Escuela de Ciencias Exactas y Naturales

Licenciatura en Ciencias de la Computación

Trabajo práctico Final

Manuel Spreutels
Augusto Rabbia
Erik Gimenez

Probabilidad y Estadística

Febrero 2024

Trabajo Práctico Final
Probabilidad y Estadística

Problema 1

Se sabe que ciertos errores aleatorios de medición $X_1, X_2, ..., X_n$, que contribuyen en forma independiente al error total, se distribuyen exponencialmente con tasa $\lambda = 0{,}5$. Encuentre en forma empírica la función de densidad de probabilidad de tres variables aleatorias

$$A = X_1 + X_2 + ... + X_5$$
$$B = X_1 + X_2 + ... + X_{20}$$
$$C = X_1 + X_2 + ... + X_{100}$$

Caracterice gráfica y numéricamente (media aritmética y desvío estándar) los resultados obtenidos. Comente.

Solución. Se nos presentan tres variables aleatorias definidas como la suma de variables aleatorias continuas independientes, cada una con distribución exponencial con tasa $\lambda = 0{,}5$.

Para lograr encontrar una aproximación de sus funciones de densidad de forma empírica, se simula el experimento 100000 (cien mil) veces.

```r
# Cantidad de simulaciones
n_simulations <- 100000

n_A <- 5
n_B <- 20
n_C <- 100

lambda <- 0.5 # Tasa lambda

# Realizar simulaci n para cada variable aleatoria
set.seed(123)  # Fijar semilla para reproducibilidad
simulations_A <- matrix(rexp(n_simulations * n_A, rate = lambda), ncol = n_A)
simulations_B <- matrix(rexp(n_simulations * n_B, rate = lambda), ncol = n_B)
simulations_C <- matrix(rexp(n_simulations * n_C, rate = lambda), ncol = n_C)

# Calcular sumas para cada simulacion
sums_A <- rowSums(simulations_A)
sums_B <- rowSums(simulations_B)
sums_C <- rowSums(simulations_C)

# Estad sticas descriptivas
mean_A <- mean(sums_A)
sd_A <- sd(sums_A)

mean_B <- mean(sums_B)
sd_B <- sd(sums_B)

mean_C <- mean(sums_C)
sd_C <- sd(sums_C)
```

Luego, imprimendo por pantalla los resultados se obtiene:

Variable A:

Media: 10.00416

Desvío Estándar: 4.448108

Variable B:

Media: 39.97253

Desvío Estándar: 8.93313

Variable C:

Media: 200.0222

Desvío Estándar: 19.87303

Lo cual coincide con los resultados que podrían esperarse de una variable de estas características, que se pueden obtener utilizando las fórmulas correspondientes, sabiendo que:

- Si se tiene una v.a. definida como una suma de n v.a.s,

$$Z = \sum_{i=1}^{n} a_i X_i$$

entonces vale que

$$E(Z) = \sum_{i=1}^{n} a_i E(X_i)$$

y, si X_i independientes dos a dos para todo i,

$$V(Z) = \sum_{i=1}^{n} a_i^2 V(X_i)$$

- La esperanza matemática y varianza para la función exponencial son $E(X) = 1/\alpha$ y $V(X) = 1/\alpha^2$ respectivamente, con α la tasa de la distribución.

Luego, como la tasa es $\lambda = 0{,}5$, se tiene que:

$$E(X_i) = \frac{1}{0{,}5} = 2, \qquad \forall i = 1, 2, ..., n$$

y

$$V(X_i) = \frac{1}{0{,}5^2} = 4, \qquad \forall i = 1, 2, ..., n$$

Ahora, utilizando las ecuaciones obtenidas en el primer punto, con $a_i = 1, \forall i = 1, 2, ..., 100$:

$$E(A) = \sum_{i=1}^{5} E(X_i) = \sum_{i=1}^{5} 2 = 10 \qquad V(A) = \sum_{i=1}^{5} V(X_i) = \sum_{i=1}^{5} 4 = 20$$

$$E(B) = \sum_{i=1}^{20} E(X_i) = \sum_{i=1}^{20} 2 = 40 \qquad V(B) = \sum_{i=1}^{20} V(X_i) = \sum_{i=1}^{20} 4 = 80$$

$$E(C) = \sum_{i=1}^{100} E(X_i) = \sum_{i=1}^{100} 2 = 200 \qquad V(C) = \sum_{i=1}^{100} V(X_i) = \sum_{i=1}^{100} 4 = 400$$

Finalmente, obteniendo la raíz positiva de las respectivas varianzas, se haya el desvío estándar para cada variable

$$\sigma_A = \sqrt{V(A)} = 4{,}47$$

$$\sigma_B = \sqrt{V(B)} = 8,94$$

$$\sigma_C = \sqrt{V(C)} = 20$$

Como era de esperar, los valores en las simulaciones son muy similares a los calculados analíticamente.

Por otro lado, se desea ver la forma del gráfico obtenido a partir de estas simulaciones:

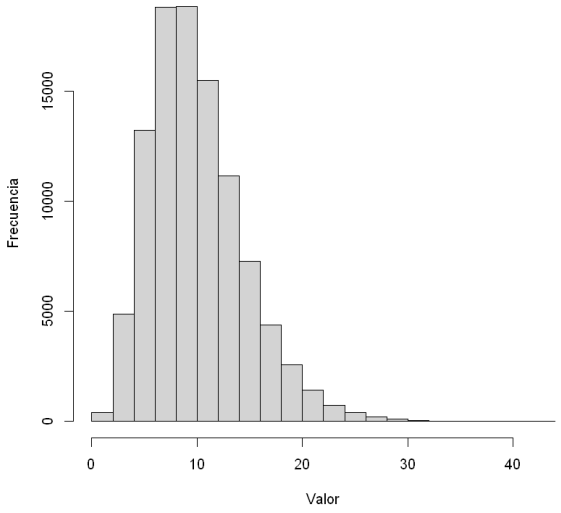
```
histogramas <- function(sums, n_sims, dist, E, SD) {
  title <- paste("Distribuci n de ", nom_dist)
  # Crear histogramas de las distribuciones
  hist = hist(sums, breaks = 25, main = title, xlab = "Valor", ylab = "
   Frecuencia")
  hist$counts <- hist$counts / n_sims
  title2 <- paste("Distribuci n Normalizada de ", nom_dist)
  # Crear graficos de las distribuciones normalizadas, y comparar con la
   distribucion normal
  plot(hist, main = title2, xlab = "Valor", ylab = "Probabilidad", ylim = c
   (0,1))
  binwidth <- hist$breaks[2] - hist$breaks[1]
  x_values <- seq(min(sums), max(sums), length.out = 100)
  lines(x_values, dnorm(x_values, mean=E, sd=SD) * binwidth, type="l", lwd
   =2, col="darkblue")
       # Anadimos leyendas de los colores correspondientes
  legend("topright", legend=c("Resultado de la simulaci n", "Curva normal")
   , col=c("grey", "darkblue"), pch=15)
}
histogramas(sums_A, n_simulations, "A", mean_A, sd_A)
histogramas(sums_B, n_simulations, "B", mean_B, sd_B)
histogramas(sums_C, n_simulations, "C", mean_C, sd_C)
```

Y a partir de este código, se derivan los resultados que se observan en la figura 1.

