

# Trabajo Práctico Final

## Probabilidad y Estadística

**Alumnos:**

**Cipullo, Inés**

**Bisiach, Ezequiel**

Universidad Nacional de Rosario

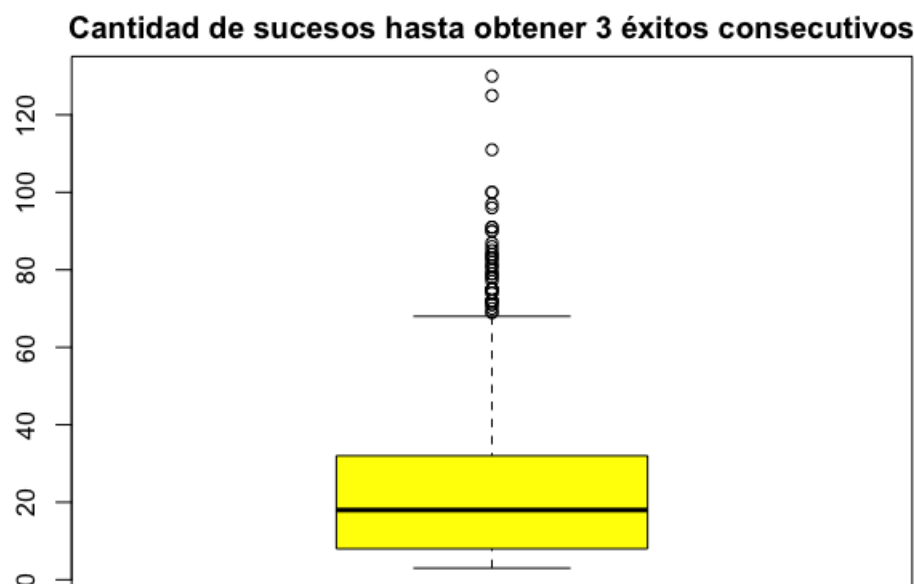
2021

## Problema 1

Tenemos un suceso cuya probabilidad de éxito es  $p$ . Se realizan intentos independientes de este suceso hasta obtener  $k$  éxitos consecutivos. Definimos el proceso  $N_k$  que denota el número de ensayos necesarios para obtener  $k$  éxitos consecutivos.

a)

Simulamos 1000 trayectorias del proceso  $N_3$ , es decir, buscando 3 éxitos consecutivos, con  $p = 0.4 < 0.5$ . Graficamos en un boxplot los resultados obtenidos.



Vemos que la mínima cantidad de sucesos es 3 y, en principio, no tiene límite superior.

Se presenta, además, una tabla con dicho resultado al final del trabajo, que incluye frecuencia absoluta y relativa según la cantidad de sucesos que se obtuvieron en las trayectorias simuladas.

**b)**

A partir de lo obtenido anteriormente, estimamos la esperanza de  $N_3$  haciendo la suma de cada cantidad de sucesos por su respectiva frecuencia relativa, quedando  $E(N_3) = 23.255$ .

## Problema 2

Tenemos el proceso  $D_n = 2 \cdot I_n - 1$ , que representa el cambio de posición de una partícula que se mueve a lo largo de una línea recta con saltos de magnitud 1 en cada momento. Esto se debe a que  $D_n$  depende del proceso de Bernoulli  $I_n$ : “la señal emitida en el momento  $n$  es correcta”. Lo definimos como sigue:

- $I_n = 1$  si la señal emitida en el  $n$ -ésimo momento es correcta
- $I_n = 0$  si la señal emitida en el  $n$ -ésimo momento NO es correcta

Tomamos el estado inicial del proceso como  $N_0 = 0$ , el comienzo de la trayectoria.

**a)**

Una simulación de 50 pasos de una trayectoria del proceso  $D_n$ , para un proceso Bernoulli  $I_n$  con una probabilidad de éxito  $p = 0.75 > 0.7$ , es como sigue:

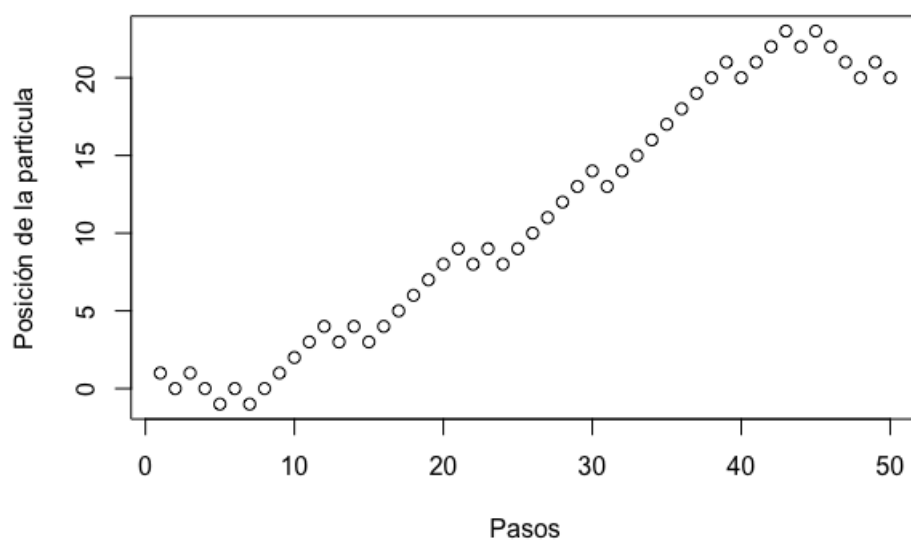
$$D_n = [1, -1, 1, -1, -1, 1, -1, 1, 1, 1, 1, 1, -1, 1, -1, 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, -1, 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, 1, 1, 1, -1, 1, 1, 1, -1, 1, -1, -1, -1, 1, -1]$$

b)

Definimos otro proceso,  $S_n$ , proceso estocástico, como la posición de la partícula en el momento  $n$ . Notamos que  $S_i = D_1 + D_2 + \dots + D_i$ , es decir,  $S_i$  es la suma acumulada hasta  $D_i$ .

