Notas XII Escuela de Probabilidad y Estadística[1]

Víctor RIVERO [2]

14 de marzo de 2014

[1]En estas notas se describen algunos resultados que fueron mencionados o utilizados en charlas de la XII Escuela de Probabilidad y Estadística, fueron extraídas de un borrador utilizado por el autor en cursos de Modelación Estocástica.

[2]Centro de Investigación en Matemáticas A.C. Calle Jalisco s/n Col. Mineral de Valenciana C.P. 36240 Guanajuato, Guanajuato, México. E–mail: rivero@cimat.mx

Capítulo 1

Caminatas aleatorias

En esta sección nos interesaremos por nuestro primer ejemplo de proceso estocástico, la caminata aleatoria. Empecemos por definir que entenderemos por un proceso estocástico a una familia de v.a. digamos $\{Z_t, t \in T\}$ definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ e indexadas por un conjunto no vacio T. En esta sección nos interesaremos por procesos estocásticos indexados por $T = \mathbb{Z}^+$ y tal que todas las v.a. toman valores en un subconjunto de $\mathbb{Z}^+$. Por supuesto se puede tambien considerar preocesos estocasticos cuyo conjunto de indices sea, $\mathbb{Z}, \mathbb{R}^+$ o cualquier otro conjunto ordenado y el espacio de valores en donde toma valores Z_r puede ser cualquier conjunto como por ejemplo $\mathbb{R}$, $\mathbb{R}^d$ etc..

Supongamos que el domingo nos encontramos con el merenguero y jugamos volados, tenemos una fortuna inicial S_0 y cada vez que jugamos un volado podemos ganarlo con probabilidad p perderlo con probabilidad $1-p$, puede ocurrrir que $p \neq 1/2$. Denotemos por X_n la suma que ganamos en el n-ésimo juego, es decir que las X_n son v. a. del tipo Bernoulli(p). Tenemos que nuestra fortuna despues de n volados está dada por

$$S_n = S_0 + \sum_{i=1}^{n}, \qquad n \geq 1,$$

y nuestra fortuna al tiempo $n+m$ estará dada por

$$S_{n+m} = S_n + \sum_{i=n+1}^{n+m} X_i, \qquad n, m \geq 1.$$

Supondremos que el resultado de los volados son independientes entre si. es decir que las $X_1, X_2, \ldots$ son independientes, esto quiere decir que para cualquier n-éada de enteros $(k_1, k_2, \ldots, k_n)$, distintos se tiene que las v. a. $X_{k_1}, X_{k_2}, \ldots, X_{k_n}$ son independientes. Es decir que la riqueza con la que contamos está dada por una suma de v.a. independientes. Este modelo puede ser muy útil para modelar muchas cosas, por ejemplo podemos usarlo para modelar la posición de una partícula que se mueve en los enteros, a cada paso esta partícula puede avanzar de un paso o retroceder de un paso con probabilidad p ó $1-p$ respectivamente, y la dirección que escoja seguir es independiente del resto de los pasos que ha dado. Por supuesto no hay ninguna razón para restringirse a v.a. X_i que solo pueden tomar el valor $+1$ o -1, igualmente podemos tomar v.a. que toman valores en todo $\mathbb{Z}^+$ y con cualquier densidad o función de distribución.

Esto nos puede ser útil para modelar fenómenos que evolucionan en el tiempo en los cuales podemos suponer que la $n+1$-ésima observación difiere de la n-ésima por una cantidad que es independiente de las observaciones anteriores pero que las variaciones siguen la misma ley de probabilidad.

Por ejemplo la posición de una partícula de polen en un fluido observado cada segundo, o la variación del indice de precios y cotizaciones de la bolsa de valores Mexicana al final de cada día, el numero de clientes atendidos en un negocio cada día, etc..

Las caminatas aleatorias se gráfican en el plano cartesiano con los puntos $(n, S_n, n \geq 0)$ uniendo los puntos vecinos mediante lineas rectas de pendiente 1 o -1. A la gráfica obtenida se le llama trayectoria. Para cada $\omega \in \Omega$ se obtiene una trayectoria.

Las caminatas aleatorias simples tienen las siguientes propiedades fundamentales.