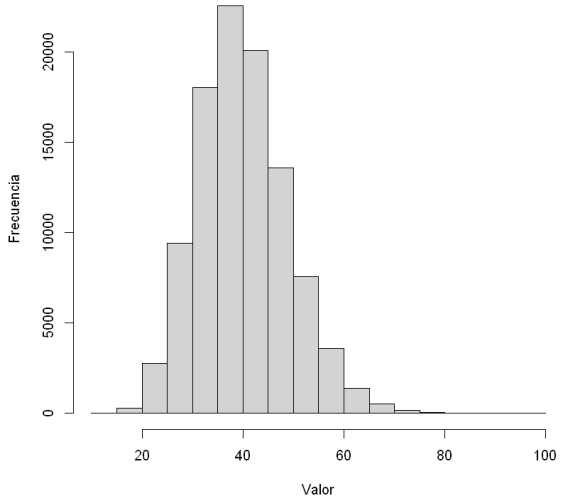
Se observa que a medida que crece el número de variables involucradas, la distribución se vuelve cada vez más simétrica, asimilándose a una distribución normal, lo cual se nota especialmente cuando se grafica la función de densidad de probabilidad de esta última, dada por la ecuación $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, junto a sus distribuciones normalizadas.

Esto se explica teniendo en cuenta el teorema central del límite (TCL). Este establece que dadas variables cualesquiera $Y_1, Y_2, ..., Y_m$ independientes (hipótesis la cual $X_1, X_2, ..., X_n$

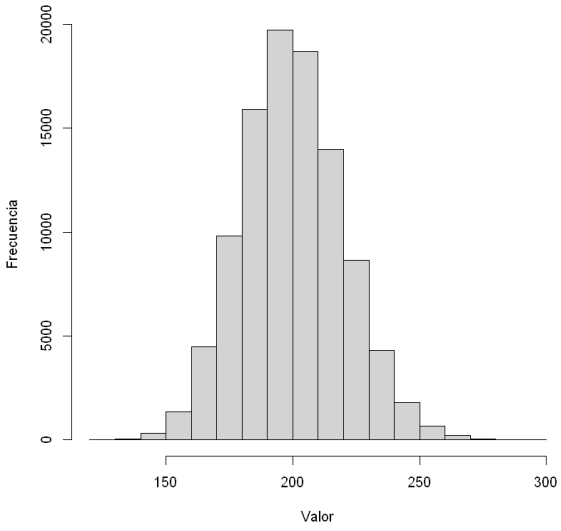
cumplen) con $E(Y_i) = \mu_i$, $V(Y_i) = {\sigma_i}^2$ finitas (nuevamente, las variables estudiadas lo cumplen), si $m \to \infty$ (lo cual se cumple en mayor medida a mayor es n), la distribución de $Y_1 + Y_2 + ... + Y_m$ será aproximadamente normal con parámetros $(\sum_{j=0}^{m} \mu_i, \sum_{j=0}^{m} {\sigma_i}^2)$. En particular, para el caso de las variables estudiadas, se puede ver que estos valores coinciden casi perfectamente con los anteriormente calculados con R. Es posible notar que al tener solo 5 o 20 sumandos, la gráfica es asimétrica mientras que al trabajar con la variable C, suma de 100 variables aleatorias, la gráfica es mucho más simétrica, asemejandose a la gráfica de de la fdp de una variable con distribución normal. Esto último se da porque el TCL es aplicable, en la práctica, cuando una variable es suma de más de 30 variables aleatorias (que además cumplan con las condiciones expuestas más arriba).



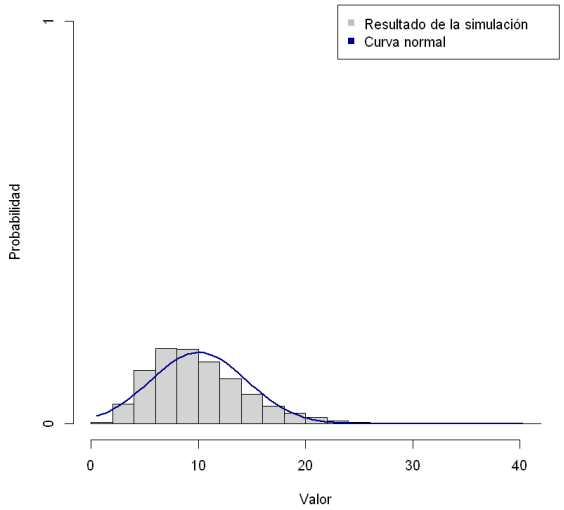
(a) Distribución de A.



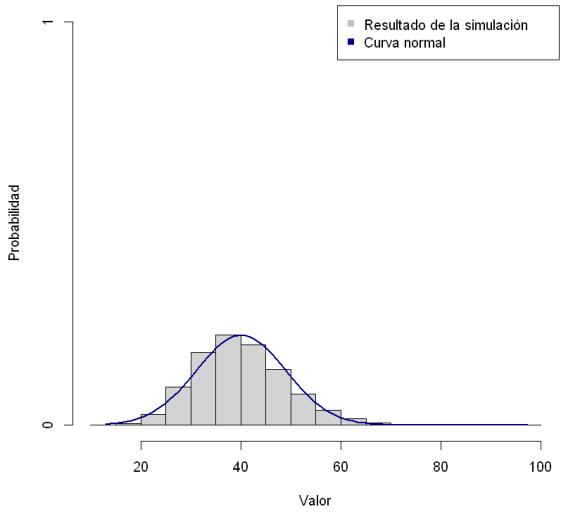
(b) Distribución de B.