Del resultado antes expuesto, obtenemos una trayectoria  $S_n$  como sigue:

**Simulación de 50 pasos del proceso  $S_n$**



$S_n = [1, 0, 1, 0, -1, 0, -1, 0, 1, 2, 3, 4, 3, 4, 3, 4, 5, 6, 7, 8, 9, 8, 9, 8, 9, 10, 11, 12, 13, 14, 13, 14, 15, 16, 17, 18, 19, 20, 21, 20, 21, 22, 23, 22, 23, 22, 21, 20, 21, 20]$

### Problema 3

Tenemos un proceso estocástico  $N_n$  que representa el estado actual de la apuesta de un jugador luego de  $n$  juegos. El estado inicial del proceso es  $N_0 = k$ , la apuesta inicial. Por cada juego, el jugador puede perder 1 o ganar 1 de su apuesta. Por lo tanto podemos definir el proceso  $X_n$  como  $X_n = 2 \cdot J_n - 1$ , donde  $J_n$  es el proceso de Bernoulli que representa: “el  $n$ -ésimo juego se ganó”, definido como sigue:

- $J_n = 1$  si el  $n$ -ésimo juego se ganó
- $J_n = 0$  si el  $n$ -ésimo juego NO se ganó

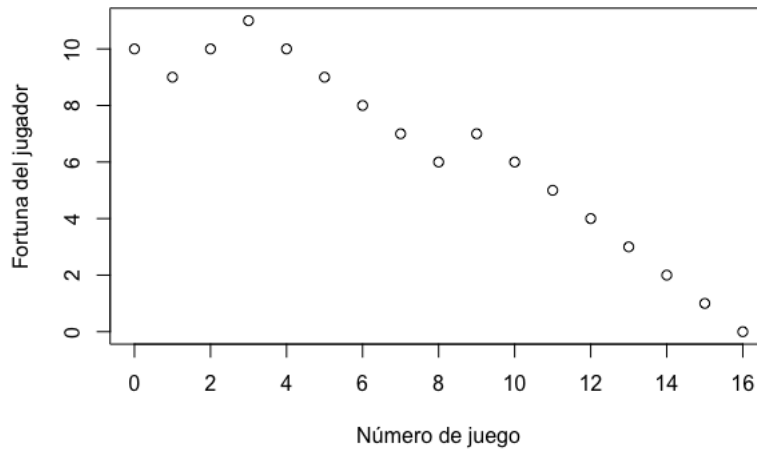
Para las simulaciones se toman los valores  $k = 10$  y  $s = 20$ , siendo  $p$  variable.

**a)**

Se realizan simulaciones de la evolución del capital del jugador, para distintos valores de  $p$  (probabilidad de éxito de  $J_n$ ), logrando los resultados que siguen:

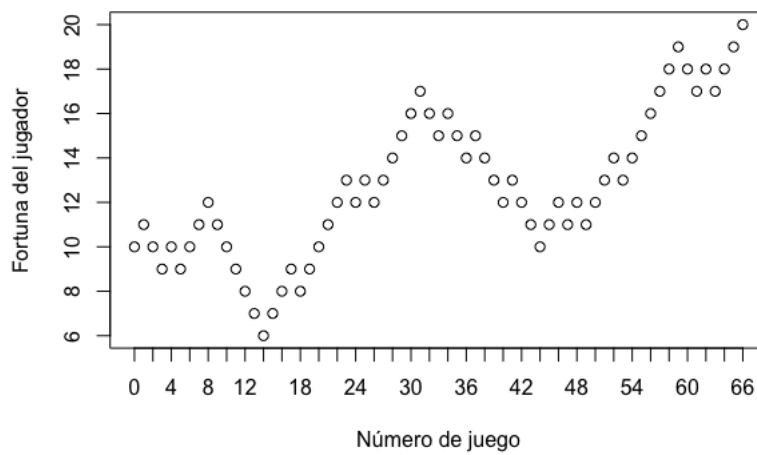
Para  $p = 0.25 < 0.5$ ,

**Simulación del proceso  $N_n$  con  $p = 0.25$**



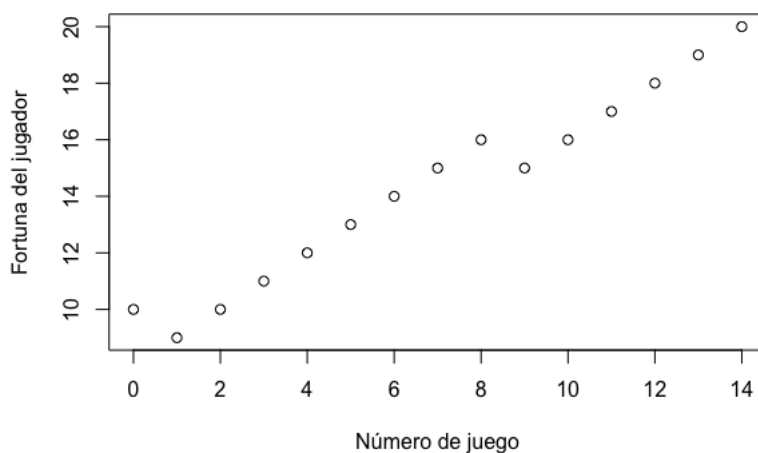
Para  $p = 0.5$ ,

### Simulación del proceso $N_n$ con $p = 0.5$



Para  $p = 0.8 > 0.5$ ,

### Simulación del proceso $N_n$ con $p = 0.8$



b)

Se simula una cantidad significativa de trayectorias del capital del jugador para luego calcular la probabilidad de ruina (que su fortuna llegue a 0), para los distintos valores de  $p$ , logrando los resultados que siguen:

Para  $p = 0.25 < 0.5$ , se obtiene una probabilidad de ruina de 1.

Para  $p = 0.5$ , se obtiene una probabilidad de ruina de 0.4975.

Para  $p = 0.8 > 0.5$ , se obtiene una probabilidad de ruina de 0.

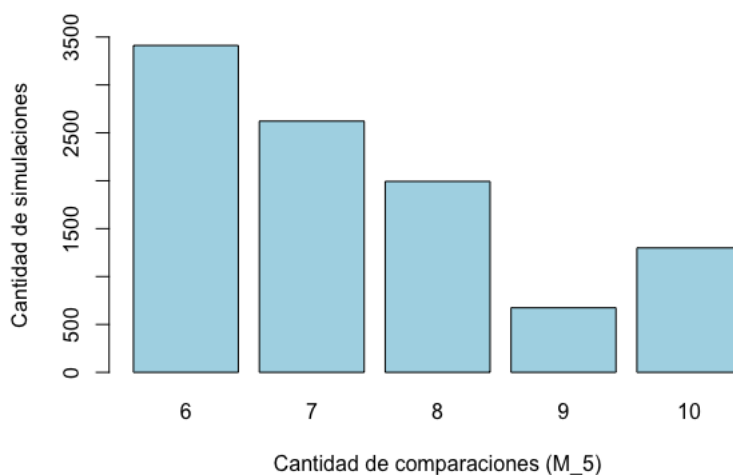
## Problema 4

Se busca medir la efectividad de un algoritmo que opera sobre conjuntos midiendo la cantidad de comparaciones que realiza. Denominamos  $M_n$  al número esperado de comparaciones que realiza el algoritmo para un conjunto de  $n$  elementos.

a)

Se toma  $n = 5$ , y luego de hacer 10000 simulaciones del valor de  $M_5$ , obtenemos las siguientes frecuencias absolutas de las cantidades de comparaciones posibles:

### Frecuencia absoluta del valor de $M_5$ en 10000 simulaciones



, donde la cantidad de comparaciones promedio es 7.3826. Vemos también que el espacio muestral de  $M_5$  es  $\{6, 7, 8, 9, 10\}$ .

b)



Análisis formal de la distribución de probabilidades de la variable aleatoria  $M_5$ :

Sea una lista  $l = [x_1, x_2, x_3, x_4, x_5]$ . Se tiene la misma probabilidad de elegir cualquier  $x_i, i = 1, \dots, 5$ , por lo tanto esta probabilidad es  $\frac{1}{n} = \frac{1}{5}$ .

Luego de elegir un  $x_i$ , se realizan  $n - 1$  comparaciones, en este caso son 4. Dependiendo la elección de  $x_i$ , se pueden generar diferentes listas a ordenar en el llamado recursivo:

- 0 elementos menores a  $x_i$  y 4 elementos mayores a  $x_i$
- 1 elemento menor a  $x_i$  y 3 elementos mayores a  $x_i$
- 2 elementos menores a  $x_i$  y 2 elementos mayores a  $x_i$
- 3 elementos menores a  $x_i$  y 1 elemento mayor a  $x_i$
- 4 elementos menores a  $x_i$  y 0 elementos mayores a  $x_i$

Sabemos que ordenar una lista vacía o con un elemento no requiere comparaciones, y que para ordenar una lista de 2 elementos se realiza 1 comparación. Además, ordenar una lista de  $m$  elementos tiene el mismo costo, independientemente de sus elementos.

Por lo tanto, llegamos a que hay:

- $\frac{2}{5}$  de probabilidad de ordenar una lista de 4 elementos
- $\frac{2}{5}$  de probabilidad de ordenar una lista de 3 elementos
- $\frac{1}{5}$  de probabilidad de realizar 2 comparaciones

Vemos que es necesario conocer la cantidad de comparaciones que se realizan sobre listas de 3 y 4 elementos.

Para listas de 3 elementos, luego de elegir un elemento al azar y realizar 2 comparaciones, hay  $\frac{2}{3}$  de probabilidad de que se seleccione el menor o mayor elemento de la lista y por eso realizar 3 comparaciones en total (se le suma solo una comparación a las 2 iniciales), y hay  $\frac{1}{3}$  de probabilidad de que se seleccione el elemento central, en cuyo caso no se deberán realizar más comparaciones, debiendo realizar 2 comparaciones en total.