Lema 1. *Toda caminata aleatoria simple $\{S_n, n \geq 0\}$ tiene las siguientes propiedades*

- *Homogeneidad espacial*

$$\mathbb{P}(S_n = j | S_0 = a) = \mathbb{P}(S_n = j + b | S_0 = a + b).$$

- *Homogeneidad Temporal*

$$\mathbb{P}(S_n = j | S_0 = a) = \mathbb{P}(S_{n+m} = j | S_m = a).$$

- *Propiedad de Markov*

$$\mathbb{P}(S_{n+m} = j | S_0, S_1, \ldots, S_n) = \mathbb{P}(S_{n+m} = j | S_n).$$

Demostración. Demostremos la homogeneidad espacial: se tiene que el lado izquierdo es igual a $\mathbb{P}\left(\sum_{i=1}^{n} X_i = j - a\right)$ mientras que el lado derecho es igual a $\mathbb{P}\left(\sum_{i=1}^{n} X_i = j + b - (a + b)\right)$. Ahora pasemos a la homogeneidad temporal: análogo a la propiedad anterior, es una consecuencia fácil de la independencia y del hecho que las v. a. X_i son idénticamente distribuidas que el lado derecho es igual a:

$$\frac{\mathbb{P}\left(S_0 + \sum_{i=1}^{n+m} X_i = j, S_0 + \sum_{i=1}^{m} X_i = a\right)}{\mathbb{P}\left(S_0 + \sum_{i=1}^{m} X_i = a\right)} = \mathbb{P}\left(\sum_{i=m+1}^{m+n} X_i = j - a\right) = \mathbb{P}\left(\sum_{i=1}^{n} X_i = j - a\right);$$

un calculo elemental prueba que el lado izquierdo es idéntico a esta cantidad.

La propiedad de Markov se demuestra observando que S_{n+m} es una función de S_n y de $X_{n+1}, \ldots X_{n+m}$ pues $S_{n+m} = S_n + \sum_{i=n+1}^{n+m} X_i$, y además el vector $(X_{n+1}, \ldots X_{n+m})$ es independiente de $S_0, \ldots, S_n$ ya que estas son función de $(X_1, \ldots, X_n)$. Podemos afirmar que condicionar S_{n+m} con respecto a $S_0, \ldots, S_n$ es igual a condicionar a S_{n+m} con respecto a S_n. Veamoslo, sean

$s_0, \ldots, s_n$ enteros tales que $\mathbb{P}(S_0 = s_0, \ldots, S_n = s_n)$, se tiene que

$$\frac{\mathbb{P}(S_0 = s_0, \ldots, S_n = s_n, S_{n+m} = s_{n+m})}{\mathbb{P}(S_0 = s_0, \ldots, S_n = s_n)}$$
$$\frac{\mathbb{P}(S_0 = s_0, X_1 = s_1 - s_0, \ldots, X_{n-1} = s_{n-1} - s_{n-2}, X_n = s_n - s_{n-1}, \sum_{i=n+1}^{n+m} X_i = s_{n+m} - s_n)}{\mathbb{P}(S_0 = s_0, X_1 = s_1 - s_0, \ldots, X_{n-1} = s_{n-1} - s_{n-2}, X_n = s_n - s_{n-1})}$$
$$= \mathbb{P}(\sum_{i=n+1}^{n+m} X_i = s_{n+m} - s_n)$$
$$= \frac{\mathbb{P}(\sum_{i=n+1}^{n+m} X_i = s_{n+m} - s_n, S_n = s_n)}{\mathbb{P}(S_n = s_n)}$$
$$= \mathbb{P}(S_{n+m} = s_{n+m} | S_n = s_n).$$

□

Se dirá que la probabilidad $\mathbb{P}(S_n = j | S_0 = a)$ es la probabilidad de transición del estado a al j en n pasos. Veamos como calcular las probabilidades de este tipo:

Lema 2. *Se tiene que para todo a, b enteros y $n \geq 0$*

$$\mathbb{P}(S_n = j | S_0 = a) = \begin{cases} \binom{n}{(n+b-a)/2} p^{(n+b-a)/2} q^{(n-b+a)/2} & \textit{si } (n+b-a)/2 \in \mathbb{Z} \\ 0 & \textit{en otro caso.} \end{cases}$$

Demostración. Se tiene que una trayectoria que lleva del punto $(0, a)$ al punto $(0, b)$ en n pasos tiene r pasos para arriba $(+1)$ y l pasos hacia abajo (-1), estos son tales que $r + l = n$ y $r - l = b - a$. (Pues $S_n = r(+1) + l(-1) = b - a$.) Estas ecuaciones determinan a l y r, lo que implica que $r = (n + b - a)/2$ y $l = (n - b + a)/2$. cada trayectoria que lleva de a a b en n pasos tiene probabilidad $p^r q^l$, y hay $\binom{n}{(n+b-a)/2}$ trayectorias posibles. El resultado se sigue. □

Una propiedad interesante de las caminatas aleatorias, que ha atraído no solo a probabilistas, sino también a investigadores en combinatoria, teoría de juegos física, etc.. es el hecho de que muchas probabilidades pueden ser calculadas contando trayectoria, como lo acabamos de hacer. Antes de pasar al calculo de otras probabilidades interesantes, calcularemos algunas probabilidades usando condicionamiento.

