Probabilidad y Estadística LCC
Trabajo Práctico Final 2020

1. Considere el proceso I_n donde cada realización corresponde al experimento de tirar una moneda al aire y registrar si aparece cara.

   a. Simule 100 pasos del proceso I_n considerando una moneda “equilibrada” y registre el número de caras al momento n (X_n).

   b. Use su simulación para estimar la P(X=1) y la E(X).

   c. ¿Cuántas veces deberá realizar el experimento a fin de que cara aparezca por tercera vez?

   d. Modifique su código a fin de permitir que la moneda sea “sesgada” y responda el item b.

2. Considere el proceso $D_n = 2I_n - 1$, donde D_n representa el cambio de posición de una partícula que se mueve a lo largo de una linea recta con saltos de magnitud 1 en cada momento, ya que depende del proceso I_n definido en el punto 1.

   a. Simule 10 pasos de una trayectoria de D_n

   b. Simule una realización del proceso S_n que denota la posición de la partícula en el momento n y describa dicha trayectoria.

3. Considere a un jugador que tiene una apuesta inicial de k dólares, y repetidamente apuesta 1 dólar en un juego en el cual la probabilidad de ganar es p y la probabilidad de perder es 1-p. La fortuna sucesiva del jugador es una simple caminata aleatoria iniciada en k. Supongamos que el jugador decide detenerse cuando su fortuna alcanza S dólares (S > k), o cae a 0 dólar, lo que ocurra primero.

   a. Utilice el lenguaje R, a fin de simular y visualizar la evolución del capital del jugador.

   b. Grafique una trayectoria del proceso.

   c. Estime, mediante la simulación de un número adecuado de trayectorias del capital del jugador, la probabilidad de ruina de dicho jugador. Elija un estado inicial $X_0 = k$, un valor para S y una probabilidad “p” de que el jugador gane en cada etapa del juego (elija un valor de p > 0.5).

Nota: Este es el clásico problema de la ruina del jugador, discutido por primera vez por los matemáticos Blaise Pascal y Pierre Fermat en 1656 y el cual sirve para describir muchos fenómenos similares en la actualidad.

4. Un ratón de laboratorio se encuentra atrapado en un laberinto y, al inicio, tiene que elegir entre dos posibles direcciones. Si elige ir a la derecha, entonces se paseará en el laberinto durante los siguientes tres minutos y luego volverá a su posición inicial. Si elige ir a la izquierda, entonces con probabilidad 1/3 saldrá del laberinto después de dos minutos, y con probabilidad 2/3 volverá a su posición inicial al cabo de cinco minutos. Asumiendo que el ratón, en todo momento, elige con la misma probabilidad ir a la izquierda o a la derecha,

   a. Simule el problema para estimar el número promedio de minutos que el ratón aparecerá atrapado en el laberinto antes de conseguir la libertad.

   b. Explicite en forma analítica cómo puede obtener ese resultado sin recurrir al proceso de simulación.

5. Desde comienzos de los años 80, el SIDA se ha convertido en una de las mayores pandemias de nuestra época (hasta la aparición del Covid19!). En la actualidad, España es uno de los países más afectados por el virus, por lo que el conocimiento de la evolución futura de la epidemia es un objetivo prioritario para establecer políticas sanitarias adecuadas. Uno de los modelos más sencillos para el estudio de la epidemia de SIDA establece que cada sujeto de la población puede estar exclusivamente en uno de los siguientes estados: susceptible (S), infectado por VIH (VIH), con SIDA (SIDA) o muerto como consecuencia de la enfermedad (M). La Figura 1 muestra las posibles transiciones entre estados.