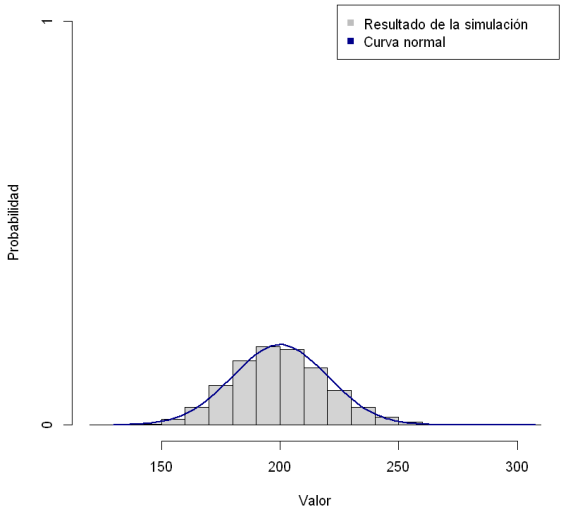
(c) Distribución de C.



(d) Distribucion Normalizada de A.



(e) Distribución Normalizada de B.



(f) Distribución Normalizada de C.

Figura 1: Gráficos de los resultados obtenidos.

Problema 2

Considere el espacio de estados $E = \{0, 1, 2, 3, 4, 5, 6, 7\}$. Cada uno de esos estados representa una esquina en una ciudad unidimensional. Consideremos un borracho que camina aleatoriamente en esta ciudad de acuerdo con las siguientes reglas:

- Si en el tiempo n el borracho se encuentra en alguna de las esquinas intermedias $1, 2, 3, 4, 5, 6$ entonces en el tiempo $n + 1$ avanza una cuadra con probabilidad p o retrocede una cuadra con probabilidad $1 - p$. Nunca se queda quieto.
- Toda vez que el borracho alcanza la esquina 0 (que corresponde a su casa) o la esquina 7 (que corresponde a un bar), allí se queda.

Simule y visualice 4 trayectorias distintas caminadas por el borracho para distintos valores de p ($p < 0{,}5$, $p = 0{,}5$ y $p > 0{,}5$) considerando que inicialmente se encuentra en la esquina k (elija algún valor de k entre 1 y 6 y manténgalo fijo para todas las simulaciones). Comente los resultados.

Solución. El problema en cuestión se puede modelar por medio de un proceso estocástico. Definiendo X = $\{X_n : n \in \mathbb{N}\}$, donde X_n = "la esquina en la que se encuentra el borracho en el momento n". En este caso, el espacio de estados es el proporcionado, $E = \{0, 1, 2, 3, 4, 5, 6, 7\}$, el cual es finito y discreto. De la misma manera, el espacio paramétrico T son los momentos, 0, 1, 2, ..., y es discreto, al ser infinito numerable. Además la esquina siguiente a la que se moverá el borracho depende únicamente de la esquina en la que se encuentra, por lo que: $\forall j_1, j_2, ...j_{n+1} \in E, \forall n \in T$,

$$P(X_{n+1} = j_{n+1} | X_1 = j_1, X_2 = j_2, ..., X_n = j_n) = P(X_{n+1} = j_{n+1} | X_n = j_n)$$

Es decir, este proceso cumple con la propiedad Markoviana, y T y E son discretos, y al ser un proceso estocástico que cumple con estas 3 propiedades, se tratará justamente de una Cadena de Markov.
Se define entonces $P(i, j) = P(X_{n+1} = j | X_n = i)$, que representa la probabilidad de transicionar de un estado i a otro j.

El saber que se trata de una Cadena de Markov proporciona información útil acerca del proceso: será posible comenzar a tratar el problema creando un diagrama de transiciones. Esto es, un grafo dirigido ponderado, donde cada nodo es un estado, y existirá una arista entre todo par de vértices $i, j \in E$, con peso P(i,j). Sin embargo, por convención, si P(i,j) = 0, no

dibujamos la arista en la representación.

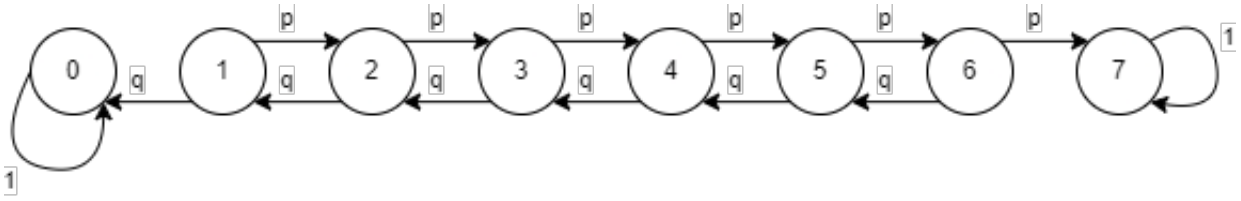


Figura 2: Diagrama de transiciones de X

Adicionalmente, a partir de la función de probabilidad de transición en un paso $P \in M^{n\times n}$, es posible determinar la matriz de transición P tal que $P_{i,j} = P(i,j)$. Se observa que esta matriz cumple las siguientes propiedades:

$$* \sum_{j\in E} P_{ij} = 1, \qquad \forall i \in E$$

$$* \; P_{ij} \geq 0, \qquad \forall i,\ j \in E$$

Por lo tanto, se tratará de la matriz estocástica del problema.

Por otro lado, se puede verificar que, al ser $0 \leq p \leq 1$ y $q = 1 - p$, que todas las filas de P, π, serán vectores de probabilidades, pues la sumatoria de sus componentes será igual a 1.

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 & 0 & 0 & 0 \\ 0 & q & 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & q & 0 & p & 0 & 0 & 0 \\ 0 & 0 & 0 & q & 0 & p & 0 & 0 \\ 0 & 0 & 0 & 0 & q & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Matriz de transición del problema del borracho.

En este problema, es posible identificar clases de recurrencia, que en este caso, serán únicamente los dos estados absorbentes: 0 y 7, pues $P(0,0) = P(7,7) = 1$. Y todos los demás estados serán transitorios, pues $F(j,j) < 1, \forall j \in E \setminus \{0,7\}$. Adicionalmente, las clases de recurrencia $C_0 = \{0\}$ y $C_1 = \{7\}$ serán a su vez los únicos conjuntos cerrados, es decir, $i \in C_k \wedge j \in E \setminus C_k \Longrightarrow \neg(i \to j)$, para $k = 0, 1$, y serán irreducibles, pues claramente ningún

subconjunto de estos será cerrado.