Esto es interesante en el caso en que la caminata aleatoria representa la fortuna de un jugador que apuesta y tiene probabilidad p de ganar y $1 - p = q$ de perder, y cada juego ganado le aporta 1 peso y perdido -1 peso.

Los primeros tiempos de llegada son sumamente útiles, definimos $T_j = \inf\{n \geq 0 : S_n = j\}$, para $j \in \mathbb{Z}$.

Lema 3 (Problema de la ruina)**.** *Sea h_j la probabilidad de que una caminata aleatoria que parte del estado j llegue al nivel 0 antes de llegar al nivel N, es decir $\mathbb{P}(T_0 < T_N | S_0 = j) = h_j$. Se tiene que*

$$h_j = \begin{cases} \frac{\left(\frac{q}{p}\right)^j - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N} & p \neq q, \\ 1 - \frac{j}{N} & p = q = 1/2 \end{cases}$$

Si T es el tiempo de absorción en $\{0,N\}$, $T=\inf\{n\geq 0 : S_n\in\{0,N\}\}$ entonces

$$\mathbb{E}\left(T|S_0=j\right)=\begin{cases}\frac{j}{q-p}-\frac{N}{q-p}\frac{1-\left(\frac{q}{p}\right)^j}{1-\left(\frac{q}{p}\right)^N}, & \textit{si } p\neq q\\ j(N-j), & \textit{si } p=q=1/2.\end{cases}$$

Demostración. Demostraremos primero la igualdad para $\{h_j, j\in\{0,N\}\}$. Utilizando la técnica de la transición inicial obtenemos el sistema de ecuaciones lineales $h_j = ph_{j+1}+qh_{j-1}$, si $j\in\{2,\ldots,N-1\}$ y $h_0=1$ y $h_N=0$.

Reescribiendo la segunda ecuación de este sistema se tiene que

$$h_n = ph_{n+1}+qh_{n-1} \qquad \Longleftrightarrow \qquad q\left(h_n-h_{n-1}\right)=p\left(h_{n+1}-h_n\right) \qquad n\geq 1.$$

En el caso en que $q=p$ tenemos que la función h_n tiene una pendiente constante $c=h_{n+1}-h_n$ y por lo tanto

$$h_n = 1+\sum_{j=1}^{n} h_j-h_{j-1} = 1+nc, \qquad 1\leq n\leq N,$$

y puesto que $h_N=0$ se puede concluir que $c=-1/N$.

Para estudiar el caso en el cual $q\neq p$, usaremos una sucesión de reales $\{x_n, n\in\mathbb{N}\}$ definida por $x_0\in\mathbb{R}$, y $x_n=h_n-h_{n-1}$, $1\leq n\leq N$. De el cálculo anterior se deduce que la sucesión $\{x_n, 1\leq n\leq N\}$ satisface la relación $x_{n+1}=\frac{q}{p}x_n$, $1\leq n\leq N$, es decir que

$$x_n=\left(\frac{q}{p}\right)^n x_0, \qquad 0\leq n\leq N.$$

Usando esto se tiene que la sucesión h_n esta dada por

$$\begin{aligned}h_n &= h_0+\sum_{j=1}^{n}(h_j-h_{j-1})\\ &= h_0+x_0\sum_{j=1}^{n}\left(\frac{q}{p}\right)^j\\ &= h_0+x_0\left(\frac{q}{p}\right)\frac{1-\left(\frac{q}{p}\right)^n}{1-\left(\frac{q}{p}\right)}.\end{aligned}$$

Sabemos que $h_N=0$ y que $h_0=1$, esto nos permite determinar a x_0 de hecho se tiene que

$$x_0=-\frac{p\left(1-\frac{q}{p}\right)}{q\left(1-\left(\frac{q}{p}\right)^N\right)},$$

de donde que

$$h_n = 1 - \frac{p\left(1-\frac{q}{p}\right)}{q\left(1-\left(\frac{q}{p}\right)^N\right)}\left(\frac{q}{p}\right)\frac{1-\left(\dfrac{q}{p}\right)^n}{1-\left(\dfrac{q}{p}\right)} = \frac{\left(\frac{q}{p}\right)^n - \left(\frac{q}{p}\right)^N}{1-\left(\frac{q}{p}\right)^N}.$$