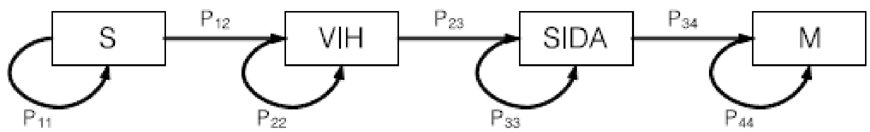

Figura 1. Estados en la epidemia de VIH-SIDA

$$P = \begin{pmatrix} 0{,}999 & 0{,}001 & 0 & 0 \\ 0 & 0{,}994 & 0{,}006 & 0 \\ 0 & 0 & 0{,}918 & 0{,}082 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

La observación de los sujetos se realiza de forma anual, por lo que tanto los estados como los instantes de tiempo del proceso son discretos. Además, el estado futuro en que se encuentre un individuo dependerá solo de su estado actual, pero no de su historia pasada. Por ello, el proceso estocástico es una cadena de Markov discreta.

a. Estudie las cadenas y sus estados.

b. Calcule para cada i, j pertenecientes a E la probabilidad de alcanzar el estado j a partir de i en algún número finito de pasos. Interprete los resultados en términos del problema.

c. Calcule el número esperado de pasos que la cadena está en los estados transitorios antes de ser absorbida. Interprete en términos del problema.

d. Analice si existe para esta cadena alguna distribución invariante.

e. Analice si existe distribución límite para esta cadena. En caso de existir, hállela.

6. El poder de los motores de búsqueda de Internet reside en su capacidad para responder a la consulta de un usuario con una lista ordenada de sitios web clasificados por importancia y relevancia. El corazón del motor de búsqueda de Google es la familia de algoritmos PageRank, que asignan un valor de importancia a cada página web, llamado rango de página. PageRank confía en la naturaleza democrática de la web utilizando su vasta estructura de enlaces como un indicador del valor de una página en concreto. Google interpreta un enlace de una página *A* a una página *B* como un voto, de la página *A*, para la página *B*. No solo se tiene en cuenta el volumen de votos, o enlaces que una página recibe; también se analiza la página que emite el voto. Los votos emitidos por las páginas consideradas "importantes", es decir con un PageRank elevado, valen más, y ayudan a hacer a otras páginas "importantes". Por lo tanto, el PageRank de una página refleja la importancia de la misma en Internet. Aunque el algoritmo real de PageRank es complejo y tiene muchos aspectos técnicos (además de detalles secretos para intentar impedir que alguien manipule el rango

de su página web), el rango de página de una página web particular se describe fácilmente por medio de un modelo estocástico.

Considere un usuario web hipotético que navega a través de Internet moviéndose de una página a otra al azar. Cuando el usuario está en una página web en particular, elige uno de los enlaces de hipertexto disponibles en esa página de manera uniforme al azar y luego pasa a esa página. El modelo puede describirse como una caminata aleatoria por parte del internauta en un gráfico gigante llamado el webgraph. En el gráfico web, los vértices (nodos) son páginas web. El vértice "x" se une al vértice "y" por un arco dirigido si hay un enlace de hipertexto en la página "x" que conduce a la página "y". Cuando la persona que está navegando se encuentra en el vértice "x", elige al azar un arco del conjunto de arcos disponibles. Imagine que el internauta ha estado caminando aleatoriamente por la web durante mucho tiempo, y se calcula la probabilidad de que esté en la página "x". Esta probabilidad a largo plazo es precisamente el rango de página de la página x. Intuitivamente, la probabilidad a largo plazo de estar en una página en particular tenderá a ser mayor para las páginas con más enlaces entrantes y más pequeños para páginas con pocos enlaces, y es una medida de la importancia o popularidad de una página. El algoritmo de PageRank se puede entender como una asignación de probabilidades a cada sitio en la web. La Figura 2 muestra una red simplificada de siete páginas.

a. Modele el comportamiento de visitas a las páginas como una cadena de Markov.

b. Simule 100 pasos de este proceso.

c. Determine el rango de cada página.

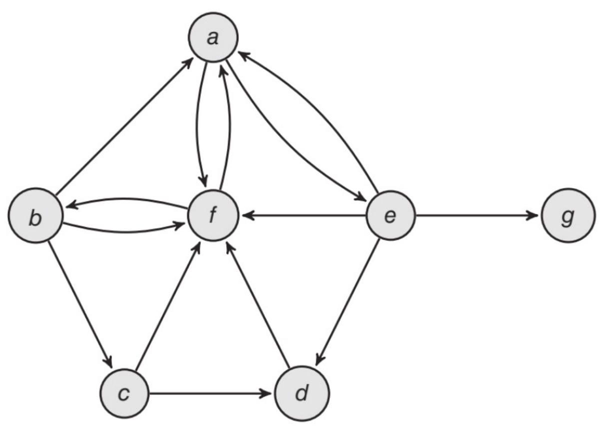

Figura 2. Red simplificada. No corresponde al grafo de la cadena.