Ahora bien, en el problema, se debe elegir un valor k como la esquina inicial. Esta será la distribución inicial X_0, y en este caso, será 3,

```r
library(markovchain)
set.seed(123) # Para reproducibilidad
n_steps = 20
sim_borracho <- function(p_val, initial_k) {
  # Definir matriz de transicion para diferentes valores de p
  transition_matrix <- matrix(0, nrow = 8, ncol = 8)
  for (i in 1:8) {
    if (i %in% c(2, 3, 4, 5, 6, 7)) {
      transition_matrix[i, i + 1] <- p_val
      transition_matrix[i, i - 1] <- 1 - p_val
    } else if (i == 1 || i == 8) {
      transition_matrix[i, i] <- 1
    }
  }
  # Crear objeto de la cadena de Markov
  mc <- new("markovchain", states = c("0", "1", "2", "3", "4", "5", "6", "7"
   ), byrow = TRUE, transitionMatrix = transition_matrix)
  # Simulaci n de trayectorias
  simulated_paths <- simulated_paths <- c(initial_k, rmarkovchain(n = n_
   steps - 1, object = mc, t0 = initial_k))

  # Visualizacion
  plot(simulated_paths, type = "b", pch = 19, xlab = "Tiempo", ylab = "
   Esquina",
    main = paste("Trayectorias con p =", p_val))
}
sim_borracho(0.85, 3)
sim_borracho(0.75, 3)
sim_borracho(0.6, 3)
# ...
```

Los resultados obtenidos de estas simulaciones son los siguientes:

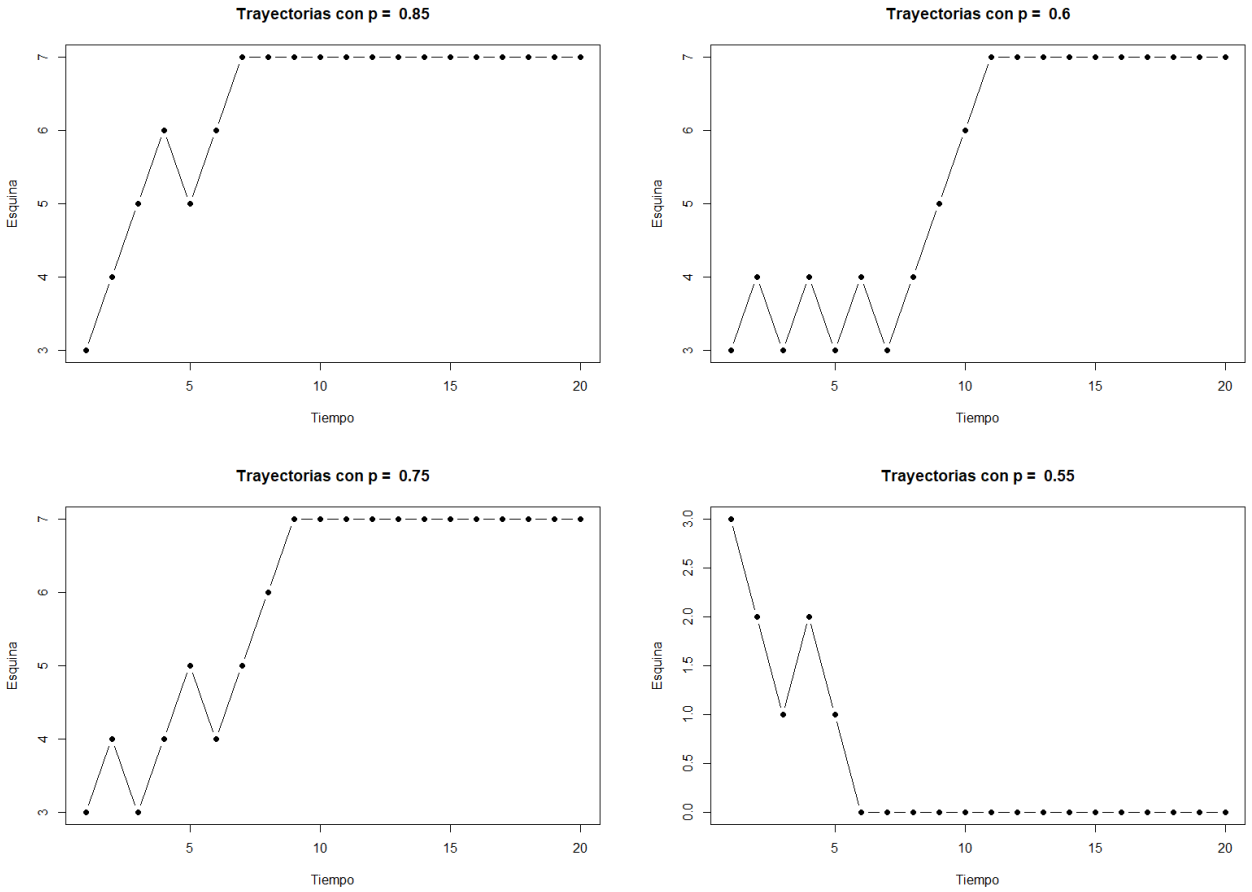


Figura 3: Gráficos de los resultados obtenidos con $p > 0{,}5$.

