para cada tipo de evento).
- *Rutina de bibliotecas*: Un conjunto de subprogramas usadas para generar observaciones aleatorias con las distribuciones especificadas en el modelo de simulación.

- *Generador de reportes*: Un subprograma que calcula las medidas de desempeño elegidas para reportar y produce un reporte cuando termina la simulación.
- *Programa principal*: Un subprograma `mm1` que llama a los diferentes subprogramas conforme se requiere (típicamente llama primero a la rutina de inicialización y posteriormente de manera repetitiva a la rutina de tiempo y a la rutina de evento).

El código de colores corresponde a las actividades que se muestran en el siguiente diagrama de flujo, que conecta las relaciones lógicas entre los componentes del modelo.

Flujo del proceso



Supuestos del modelo de simulación I

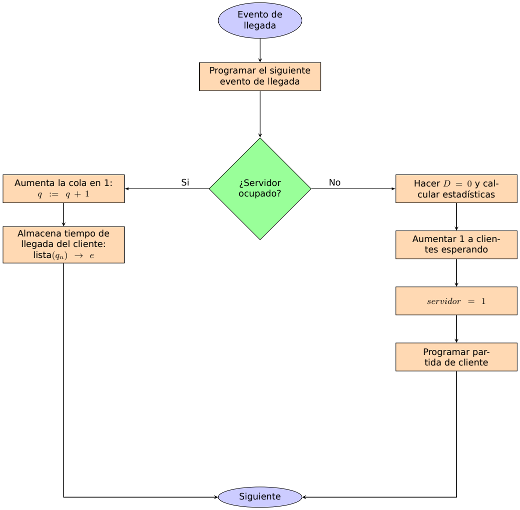
A continuación se resumen los supuestos que se utilizarán para el modelo $M/M/1$:

- El sistema se ejecuta hasta que n clientes esperan en fila: la simulación termina cuando el cliente n sale de servicio.
- Los tiempos de interarribo se modelarán como varibles aleatorias iid exp (λ_A). La distribución exponencial con media λ_A tiene densidad:

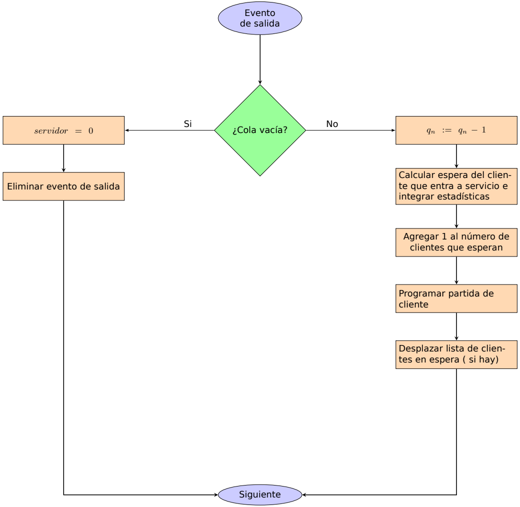
$$f(x) = \frac{1}{\lambda_A} e^{-x/\lambda_A} I(x \geq 0)$$

- Los tiempos de servicio se modelarán como variables aleatorias iid exp (λ_S).

Lógica del evento 1: Llegada de un cliente al sistema I



Lógica del evento de salida I



Funciones y Variables relevantes I

Parámetros del modelo:

- lambdaA (λ_A) y lambdaS (λ_S) son los tiempos promedio de interarribo y de servicio, respectivamente.
- n = número de clientes que entraron al sistema. Fijo de antemano.

Variables principales:

- reloj = e = reloj de simulación
- sig_tipo_evento = siguiente tipo de evento (1= llegada, 2= salida)
- tiempo_sig_evento = vector de dimensión 2, con el tiempo de llegada y el tiempo de salida de eventos conforme se registran
- Di = D_i tiempo de espera en cola de un cliente
- total_esperas = $T(n)$ = La suma de los tiempos de espera D_i de clientes en cola conforme se van generando
- q_t = longitud de la cola en un tiempo dado

Variables auxiliares:

- lista_A = lista$_A$ vector lista de tiempos de arribo (longitud dinámica)

Funciones y Variables relevantes II

- `le` = lista de eventos ocurridos, con su tiempo de ocurrencia y su tipo y la longitud de la fila en esos tiempos.
- `area_q` = variable auxiliar para calcular la longitud promedio de la fila.
- `area_status_servidor` = variable auxiliar para calcular la ocupación del servidor.
- `tiempo_ultimo_evento` = registro del último evento antes de avanzar el reloj de simulación.
- `tiempo_desde_ultimo_evento` = calcula el intervalo de tiempo entre el actual y el tiempo del último evento.
- `clientes_enespera` = Número de clientes que se han formado en la fila.

Ejecución del programa I

La siguiente simulación nos da una muestra de tamaño 1 del proceso $M/M/1$. Para sacar conclusiones definitivas, deberíamos ejecutar el modelo de simulación varias veces para obtener una muestra adecuada de las cantidades estimadas.
El programa `Queue.Rmd` lo pueden obtener de la siguiente liga de Piazza

Ejecución del programa II

```
source("../scripts/Queue/Queue.R")
(a <- mm1(lambdaA = 1, lambdaS = 0.5, n = 1000))

$promedio.espera
[1] 0.5564974

$tiempo_total_promedio_en_sistema
[1] 0.4906873

$longitud_max
[1] 8

$espera_max
[1] 556.4974

$longitud.promedio.fila
[1] 0.5501804

$utilizacion
[1] 0.490648

$tiempo.simulacion
[1] 1011.482

$le
              e tipo q
le    0.0000000    0 0
      0.3160743    1 0
      0.8321808    2 0
      1.0643127    1 0
      1.5424283    2 0
      1.7142070    1 0
      1.7397158    2 0
      3.4816925    1 0
      4.1700201    1 0
      4.4091980    2 1
```

Comparación con la teoría de los modelos $M/M/1$

De acuerdo a la teoría de los modelos de linea de espera, considerando tiempos de arribo y tiempos de servicio con medias λ_A y λ_S respectivamente, las cantidades que estimamos en el ejemplo anterior tienen los siguientes valores esperados teóricos:

- La utilización esperada del servidor es $u = \frac{\lambda_S}{\lambda_A}$.
- El número promedio de clientes en la cola: $q_n = \frac{\lambda_S^2}{\lambda_A(\lambda_A - \lambda_S)}$
- El tiempo promedio de espera en cola es: $D_n = \frac{\lambda_S^2}{\lambda_A - \lambda_S}$

Varias simulaciones

```r
a <- numeric(100)
for(i in 1:100){
        a[i] <- mm1(lambdaA = 1, lambdaS = 0.5, n = 1000)$promedio.espera
}
mean(a) # estimador del promedio de espera basado en 100 muestras
```

```
[1] 0.5060275
```

```r
sd(a)
```

```
[1] 0.09292669
```