Para listas de 4 elementos, luego de elegir un elemento al azar y realizar 3 comparaciones, hay  $\frac{1}{2}$  de probabilidad de que se seleccione el menor o mayor elemento de la lista y que reste ordenar una lista de 3 elementos, y hay  $\frac{1}{2}$  de probabilidad de que se haga 1 comparación más por haber elegido alguno de los elementos centrales.

Multiplicando estas probabilidades con lo obtenido para listas de 3 elementos, se obtiene que hay:

- $\frac{1}{2}$  de probabilidad de que se realicen 4 comparaciones
- $\frac{1}{6}$  de probabilidad de que se realicen 5 comparaciones
- $\frac{1}{3}$  de probabilidad de que se realicen 6 comparaciones

Juntando todos los calculos realizados, tenemos que la distribución de probabilidad de  $M_5$  es:

- $\frac{1}{3}$  de probabilidad de que se realicen 6 comparaciones
- $\frac{4}{15}$  de probabilidad de que se realicen 7 comparaciones
- $\frac{1}{5}$  de probabilidad de que se realicen 8 comparaciones
- $\frac{1}{15}$  de probabilidad de que se realicen 9 comparaciones
- $\frac{2}{15}$  de probabilidad de que se realicen 10 comparaciones

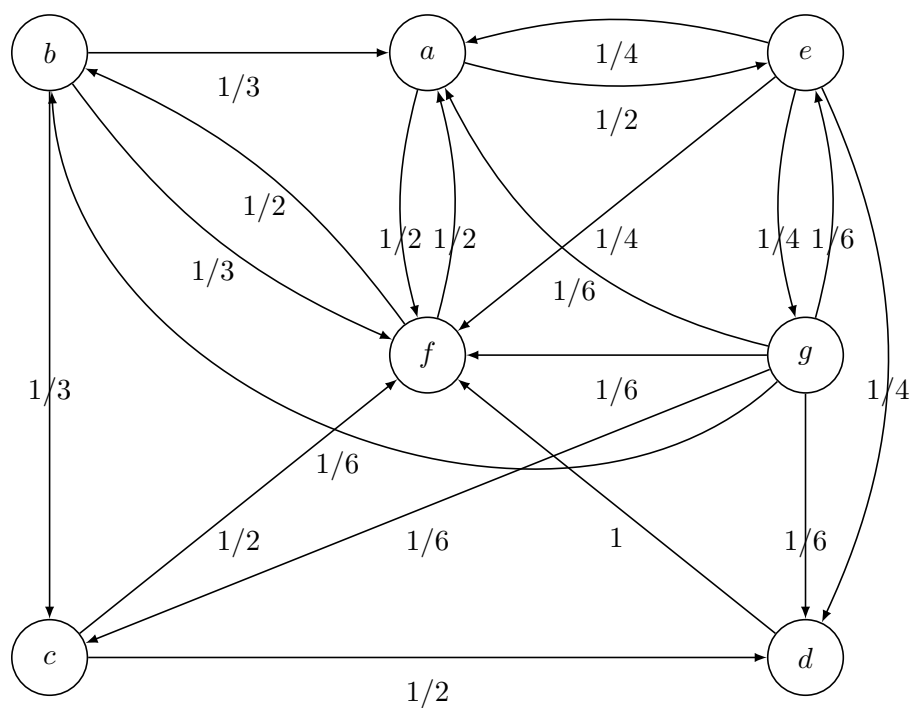
Por último, al ser cantidades numéricas, la esperanza se puede calcular como el promedio. Así, llegamos a que  $E(M_5) = 7.4$ , por lo que concluimos que el resultado obtenido en el apartado anterior es una buena aproximación. También resulta claro que la distribución de probabilidades concuerda con la simulación graficada.

## Problema 5

Contamos con una red simplificada de 7 páginas. La importancia de cada página se entiende como la probabilidad de que un usuario se encuentre en dicha página luego de un tiempo navegando la red. Esto es lo que mide el algoritmo **PageRank**, para poder asignar un orden de las páginas dentro de la red. Esta probabilidad está medida, entre otras cosas, por la cantidad de enlaces de hipertexto que dirigen al usuario a la página y el peso de estos, que depende de la importancia de la página que lo contiene y cuantos haya.

**a)**

El comportamiento de visitas de las páginas se puede modelar con la siguiente Cadena de Markov, y su respectiva matriz de transición en un paso.