En la demostración de la identidad para el primer tiempo de absorción en $\{0, N\}$ se utiliza el método de la transición inicial para demostrar establecer que el vector $\{s_j = \mathbb{E}(T|X_0 = j), j \in \{0, N\}\}$, resuelve el sistema

$$s_0 = 0 = s_N, \quad s_j = 1 + ps_{j+1} + qs_{j-1}, \quad j \in \{1, \dots, N-1\}.$$

usando el hecho de que $p + q = 1$ se llega a que el sistema anterior es equivalente a

$$s_0 = 0 = s_N, \quad (s_{j+1} - s_j) = -\frac{1}{p} + \frac{q}{p}(s_j - s_{j-1}), \quad j \in \{1, \dots, N-1\}.$$

En el caso en que $p = q = 1/2$ el sistema anterior se escribe

$$s_0 = 0 = s_N, \quad (s_{j+1} - s_j) = -2 + (s_j - s_{j-1}) = -2j + (s_1 - s_0), \quad j \in \{1, \dots, N-1\}.$$

Utilizando la relación

$$s_k - s_0 = \sum_{j=1}^{k} s_j - s_{j-1}$$

para $k = N$ vemos que

$$0 = s_N - s_0 = \sum_{j=1}^{N} (-2(j-1) + (s_1 - s_0)) = -N(N-1) + N(s_1 - s_0),$$

y por lo tanto que $s_1 = N - 1$. Remplazando éste valor en la igualdad anterior y haciendo un poquito de álgebra llegamos a

$$s_k = s_k - s_0 = -2\frac{(k-1)k}{2} + k(N-1) = k(N-k),$$

para todo $k \in \{0, \dots, N\}$. En el caso en que $p \neq q$, denotamos $x_j = s_j - s_{j-1}$, $j \in \{1, \dots, N\}$. Por lo hecho anteriormente sabemos se cumple la igualdad

$$x_{j+1} = -\frac{1}{p} + \frac{q}{p}x_j, j \in \{1, \dots, N-1\}.$$

Usando un método estandar sobre resolución de ecuaciones en recurrencia sabemos que necesariamente

$$x_j = c\left(\frac{q}{p}\right)^j + \frac{1}{q-p}, j \in \{1, \dots, N-1\}.$$

Procediendo como en el caso anterior llegamos a $c = \frac{-N}{q\left(1-\left(\frac{q}{p}\right)^N\right)}$ y por lo tanto

$$s_k = \sum_{j=1}^{k} x_j = \frac{k}{q-p} - \frac{N}{q-p}\frac{1-(\frac{q}{p})^k}{1-(\frac{q}{p})^N}, \qquad k \in 0, 1, \dots, N.$$

□

Corolario 1. *Se tiene que para todo $j \in \mathbb{N}$*

$$\mathbb{P}(T_0 < \infty | S_0 = j) = \begin{cases} 1 & si\ q \leq p \\ \left(\frac{q}{p}\right)^j, & si\ q > p. \end{cases}$$

Demostración. Para $n \geq 1$ sea A_n el evento $A_n = \{T_0 < T_n\}$. Observemos que $A_n \subseteq A_{n+1}$ puesto que $T_n \leq T_{n+1}$, para todo n. Es decir que la sucesión A_n es una sucesión de eventos creciente y además

$$\{T_0 < \infty\} = \cup_{n\geq 1}\{T_0 < T_n\},$$

y por la continuidad por debajo de la probabilidad se tiene que

$$\lim_{n\to\infty} \mathbb{P}(A_n | S_0 = j) = \mathbb{P}(T_0 < \infty | S_0 = j).$$

□

Sean $a, b \in \mathbb{Z}$ y $n \in \mathbb{N}$. Denotaremos por $N_{n(a,b)}$ al número de trayectorias que van de a a b en n pasos y a $N_n^0(a,b)$ a aquellas que van de a a b en n pasos pasando por 0 por lo menos una vez. Un resultado fundamental de la teoría de caminatas aleatorias simples es el principio de reflexión.

Teorema 1. *Para $a, b > 0$ se tiene que*

$$N_n^0(a,b) = N_n(-a,b).$$

Demostración. Hacer un dibujo para ver que cada trayectoria que lleva de $(0,-a)$ a (b,n) cruza el eje x por lo menos una vez, denotemos por $(k,0)$ el punto en el cual esto ocurre por la primera vez. Reflejando el segmento de la trayectoria anterior a $(k,0)$ se obtiene una trayectoria lleva de $(0,a)$ a (b,n) y que intercepta al eje x por lo menos una vez. Haciendo lo mismo en el sentido contrario □

Lema 4. *Para todo $a, b \in \mathbb{Z}$ se tiene que*

$$N_n(a,b) = \binom{n}{\frac{1}{2}(n+b-a)}.$$

Demostración. La prueba de este resultado es como en la prueba del teorema anterior. □

Tenemos el siguiente teorema consecuencia importante del lema anterior.