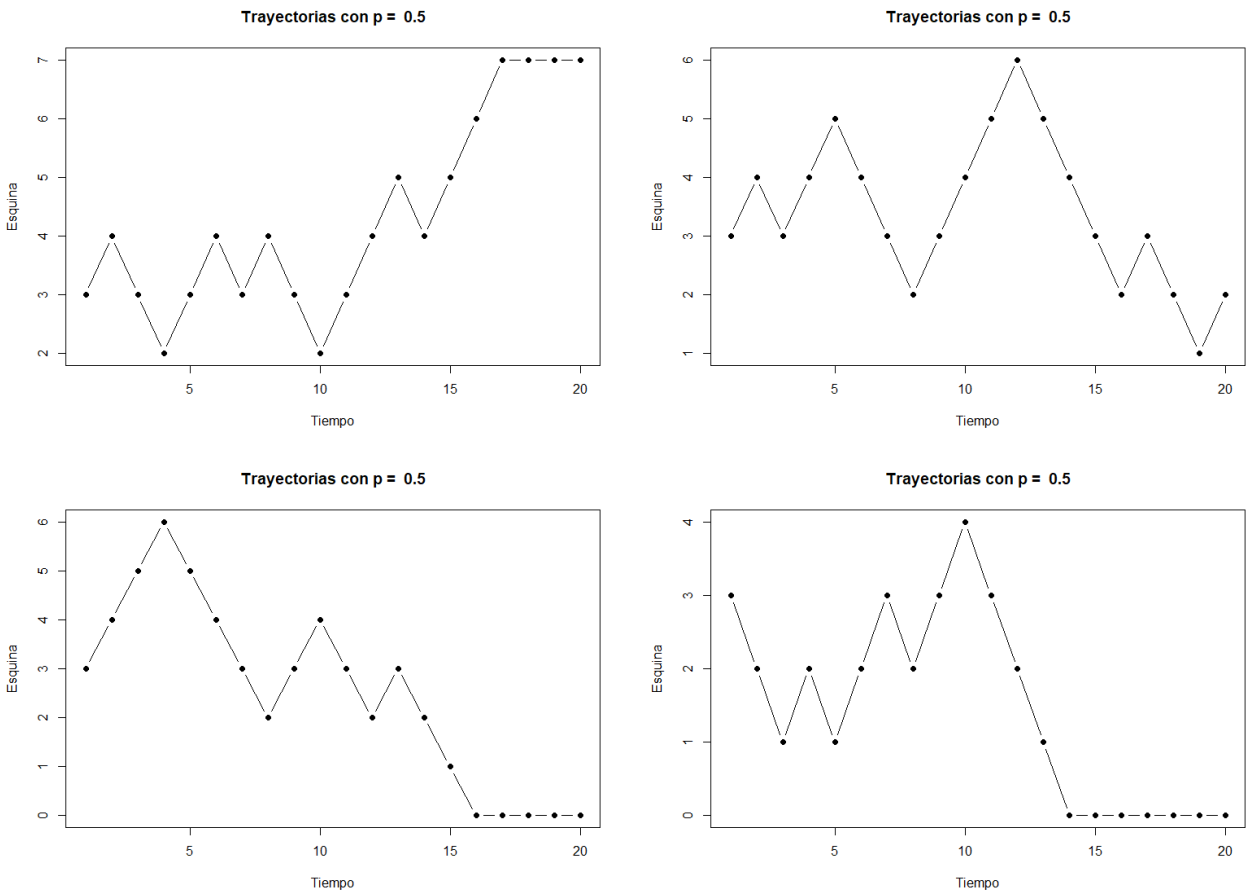


Figura 4: Gráficos de los resultados obtenidos con $p = 0{,}5$.

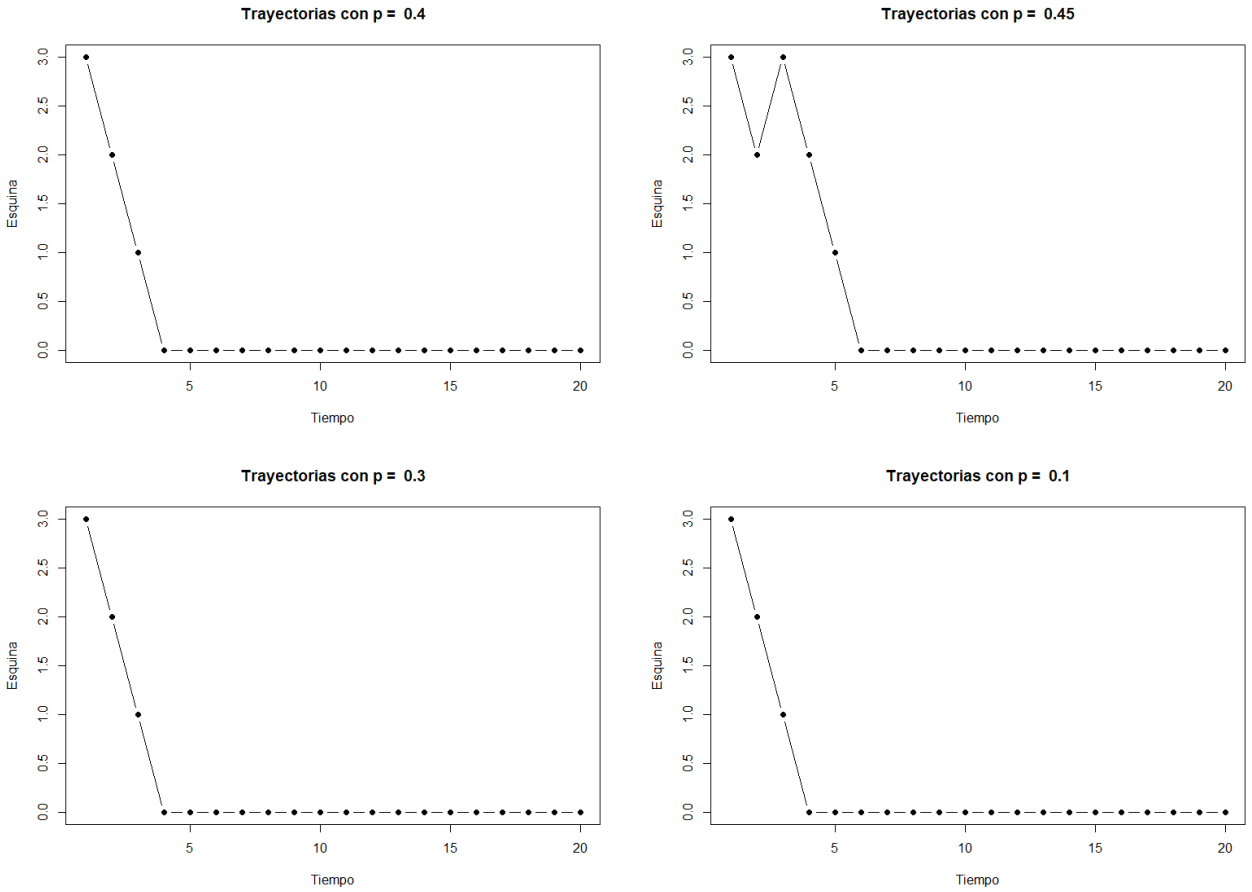


Figura 5: Gráficos de los resultados obtenidos con $p < 0{,}5$.

Se puede notar que la caminata del borracho es una CM similar a la de la caminata aleatoria vista en la teoría. Como era previsible, cuando $p > 0{,}5$ luego de unos instantes de observación, la tendencia es llegar al estado 7 rápidamente (evidentemente, existe la posibilidad de llegar al estado 0, pero no es lo más común). Por otro lado, cuando es $p < 0{,}5$, ocurre lo contrario, se llega rápidamente al estado 0. El caso cuando es $p = 0{,}5$ varía un poco, no hay una tendencia clara referente a qué estado absorbente es más seguro alcanzar y además, puede observarse que la longitud de la trayectoria es más larga (se tardan más instantes de tiempo en llegar a un estado absorbente).