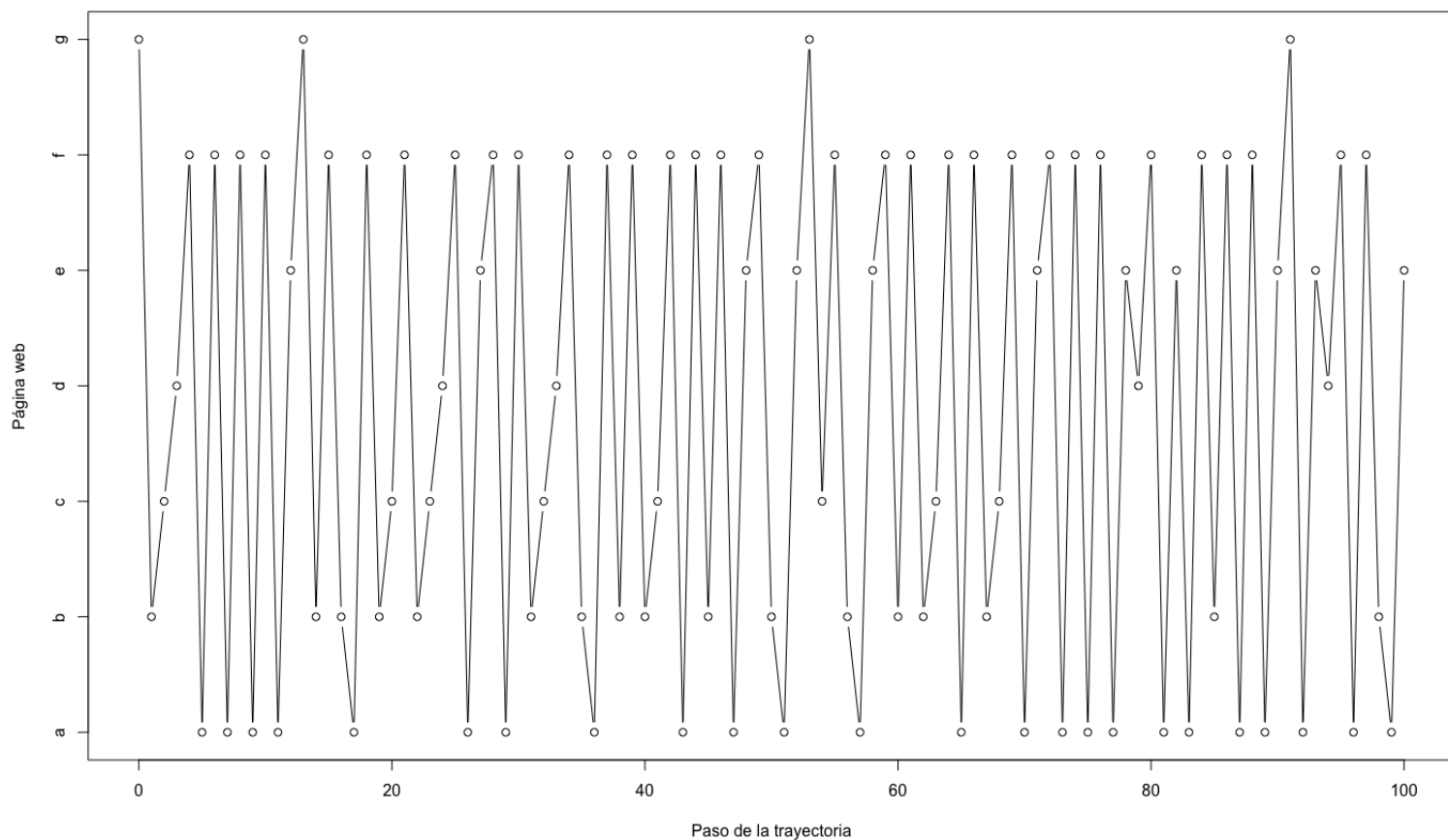


$$P = \begin{matrix} & \begin{matrix} a & b & c & d & e & f & g \end{matrix} \\ \begin{matrix} a \\ b \\ c \\ d \\ e \\ f \\ g \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1/4 & 0 & 0 & 1/4 & 0 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \end{pmatrix} \end{matrix} \quad (1)$$

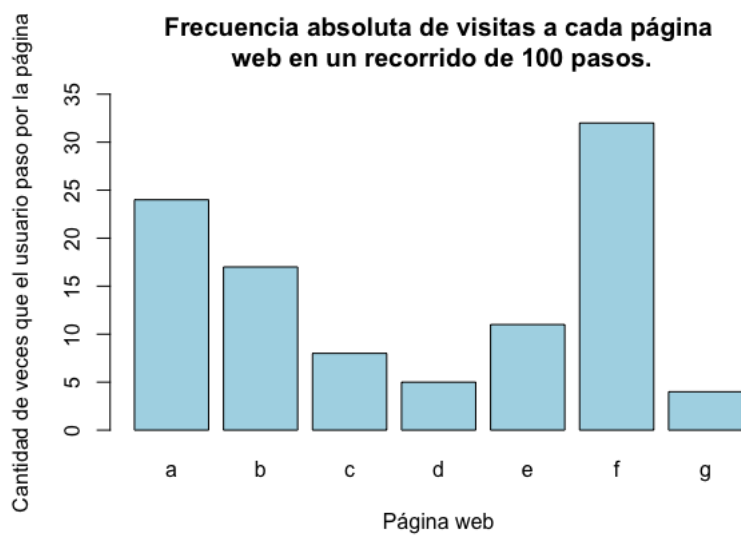
b)

Se simula una trayectoria de 100 pasos de un usuario en esta red. La página inicial es elegida al azar y todas las páginas tienen igual probabilidad de ser elegidas como tal.

**Simulación de una trayectoria de 100 pasos  
de un usuario en la red.**



Para que se visualicen un poco mejor ciertos datos, se muestra también un gráfico de barras sobre la frecuencia absoluta de la cantidad de visitas que recibió cada página web de la red en la trayectoria de 100 pasos del usuario que se simuló.



c)

El rango de página de una página web es la probabilidad a largo plazo de que un usuario se encuentre en dicha página, luego de haber estado navegando en la red. Analíticamente, esto equivale a calcular la distribución invariante  $\pi$ , la cual podemos asegurar que es única porque la cadena es cerrada, finita, irreducible y aperiódica.