7. La Figura 3 muestra una trayectoria del proceso de conteo en el cual los eventos ocurren en los tiempos t_1, t_2, t_3, t_4, t_5. Suponga que el evento bajo estudio es el número de mensajes de WhatsApp que recibe una persona a una tasa λ (mensajes/hora).

a. Juan recibe mensajes comenzando a las 10 am a una tasa de 10 textos por hora, de acuerdo a un Proceso Poisson. Encuentre la probabilidad de que reciba exactamente 18 mensajes hasta las 12 am y 70 mensajes hasta las 5 pm

b. Simule una trayectoria muestral durante un intervalo de tiempo suficientemente largo [0,T], Calcule los tiempos entre llegadas. Represente gráficamente los tiempos observados entre llegadas a través de un histograma y compare el resultado obtenido con la función de densidad teórica del tiempo entre llegadas determinada por el modelo elegido.

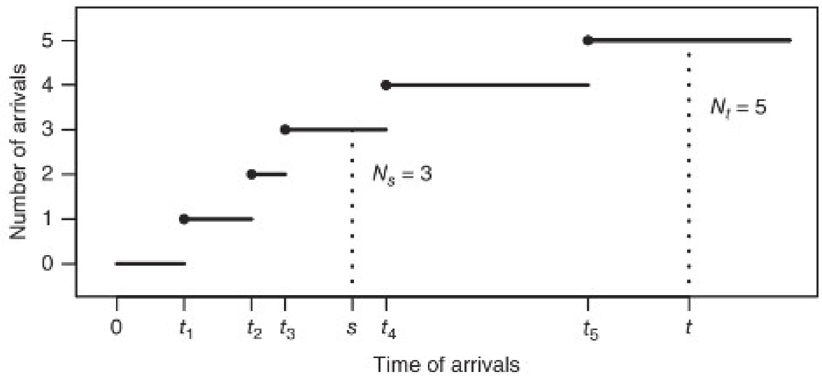

Figura 3. Trayectoria de proceso de conteo.

8. En aplicaciones de seguridad informática, un *honeypot* (o sistema trampa) es una herramienta dispuesta en una red o sistema informático para ser el objetivo de un posible ataque informático, y así poder detectarlo y obtener información del mismo y del atacante. Los datos del *honeypot* son estudiados utilizando cadenas de markov. Se obtienen datos desde una base de datos central y se observan ataques contra cuatro puertos de computadoras - 80, 135, 139 y 445- durante un año. Los estados de la cadena de Markov son los cuatro puertos y se incluye un nodo indicando que ningún puerto está siendo atacado. Los datos de monitorean semanalmente y el puerto más atacado durante la semana es guardado. La matriz de transición para la cadena estimada para los ataques semanales es:

$$
\begin{array}{r c}
 & \begin{array}{ccccc} \quad 80 & \quad 135 & \quad 139 & \quad 445 & \text{No attack} \end{array} \\
\boldsymbol{P} = \begin{array}{r} 80 \\ 135 \\ 139 \\ 445 \\ \text{No} \end{array} & \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 8/13 & 3/13 & 1/13 & 1/13 \\ 1/16 & 3/16 & 3/8 & 1/4 & 1/8 \\ 0 & 1/11 & 4/11 & 5/11 & 1/11 \\ 0 & 1/8 & 1/2 & 1/8 & 1/4 \end{pmatrix},
\end{array}
$$

con distribución inicial π_0 = (0,0,0,0,1).

a. Después de dos semanas, ¿cuáles son los puertos con más y menos probabilidad de ser atacados?

b. Encuentre la distribución límite (si es que existe) de los puertos atacados. Justifique.