Esto podría también formalizarse hayando el valor de $F(3, k), k = 0, 7$, que determina la probabilidad de que se llegue al estado 0 y 7 en una cantidad finita de pasos:

$$F(i, j) = \sum_{n=0}^{\infty} F_n(i, j)$$

Donde F_n se define como

$$F_1(i,j) = P(i,j)$$

$$F_n(i,j) = \sum_{b \in E \setminus \{j\}} P(b,j) * F_{n-1}(i,b)$$

y es la función que determina la probabilidad de llegar desde el estado i a j en n pasos.

Problema 3

Una persona está buscando cierta información en un universo de tan solo 5 páginas Web. Cada unade estas páginas puede tener uno o varios links a alguna otra. También puede haber páginas que no posean links. La persona elegirá, a partir de la página actual que está mirando, la próxima a visitar seleccionando con igual probabilidad alguna de las linkeadas. Si la página actual no posee link alguno, entonces seleccionará con igual probabilidad cualquiera de las 5 páginas Web existentes. El esquema de red simplificado se presenta en la figura 6.

a) Modele el comportamiento de visitas a las páginas como una cadena de Markov. Especifique la matriz de transición en un paso y realice el grafo correspondiente.

b) Determine la probabilidad de visitar la página j, para $j = 1, ..., 5$.

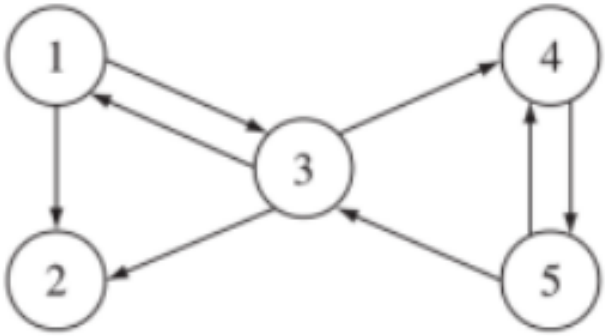


Figura 6: Red simplificada.

Solución.

a) Si se plasma el comportamiento del sistema en una cadena de Markov con las pautas indicadas, se obtiene una cadena finita de características como la de la de la figura 7.
Se puede ver que la cadena es **irreducible** ya que E es irreducible puesto que todos los estados están comunicados. Además, E es cerrado y finito por lo que todos los estados son recurrentes, es decir, **la cadena es recurrente**.

Además, el estado 2 tiene un lazo, por lo tanto necesariamente es aperiódico. Esto se mantiene porque un estado j tendrá período δ si y sólo si $\delta = mcd\{n : P^n(j,j) > 0\}$ y además $\delta \geq 2$, sin embargo, como se puede observar en la cadena, y más fácilmente en la matriz de transición, $P^1(2,2) = \frac{1}{5}$. Luego δ no puede ser mayor a 1 y el estado 2 no puede ser periódico. Es posible entonces, por teorema, establecer que, al tratarse de una cadena irreducible con al

menos un estado aperiódico, todos sus estados serán aperiódicos, y **la cadena es aperiódica**. Se ha establecido que la cadena es irreducible, recurrente y aperiódica. Es decir, cumple con la definición de **cadena ergódica**.

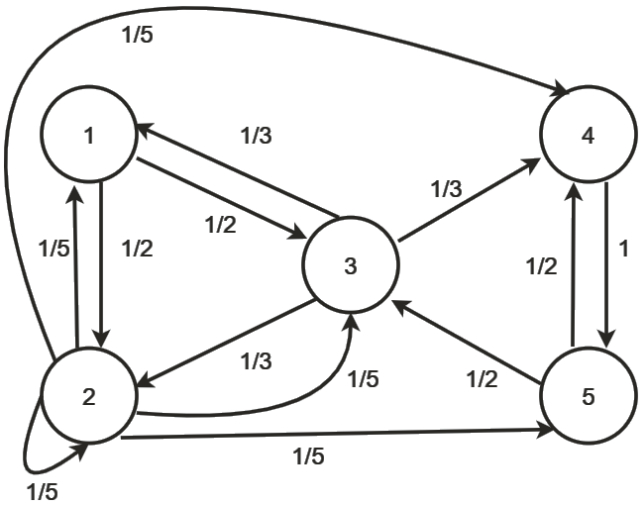


Figura 7: Cadena de Markov que modela la red.

$$
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\
\frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\
\frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & \frac{1}{2} & \frac{1}{2} & 0
\end{pmatrix}
\end{array}
$$

Matriz de transición en un paso del problema.

b) Por teorema, se tiene que si dos estados i, j pertenecen a un mismo conjunto irreducible, entonces $F(i, j) = 1$. Esta hipótesis se cumple para cualquier par de estados de la cadena, por lo cual todos los estados tendrán probabilidad 1 de ser visitados en un número finito de pasos, independientemente de las características particulares de la distribución inicial.

Si analizamos la distribución estacionaria, sabemos que las cadenas de Markov ergódicas son herramientas algorítmicas útiles porque, independientemente de su estado inicial, acaban alcanzando una distribución estacionaria única. Calculamos la misma haciendo uso de R:

```r
P <- matrix(c(0, 1/2, 1/2, 0, 0,
              1/5, 1/5, 1/5, 1/5, 1/5,
              1/3, 1/3, 0, 1/3, 0,
              0, 0, 0, 0, 1,
              0, 0, 1/2, 1/2, 0), nrow = 5, byrow = TRUE)


markov_chain <- new("markovchain", name = "Cadena", states = c("1", "2", "3"
    , "4", "5"), transitionMatrix = P)

print(steadyStates(markov_chain))
```

Luego, la distribución estacionaria es:

$$\pi \approx \begin{bmatrix} 0{,}105 & 0{,}158 & 0{,}221 & 0{,}242 & 0{,}274 \end{bmatrix}$$

Se verifica que las componentes de π suman 1. $\pi(j)$ representa la probabilidad de quedar en la página j a largo plazo. También, puede interpretarse como la proporción de tiempo que la persona se mantendrá en la página j.

Problema 4

Los clientes llegan a una tienda de acuerdo con un proceso de Poisson de tasa $\lambda = 4$ por hora. Si la tienda abre a las 9 a.m.:

a) Simule una trayectoria de dicho proceso para la primera hora de la jornada.

b) Grafique dicha trayectoria, tratando de que la gráfica refleje las características fundamentales de dicho proceso.

c) Simule una trayectoria del proceso para las horas hábiles de la tienda en una semana (de 9 a.m. a 6 p.m. de lunes a viernes) y registre la variable T: Tiempo transcurrido entre la llegada de dos clientes sucesivos.

d) Realice un análisis descriptivo gráfico y numérico conveniente para la variable T.

e) ¿Qué modelo de probabilidad podría ajustarse a la variable T?