Teorema 2. *Teorema de las votaciones (Ballot Theorem) Si $b > 0$, entonces el número de trayectorias que llevan de $(0,0)$ a (n,b) y que no visitan el eje x después del primer paso es igual a*

$$\frac{b}{n} N_n(0,b).$$

Demostración. Observemos que el primer paso de dichas trayectorias lleva $(1,1)$, por lo tanto el número que buscamos calcular está dado por

$$\begin{aligned}
N_{n-1}(1,b) - N_{n-1}^{0}(1,b) &= N_{n-1}(1,b) - N_{n-1}(-1,b) \\
&= \frac{(n-1)!}{\left(\frac{n-b}{2}\right)!\left(\frac{n+b-2}{2}\right)!} - \frac{(n-1)!}{\left(\frac{n-b-2}{2}\right)!\left(\frac{n+b}{2}\right)!} \\
&= \frac{(n-1)!}{\left(\frac{n-b}{2}\right)!\left(\frac{n+b}{2}\right)!}\left(\frac{n+b}{2} - \frac{n-b}{2}\right) \\
&= \frac{b}{n} N_n(0,b).
\end{aligned}$$

□

¿Por que se le llama a este resultado el teorema de las votaciones? Supongamos que tenemos dos candidatos F y A, y que F gana α votos y A gana β votos, y que $\alpha > \beta$. ¿Cual es la probabilidad de que F lleve la ventaja durante toda la votación? Supongamos que $X_i = 1$ si el i-ésimo individuo vota por F y vale -1 si vota por A. Supongamos que cualquier combinación de votos es igualmente probable, es decir que cada una tiene probabilidad $\binom{\alpha+\beta}{\alpha}$. La trayectoria que deben seguir las votaciones para que F tenga la mayoria durante toda la jornada de votaciones va del punto $(0,0)$ al punto $(\alpha+\beta, \alpha-\beta)$. Por lo tanto la probabilidad buscada está dada por

$$\frac{\alpha-\beta}{\alpha+\beta} N_{\alpha+\beta}(0, \alpha-\beta) \frac{1}{\binom{\alpha+\beta}{\alpha}} = \frac{\alpha-\beta}{\alpha+\beta}.$$

Veamos ahora una aplicación del teorema de las votaciones a las caminatas aleatorias.

Teorema 3. *Supongamos que $S_0 = 0$, entonces para $n \geq 1$*

$$\mathbb{P}(S_1 S_2 \cdots S_n \neq 0, S_n = b) = \frac{b}{n}\, \mathbb{P}(S_n = b),$$

y por lo tanto

$$\mathbb{P}(S_1 S_2 \cdots S_n \neq 0) = \frac{\mathbb{E}(|S_n|)}{n}.$$

Demostración. Supongamos que $S_0 = 0$ y que $b > 0$. el evento al que le queremos calcular probabilidad es aquel en que la caminata aleatoria parte de $(0,0)$ y no visita el eje x en el intervalo $[1,n]$. una tal trayectoria tiene r pasos para arriba y l para abajo, en consecuencia $r = (n+b)/2$ y $l = (n-b)/2$. Cada una de estas trayectoria tiene probabilidad $p^{(n+b)/2} q^{(n-b)/2}$ y por el teorema de las votaciones hay $\frac{b}{n} N_n(0,b)$ trayectorias de este tipo. La probabilidad que queremos calcular es igual a

$$\mathbb{P}(S_1 S_2 \cdots S_n \neq 0, S_n = b) = \frac{b}{n}\binom{n}{\frac{n+b}{2}} p^{(n+b)/2} q^{(n-b)/2} = \frac{b}{n}\, \mathbb{P}(S_n = b).$$

Un argumento similar se vale para $b < 0$. □

Otro resultado particularmente interesante concierne el valor máximo que alcanza la caminata aleatoria. Sea $M_n = \text{máx}\{S_j, 1 \leq j \leq n\}$, para $n \geq 0$ y supongamos que $S_0 = 0$. En particular se tiene que $M_n \geq 0$.