Para ello, planteamos el siguiente sistema lineal de ecuaciones:

$$\begin{cases} \pi_a = \frac{1}{3} \cdot \pi_b + \frac{1}{4} \cdot \pi_e + \frac{1}{2} \cdot \pi_f + \frac{1}{6} \cdot \pi_g \\ \pi_b = \frac{1}{2} \cdot \pi_f + \frac{1}{6} \cdot \pi_g \\ \pi_c = \frac{1}{3} \cdot \pi_b + \frac{1}{6} \cdot \pi_g \\ \pi_d = \frac{1}{2} \cdot \pi_c + \frac{1}{4} \cdot \pi_e + \frac{1}{6} \cdot \pi_g \\ \pi_e = \frac{1}{2} \cdot \pi_a + \frac{1}{6} \cdot \pi_g \\ \pi_f = \frac{1}{2} \cdot \pi_a + \frac{1}{3} \cdot \pi_b + \frac{1}{2} \cdot \pi_c + \pi_d + \frac{1}{4} \cdot \pi_e + \frac{1}{6} \cdot \pi_g \\ \pi_g = \frac{1}{4} \cdot \pi_e \\ \pi_a + \pi_b + \pi_c + \pi_d + \pi_e + \pi_f + \pi_g = 1 \end{cases} \quad (2)$$

Luego de resolver el sistema (y comparar el resultado con la aproximación de R), el rango de cada página es:

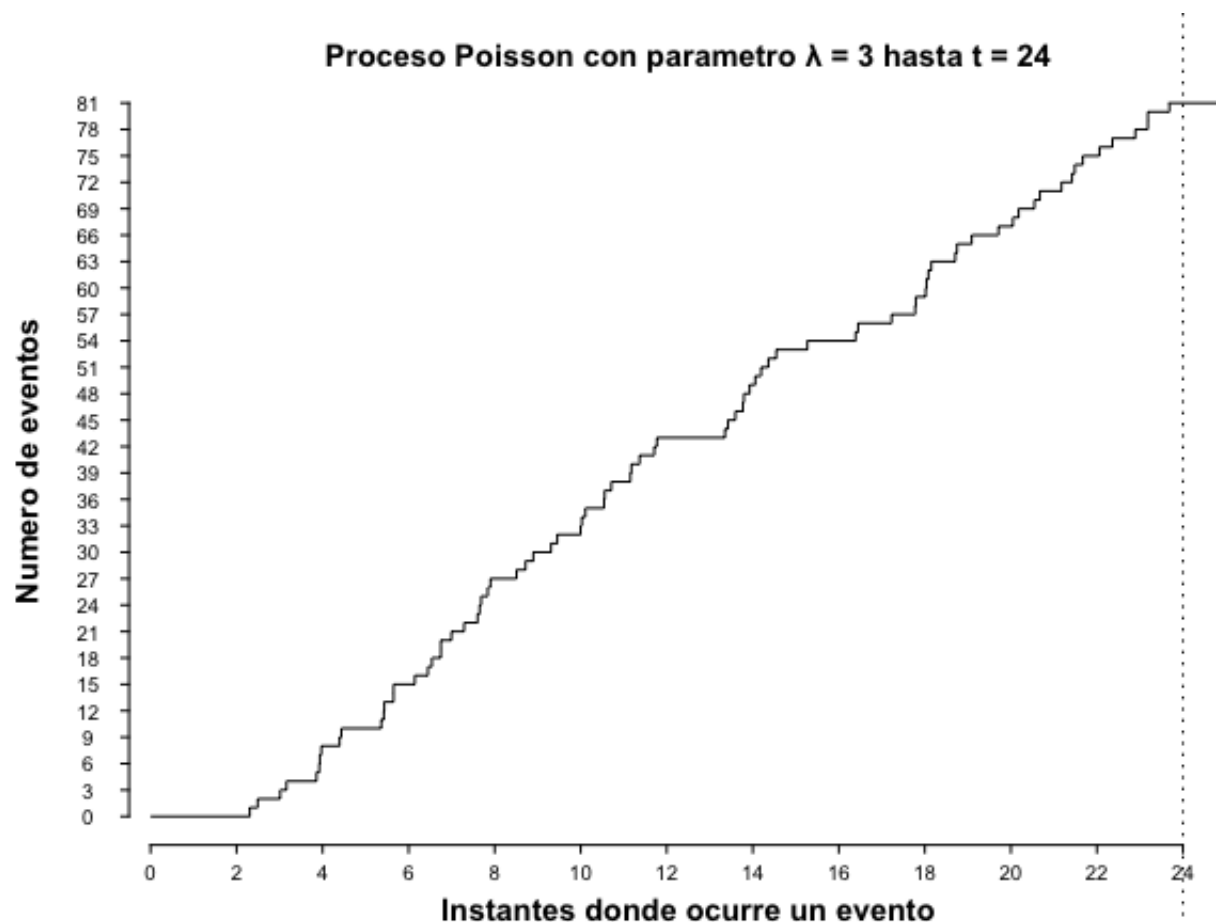
$$\pi = \begin{pmatrix} a & b & c & d & e & f & g \\ 0.2453 & 0.16 & 0.0587 & 0.0667 & 0.128 & 0.3093 & 0.032 \end{pmatrix} \quad (3)$$

## Problema 6

El evento bajo estudio es el número de vuelos que recibe un aeropuerto a una tasa  $\lambda$  aterrizajes/hora.