Solución.

a) b) Hacemos en R la simulación pedida.

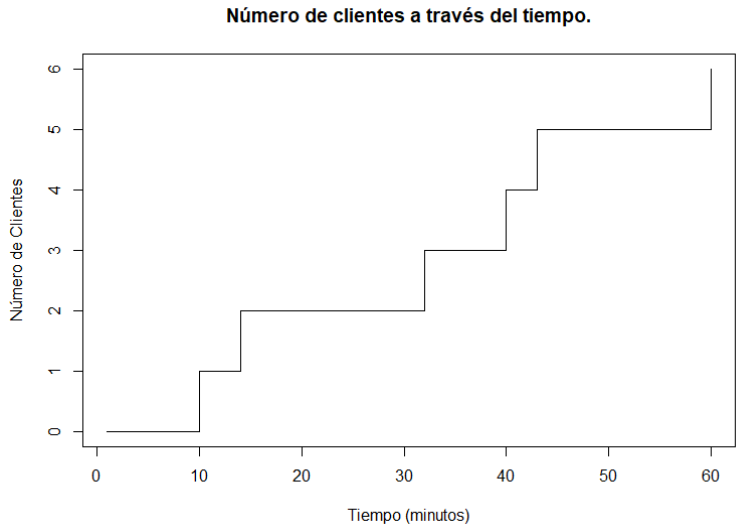


Figura 8: Simulación de la llegada de clientes durante la primera hora de trabajo.

Notemos que la gráfica refleja ciertas carácterísticas comúnes a los experimentos cuya variable aleatoria asociada se distribuye respondiendo a un proceso de Poisson. Como lo son:

- Ordenada al origen en 0.

- Gráfica escalonada no decreciente.
- Saltos de longitud unitaria que representan un momento de llegada.
- El número de clientes es un valor discreto. Mientras que el tiempo transcurrido durante la hora es un valor continuo.

Estas características se deben a las hipótesis de trabajo de un proceso Poisson

a) $N_0 = 0$. Esto es, en el momento $t = 0$ aún no se han producido arribos.

b) $\forall t > 0,\ N_t\ Po(\lambda t)$, donde N_t es la cantidad de arribos en el intervalo [0, t).

c) En ningún instante $t \in [0, \infty)$ se puede producir más de un arribo.

d) Para intervalos de tiempos no superpuestos, la cantidad de arribos en un intervalo es independiente de la cantidad de arribos en cualquier otro.En este sentido decimos que el proceso de Poisson no tiene memoria.

e) La cantidad de arribos en cualquier intervalo acotado es finita.

f) La cantidad de arribos en un intervalo de longitud t de la forma $(s, s + t]$, $s, t \geq 0$, no depende del inicio del intervalo (s), sino sólo de la longitud del mismo (t).

c) d) e) Se calcula la cantidad de horas que componen una semana laboral y simula el número de clientes que llegan a la tienda durante esa cantidad de horas. Con esto, se puede realizar la gráfica 9. Además, con la información de la cantidad de clientes que ingresaron a la tienda durante la semana, se puede graficar el tiempo transcurrido entre los arribos de dos clientes consecutivos (variable T).

Es preciso notar que la variable T: 'Tiempo transcurrido entre la llegada de dos clientes sucesivos' tendrá una distribución exponencial. La distribución de Poisson se estudió como una distribución de un solo parámetro λ, donde λ puede interpretarse como el número promedio de sucesos en una unidad de "tiempo". Se considera ahora la variable aleatoria descrita por el *tiempo necesario para que se produzca el primer suceso*. Utilizando la distribución de Poisson, se ve que la probabilidad de que no ocurra ningún arribo en el intervalo hasta el tiempo t viene dada por:

$$\frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}$$

Ahora se puede hacer uso de lo anterior y poner la variable X como el tiempo transcurrido hasta el primer evento de Poisson. La probabilidad de que el tiempo transcurrido hasta el

primer arribo supere x es es la misma que la probabilidad de que no ocurra haya ocurrido ningún arribo de Poisson hasta x. Esto último viene dado por $e^{-\lambda t}$. Como resultado, se tiene:

$$P(X > x) = e^{-\lambda x}$$

Así, la función de distribución acumulada para X viene dada por:

$$P(0 \leq X < x) = 1 - e^{-\lambda x}$$

Cuando se deriva esta fórmula, se obtiene:

$$f(x) = \lambda e^{-\lambda x}$$

Que es la función de densidad de probabilidad dada en teoría para una variable aleatoria continua con distribución exponencial.

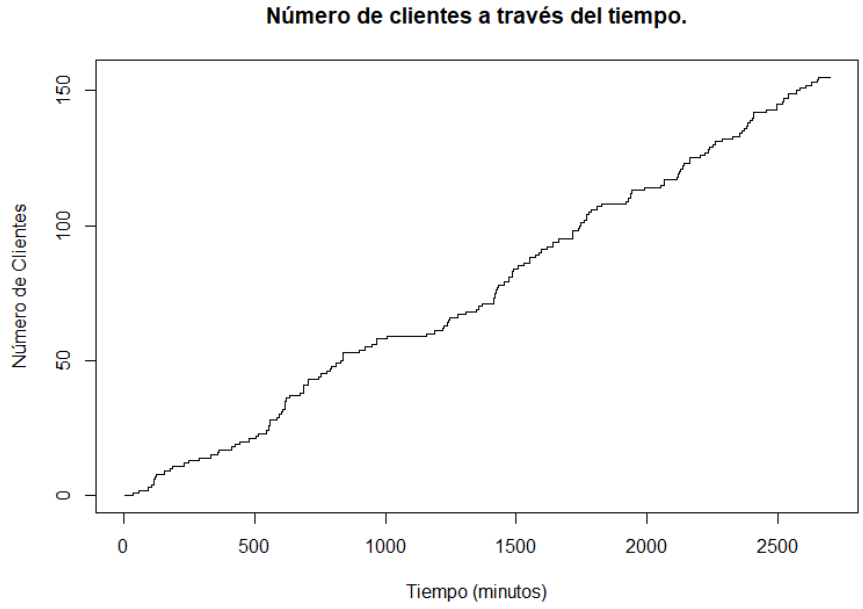


Figura 9: Simulación de la llegada de clientes durante una semana de trabajo.