Tenemos el siguiente teorema

Teorema 4. *Supongamos que $S_0 = 0$. Entonces para $r \geq 1$,*

$$\mathbb{P}(M_n \geq r, S_n = b) = \begin{cases} \mathbb{P}(S_n = b) & si\ b \geq r \\ \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r - b) & si\ b < r. \end{cases}$$

Una consecuencia del teorema anterior es la formula

$$\mathbb{P}(M_n \geq r) = \mathbb{P}(S_n \geq r) + \sum_{b=-\infty}^{r-1} \left(\frac{q}{p}\right)^{r-b} \mathbb{P}\left(S_n = 2r - b\right)$$

$$\mathbb{P}(S_n = r) \sum_{l=r+1}^{\infty} \mathbb{P}(S_n = l) \left(1 + (q/p)^{l-r}\right)$$

Demostración. Supongamos que $r \geq 1$ y que $b < r$. Sea $N_n^r(0,b)$ el número de trayectorias que van de $(0,0)$ a (n,b) pasando por r en n pasos. Una de estas trayectorias visita el nivel r por lo menos una vez, denotemos por i_r el más pequeño de estos instantes, reflejando la trayectoria entre i_r y n en la recta r se obtiene una trayectoria que va de $(0,0)$ a $(n, 2r - b)$. A una trayectoria de esta forma le aplicamos la transformación inversa y obtenemos una trayectoria que va de $(0,0)$ a (n,b) pasando por el nivel r, de longitud n. Por lo tanto podemos afirmar que

$$N_n^r(0,b) = N_n(0, 2r - b).$$

Además cada una de estas trayectorias tiene probabilidad

$$p^{(n+b)/2} q^{(n-b)/2},$$

por lo que podemos concluir que

$$\begin{aligned} \mathbb{P}(M_n \geq r, S_n = b) &= N_n^r(0,b) p^{(n+b)/2} q^{(n-b)/2} \\ &= N_n(0, 2r - b) p^{(n+b)/2} q^{(n-b)/2} \\ &= \left(\frac{q}{p}\right)^{r-b} N_n(0, 2r - b) p^{(n+2r-b)/2} q^{(n-2r+b)/2} \\ &= \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r - b) \end{aligned}$$

□

Observemos que en particular si la caminata aleatoria es simetrica

$$\mathbb{P}(M_n \geq r, S_n = b) = \mathbb{P}(S_n = 2r - b).$$

Cual es la probabilidad de que una caminata aleatoria alcance un máximo en un instante n dado? Más precisamente, cual es la probabilida de que una caminata aleatoria que parte de 0 alcance un nivel b por la primera vez al tiempo n Denotemos por $f_b(n)$ a esta probabilidad. Tenemos el siguiente teorema.

Teorema 5. *La probabilidad de que una caminata aleatoria simple alcance el nivel b por la primera vez al tiempo n habiendo empezado de 0 está dada por*

$$f_b(n) = \frac{|b|}{n}\,\mathbb{P}(S_n = b), \quad n \geq 1.$$

Demostración.

$$\begin{aligned} f_b(n) &= \mathbb{P}(M_{n-1} = S_{n-1} = b-1, S_n = b) \\ &= \mathbb{P}(, S_n = b | M_{n-1} = S_{n-1} = b-1)\,\mathbb{P}(M_{n-1} = S_{n-1} = b-1) \\ &= p\,(\mathbb{P}(M_{n-1} \geq b-1, S_{n-1} = b-1) - \mathbb{P}(M_{n-1} \geq b, = S_{n-1} = b-1)) \\ &= p\left(\mathbb{P}(S_{n-1} = b-1) - \frac{q}{p}\,\mathbb{P}(S_{n-1} = b+1)\right) \\ &= \frac{b}{n}\,\mathbb{P}(S_n = b). \end{aligned}$$

El mismo razonamiento es valido si $b < 0$. □

Ejercicio 1. Sea S_n una caminata aleatoria simple con $S_0 = 0$. Sea $r \neq 0$, y definimos a la variable aleatoria V_r el número de visitas al estado r antes de que la cadena aregrese a su punto de partida 0. Demostrar que $\mathbb{E}(V_r) = 1$. Dar un criterio para determinar si el numero esperado de vistas a 0 es finito o infinito.

Demostración. Solución Observemos que $V_r = \sum_{n\geq 1} I_{A_n}$ donde I_n vale 1 si al tiempo n la caminata aleatoria visita el estado r y no a vistado el estado 0 hasta ese instante y vale 0 en otro caso. Se tiene que

$$\mathbb{E}(I_{A_n}) = \mathbb{P}(A_n) = \mathbb{P}(S_n = r, S_1 \cdots S_n \neq 0) = \frac{|r|}{n}\,\mathbb{P}(S_n = 0) = f_r(n).$$

De donde que

$$\mathbb{E}(V_r) = \mathbb{E}(\sum_{n\geq 1} I_{A_n}) = \sum_{n\geq 1} \mathbb{P}(A_n) = \sum_{n\geq 1} f_r(n) = 1.$$