**a)**

Con una tasa de 3 aterrizajes por hora de acuerdo a un proceso de Poisson, el aeropuerto recibe vuelos a partir de las 2 am. Para visualizar el comportamiento de dicho proceso por 24 horas, se presenta el siguiente gráfico de la trayectoria obtenida.



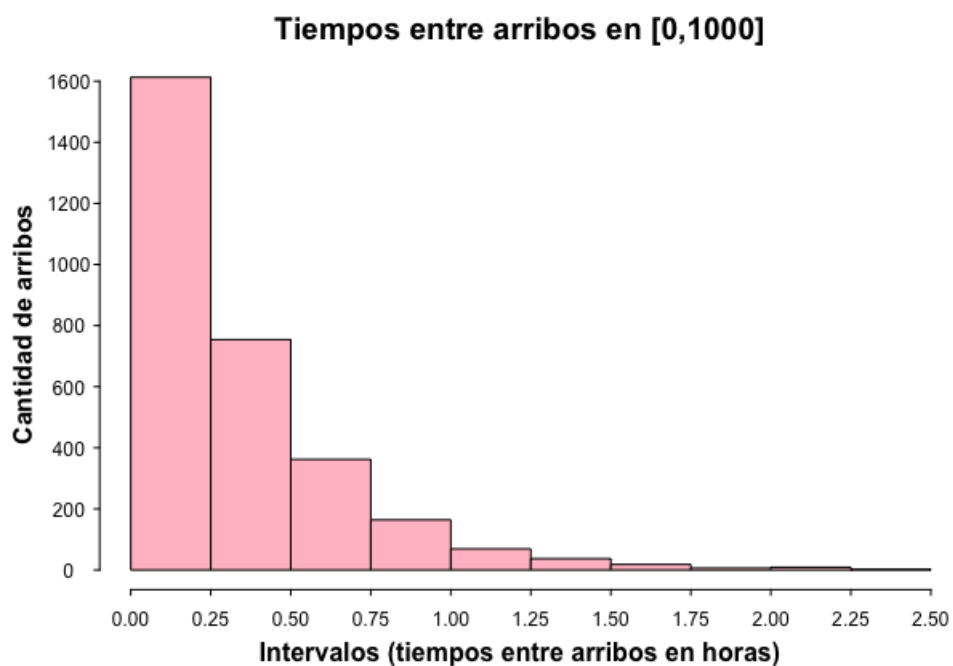
Esta trayectoria se obtiene de la siguiente forma: el número aleatorio de la distribución de Poisson con parámetro  $\lambda \cdot t$  retorna la cantidad de eventos que ocurren en el tiempo  $t$  en un Proceso de Poisson a una tasa  $\lambda$ . Luego las probabilidades de que un evento ocurra a un momento determinado es la misma para cada evento, por lo tanto se generan números aleatorios de la distribución uniforme hasta  $t$ , los cuales serán los instantes en los que ocurre un evento.

b)

Se realiza una simulación del proceso de Poisson antes descripto para un in-



tervalo de tiempo de  $T = 1000$  horas. Una vez que se saben los instantes en los que ocurre cada evento, se calculan las diferencias entre cada par de instantes consecutivos, obteniendo así los tiempos entre arribos. De este resultado, surge el siguiente histograma:



Además, sabemos que el tiempo transcurrido entre dos arribos sucesivos es independiente de los arribos de los que se trate y de los instantes de los arribos previos, y estos tienen una distribución exponencial. Por ello, su esperanza es  $\frac{1}{3}$ , lo cual concuerda con lo visto en el gráfico.

## Problema 7

Tenemos un honeypot, cuyos datos son estudiados utilizando cadenas de Markov. Se obtienen los 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 (NA). 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:

$$P = \begin{matrix} & \begin{matrix} 80 & 135 & 139 & 445 & NA \end{matrix} \\ \begin{matrix} 80 \\ 135 \\ 139 \\ 445 \\ NA \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 8/13 & 3/13 & 1/13 & 1/13 \\ 1/16 & 3/16 & 3/8 & 1/4 & 1/8 \\ 0 & 1/11 & 4/11 & 5/11 & 1/11 \\ 0 & 1/8 & 1/2 & 1/8 & 1/4 \end{pmatrix} \end{matrix} \quad (4)$$

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

a)

$P$  es la matriz de transición en un paso (una semana) y  $\pi_0$  la distribución inicial, por lo tanto  $\pi_0 * P^3$  será la distribución de probabilidad de que los puertos sean atacados luego de 3 semanas.

$$\pi_0 * P^3 = \begin{matrix} & \begin{matrix} 80 & 135 & 139 & 445 & NA \end{matrix} \\ \begin{matrix} 80 \\ 135 \\ 139 \\ 445 \\ NA \end{matrix} & \begin{pmatrix} 0.02417504 & 0.2422698 & 0.3482356 & 0.232574 & 0.1527455 \end{pmatrix} \end{matrix} \quad (5)$$

Para estar seguros de que corresponde a una distribución de probabilidades, notamos que su suma da 1.