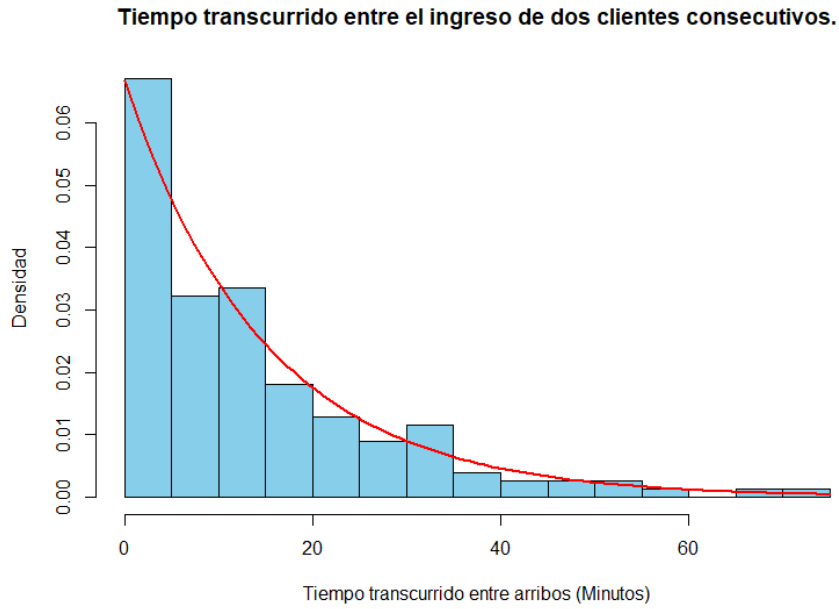


Figura 10: Simulación de la variable T: 'Tiempo transcurrido entre la llegada de dos clientes sucesivos'

La simulación hecha para la variable T nos arroja una media

$$\mu = 16{,}43521$$

cercana a la media esperada teórica, tomando $\lambda = \frac{4}{60}$:

$$\mu = \frac{1}{\lambda} = \frac{60}{4} = 15$$

Es decir, se espera que el tiempo promedio entre llegadas de nuevos clientes ronde los 15 minutos. Para el desvío estándar (σ), el resultado es similar puesto que $\mu = \sigma$ para la distribución exponencial.

Problema 5

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 se 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:

$$\begin{array}{c} \\ 80 \\ 135 \\ 139 \\ 445 \\ No\ attack \end{array} \begin{array}{c} \begin{array}{ccccc} 80 & 135 & 139 & 445 & No\ attack \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & \frac{8}{13} & \frac{3}{13} & \frac{1}{13} & \frac{1}{13} \\ \frac{1}{16} & \frac{3}{16} & \frac{3}{8} & \frac{1}{4} & \frac{1}{8} \\ 0 & \frac{1}{11} & \frac{4}{11} & \frac{5}{11} & \frac{1}{11} \\ 0 & \frac{1}{8} & \frac{1}{2} & \frac{1}{8} & \frac{1}{4} \end{pmatrix} \end{array}$$

con distribución inicial $\pi_0 = (0, 0, 0, 0, 1)$.

a) Después de tres 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.

Solución.

a) Sea X_i el estado en que se encuentra la cadena en la i-ésima semana. Por tratarse de una matriz de una cadena de markov, cuenta con las propiedades:

- $P(X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n) = P(X_0 = i_0)P(i_0, i_1) \cdots P(i_{n-1}, i_n)$
- $P(X_{m+n} = j | X_m = i) = P^n(i, j)\ \forall i, j \in E$

de lo que se deduce:

$$P(X_n = j) = (\pi_0 P^n)(j)\ \forall n, j \in E$$

donde $\pi_0(i) = P(X_0 = i)\ \forall i \in E$.

Así, la probabilidad de que cada puerto sea el más atacado durante la tercera semana estará dada por

$$\pi_0 P^3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{32} & \frac{61}{286} & \frac{885}{2288} & \frac{1019}{4576} & \frac{167}{1144} \\ \frac{4387}{237952} & \frac{11964081}{34027136} & \frac{5400719}{17013568} & \frac{3203451}{17013568} & \frac{2113687}{17013568} \\ \frac{3087}{146432} & \frac{5128733}{20939776} & \frac{3700267}{10469888} & \frac{2425947}{10469888} & \frac{1558587}{10469888} \\ \frac{579}{25168} & \frac{1591501}{7198048} & \frac{633335}{1799512} & \frac{1898595}{7198048} & \frac{504509}{3599024} \\ \frac{885}{36608} & \frac{1268269}{5234944} & \frac{911497}{2617472} & \frac{152189}{654368} & \frac{399807}{2617472} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{885}{36608} & \frac{1268269}{5234944} & \frac{911497}{2617472} & \frac{152189}{654368} & \frac{399807}{2617472} \end{bmatrix} \approx \begin{bmatrix} 0{,}024 & 0{,}242 & 0{,}348 & 0{,}232 & 0{,}152 \end{bmatrix}$$

Se observa que las componentes del vector efectivamente suman 1.
De esta manera, puede verse que el puerto más atacado será el puerto 139.

b) La distribución límite existe y será única. Con argumentos similares a los usados en el ejercicio 3, se tiene que la CM es:

- **Irreducible**: Puesto que todos los estados están comunicados entre sí.
- **Aperiódica**: Puesto que se tienen estados i para los cuáles $P(i, i) > 0$.
- **Recurrente**: Puesto que E es cerrado, finito e irreducible.

de esto se sigue que la CM es **ergódica**.

Luego, la CM tiene una distribución límite única. La calculamos con R:

```
P <- matrix(c(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), nrow = 5, byrow = TRUE)
markov_chain <- new("markovchain", name = "Cadena",
                    states = c("80", "135", "139", "445", "NO"),
                    transitionMatrix = P)
print(is.irreducible(markov_chain))
  # -> TRUE
print(recurrentStates(markov_chain))
  # -> "80"  "135" "139" "445" "NO"
print(period(markov_chain))
  # -> 1
print(steadyStates(markov_chain))
  # -> 0.02146667 0.2669333 0.3434667 0.2273333 0.1408
```

Con lo que obtenemos:

$$\pi \approx \begin{bmatrix} 0{,}0215 & 0{,}2669 & 0{,}3435 & 0{,}2273 & 0{,}1408 \end{bmatrix}$$

Nuevamente, las componentes suman 1. Se puede interpretar a este vector como la probabilidad estacionaria de hallarse en cada estado. Así, se concluye que aproxidamente el 34 % del tiempo el puerto más atacado será el puerto 139.