Lo más importante de este resultado es su interpretación: Supongamos que se lanza un volado sucesivamente y cada vez que la diferencia entre el número de soles y aguilas es igual a r obtenemos una recompensa de m pesos, el juego se detiene cuando el el número de soles es igual al de aguilas. El resultado nos dice que el valor esperado de la ganacia es igual a m, lo cual es independiente de r. □

1.0.1. Caminatas aleatorias reflejadas en el máximo

Sean $\{Y_n, n \geq 0\}$ variables aleatorias independientes e idénticamente distribuidas que toman valores en $\mathbb{Z}$ y cuya ley común denotaremos por $\mathbb{P}(Y_1 = z) = \mu(z)$, $z \in \mathbb{Z}$. Asociada a ésta sucesión de v.a. construimos la caminata aleatoria asociada $S_0 = Y_0$, $S_{n+1} = Y_{n+1} + S_n$ para $n \geq 0$, y su máximo presente $M_0 = 0$ y $M_{n+1} = \max\{M_n, S_{n+1}\}$ para $n \geq 0$. Demostrar que el proceso Z definido por

$$Z_n = M_n - S_n, \qquad n \geq 0,$$

es una cadena de Markov, describir su espacio de estados, E, y su matriz de transición.

(a) Suponer que $\mu(1) = p$ y $\mu(-1) = q$, con $0 < p < 1$ y $1 - p = q$, dar la matriz de transición en este caso. Demostrar que la cadena es irreducible. Haga un esbozo de las trayectorias de S_n, M_n y Z_n, sobre una misma gráfica.

(b) Bajo las hipótesis de (a) calcular $\mathbb{P}(H_0 < \infty | Z_0 = x)$ para todo $x \in E$, donde $H_0 = \inf\{k \geq 0 : Z_k = 0\}$.

(c) Bajo las hipótesis de (a) calcular $\mathbb{P}(T_0 < \infty | Z_0 = x)$ para todo $x \geq 1$, con la notación usual $T_0 = \inf\{k \geq 1 : Z_k = 0\}$.

(d) Bajo las hipótesis de (a) utilizar $\mathbb{P}(T_0 < \infty | Z_0 = 1)$ para obtener un criterio en términos de p para determinar cuando la cadena es recurrente o transitoria. Observe que $p < q$ si y solamente si $2p < 1$.

(e) Bajo las hipótesis de (a) obtenga criterios sobre recurrencia y transitoriedad para la caminata aleatoria $\{S_n, n \geq 0\}$. ¿Encuentra alguna relación entre estos y los obtenidos en (d)?

Prueba de (4). Veamos que el proceso estocástico$\{Z_n, n \geq 0\}$ lo podemos expresar mediante una recurrencia aleatoria: para todo $n \geq 0$,

$$\begin{aligned} Z_{n+1} &= M_{n+1} - S_{n+1} = \max\{M_n, S_{n+1}\} - S_n - Y_{n+1} \\ &= \max\{M_n - S_n - Y_{n+1}, 0\} = \max\{Z_n - Y_{n+1}, 0\} = F(Z_n, Y_{n+1}), \end{aligned}$$

donde $F : \mathbb{Z}^2 \to \mathbb{Z}$ es la función dada por $F(z, y) = \max\{z - y, 0\}$ para $(x, y) \in \mathbb{Z}^2$. Usando un resultado visto en clase podemos afirmar que Z_n es una cadena de Markov que toma valores en $0, 1, 2, \ldots$ Sus probabilidades de transición estan dadas para $i \in \{0, 1, 2, \ldots\}$

$$\begin{aligned} \mathbb{P}\left(Z_{n+1} = j | Z_n = i\right) &= \mathbb{P}(F(i, Y_{n+1}) = j) = \mathbb{P}(F(i, Y_{n+1}) = j) \\ &= \begin{cases} \sum_{k \geq i} \mathbb{P}(Y_{n+1} = k), & \text{si } j = 0, \\ \mathbb{P}(Y_{n+1} = i - j) & \text{si } j = 1, 2, \ldots \end{cases} \\ &= \begin{cases} \sum_{k \geq i} \mu(k), & \text{si } j = 0, \\ \mu(i - j) & \text{si } j = 1, 2, \ldots \end{cases} \end{aligned}$$

□

Solución de (4a). Bajo las hipótesis del ejercicio se tiene que las probabilidades de transición son

$$P_{0,0} = p, \quad P_{0,1} = q, \quad P_{i,j} = \begin{cases} p, & \text{si } j = i-1, \\ q, & \text{si } j = i+1, \\ 0, & \text{en otro caso.} \end{cases} \quad \text{si } i \geq 1.$$

En este caso vemos que la cadena de Markov Z es una cadena de nacimiento y muerte. Por resultados que se estudian en cursos básicos de cadenas de Markov vemos que es una cadena irreducible. □