Concluimos, entonces, que el puerto 80 es el de menor probabilidad de ser atacado y el puerto 139 es el de mayor probabilidad de ser atacado, quedando los puertos 135 y 445 en el medio.

b)

Tenemos la cadena de Markov definida por la matriz  $P$  de transición en un paso. Como  $E = \{80, 135, 139, 445, NA\}$  es un conjunto cerrado, irreducible y finito, deducimos que sus elementos son estados positivamente recurrentes. Además, tenemos que  $P^3(80, 80) = 0.03125 > 0$ ,  $P(135, 135) = 8/13 > 0$ ,  $P(139, 139) = 3/8 > 0$ ,  $P(445, 445) = 5/11 > 0$  y  $P(NA, NA) = 1/4 > 0$ . Luego todos los estados son aperiódicos.

Por lo tanto, tenemos una cadena de Markov cerrada, irreducible, finita, recurrente y aperiódica, es decir, tenemos una cadena de Markov ergódica.

Como  $\forall j \in E$ ,  $j$  es un estado positivamente recurrente y aperiódico, tenemos que

$$\forall i \in E, \lim_{n \rightarrow \infty} P^n(i, j) = F(i, j) \cdot \lim_{n \rightarrow \infty} P^n(j, j)$$

Además, sabemos que si  $a$  y  $b$  son ambos estados recurrentes y ambos pertenecen al mismo cerrado, se tiene  $F(a, b) = 1$ .

Por lo tanto,  $\forall i \in E, \lim_{n \rightarrow \infty} P^n(i, j) = \lim_{n \rightarrow \infty} P^n(j, j)$ .

Como la cadena de Markov es irreducible, finita y aperiódica, su distribución invariante ( $\pi$ ) es única y cumple que:

- $\pi \cdot P = \pi$
- $\sum_{i \in E} \pi_i = 1$

$$\blacksquare \pi_j = \lim_{n \rightarrow \infty} P^n(i, j) \quad \forall i, j \in E$$

Planteamos el correspondiente sistema lineal de ecuaciones para obtenerla.

$$\begin{cases} \pi_{80} = \frac{1}{16} \cdot \pi_{139} \\ \pi_{135} = \frac{8}{13} \cdot \pi_{135} + \frac{3}{16} \cdot \pi_{139} + \frac{1}{11} \cdot \pi_{445} + \frac{1}{8} \cdot \pi_{NA} \\ \pi_{139} = \frac{3}{13} \cdot \pi_{135} + \frac{3}{8} \cdot \pi_{139} + \frac{4}{11} \cdot \pi_{445} + \frac{1}{2} \cdot \pi_{NA} \\ \pi_{445} = \frac{1}{13} \cdot \pi_{135} + \frac{1}{4} \cdot \pi_{139} + \frac{5}{11} \cdot \pi_{445} + \frac{1}{8} \cdot \pi_{NA} \\ \pi_{NA} = \pi_{80} + \frac{1}{13} \cdot \pi_{135} + \frac{1}{8} \cdot \pi_{139} + \frac{1}{11} \cdot \pi_{445} + \frac{1}{4} \cdot \pi_{NA} \\ \pi_{80} + \pi_{135} + \pi_{139} + \pi_{445} + \pi_{NA} = 1 \end{cases} \quad (6)$$

Luego de resolverlo, obtenemos la distribución límite de los puertos atacados:

$$\pi = \begin{pmatrix} 80 & 135 & 139 & 445 & NA \\ 0.0215 & 0.2669 & 0.3435 & 0.2273 & 0.1408 \end{pmatrix} \quad (7)$$

## Anexo

Tabla correspondiente al apartado a del problema 1.

**Cantidad de sucesos hasta obtener 3 éxitos consecutivos**

Cantidad de sucesos	Frecuencia absoluta	Frecuencia relativa
3	71	0.071
4	44	0.044
5	36	0.036
6	35	0.035
7	41	0.041
8	31	0.031
9	29	0.029
10	43	0.043
11	29	0.029
12	24	0.024
13	26	0.026
14	22	0.022
15	29	0.029
16	21	0.021
17	16	0.016
18	20	0.020
19	26	0.026
20	21	0.021
21	20	0.020
22	21	0.021
23	26	0.026
24	11	0.011
25	19	0.019
26	13	0.013
27	17	0.017
28	19	0.019
29	13	0.013
30	12	0.012
31	12	0.012
32	6	0.006
33	10	0.010
34	15	0.015
35	8	0.008

36	7	0.007
37	11	0.011
38	6	0.006
39	5	0.005
40	10	0.010
41	8	0.008
42	10	0.010
43	11	0.011
44	14	0.014
45	8	0.008
46	3	0.003
47	4	0.004
48	6	0.006
49	6	0.006
50	6	0.006
51	9	0.009
52	4	0.004
53	3	0.003
54	4	0.004
55	3	0.003
56	4	0.004
57	3	0.003
58	3	0.003
59	4	0.004
60	1	0.001
61	2	0.002
63	4	0.004
64	2	0.002
65	4	0.004
66	2	0.002
67	3	0.003
68	1	0.001
69	2	0.002
70	1	0.001
71	2	0.002

72	4	0.004
74	2	0.002
75	5	0.005
77	1	0.001
78	2	0.002
79	2	0.002
80	1	0.001
81	2	0.002
82	1	0.001
83	2	0.002
84	2	0.002
85	1	0.001
86	1	0.001
87	1	0.001
90	2	0.002
91	2	0.002
96	1	0.001
97	1	0.001
100	2	0.002
111	1	0.001
125	1	0.001
130	1	0.001
Total:	1000	1