Solución de (4b). Se resolverá el caso en que $p \neq q$, y se deja al lector como ejercicio estudiar el caso $p = q$. Por resultados que se estudian en cursos básicos de cadenas de Markov se puede afirmar que las probabilidades buscadas están dadas por $\mathbb{P}_0(H_0 < H_{n+1}) = 1$, y

$$\mathbb{P}_x(H_0 < H_{n+1}) = \frac{\displaystyle\sum_{y=x}^{n} \gamma_y}{\displaystyle\sum_{y=0}^{n} \gamma_y} = 1 - \frac{\displaystyle\sum_{y=0}^{x-1} \gamma_y}{\displaystyle\sum_{y=0}^{n} \gamma_y},$$

con $\gamma_y = \prod_{k=1}^{y} \left(\frac{p}{q}\right) = \left(\frac{p}{q}\right)^y$, para $y \geq 1$ y $\gamma_0 = 1$. Se tiene que

$$\mathbb{P}(H_0 < H_{n+1} | X_0 = x) = \begin{cases} 1, & \text{si } x = 0, \\ 1 - \dfrac{1 - \left(\dfrac{p}{q}\right)^x}{1 - \left(\dfrac{p}{q}\right)^{n+1}}, & \text{si } 1 \leq x \leq n, \end{cases}$$

Haciendo tender n a infinito obtenemos que

$$\mathbb{P}(H_0 < \infty | X_0 = x) = \lim_{n\to\infty} \mathbb{P}(H_0 < H_{n+1} | X_0 = x) = \begin{cases} 1, & \text{si } x = 0, \\ \left(\frac{p}{q}\right)^x, & \text{si } 1 \leq x < \infty, \text{ y } p < q, \\ 1, & \text{si } 1 \leq x < \infty, \text{ si } p > q, \end{cases}$$

la primera ecuación de la igualdad anterior se justifica con el mismo argumento que se utilizo en clase para demostrar el segundo corolario sobre el calculo de probabilidades de primer llegada para cadenas de nacimiento y muerte. □

Solución de (4c). Puesto que la cadena solo puede avanzar de 1 ó -1 si parte de un estado $x \geq 1$, es claro que para llegar a 0 se necesita al menos un paso, entonces $T_0 = H_0$ si $X_0 = x$ para $x \geq 1$. Por lo tanto tenemos que

$$\mathbb{P}(T_0 < \infty | X_0 = x) = \mathbb{P}(H_0 < \infty | X_0 = x), \qquad x \geq 1.$$

□

Solución de (4d). Es consecuencia de los incisos anteriores que

$$\mathbb{P}(T_0 < \infty | X_0 = 1) = \begin{cases} 1, & \text{si } p > q, \\ \frac{p}{q}, & \text{si } p < q. \end{cases}$$

Para determinar si la cadena es recurrente nos basta saber si el estado 0 es recurrente puesto que la cadena es irreducible. Calculemos $\mathbb{P}(T_0 < \infty | X_0 = 0)$. En efecto, haciendo un estudio de la primera transición obtenemos

$$\begin{aligned} \mathbb{P}(T_0 < \infty | X_0 = 0) &= p\,\mathbb{P}(T_0 < \infty | X_1 = 0, X_0 = 0) + q\,\mathbb{P}(T_0 < \infty | X_1 = 1, X_0 = 0) \\ &= p + q\,\mathbb{P}(T_0 < \infty | X_0 = 1) \\ &= \begin{cases} p + q\left(\frac{p}{q}\right), & \text{si } p < q, \\ p + q, & \text{si } p > q, \end{cases} \\ &= \begin{cases} 2p, & \text{si } p < q, \\ 1, & \text{si } p > q, \end{cases} \end{aligned}$$

Observe que $p < q$ si y solamente si $p < 1/2$. Tenemos que la cadena Z es recurrente si $p > q$ y transitoria si $p < q$. □

Prueba de (4e). Se puede ver que la marcha aleatoria como la hemos definido aquí es recurrente si y solamente si $p = q$ y transitoria si $p \neq q$. Este ejercicio nos dice que si $q > p$ entonces la caminata aleatoria es transitoria y además tiene una probabilidad positiva de, a partir de un cierto tiempo aleatorio, no alcanzar un nuevo máximo; en este caso se dice que la cadena tiende a $-\infty$. Esto refuerza la idea de que hay mayor probabilidad de ir hacia los valores negativos que hacia los positivos ya que $q > p$. En el caso en que $p > q$, la caminata aleatoria es transitoria pero siempre va a alcanzar un nuevo máximo. De hecho aplicando el razonamiento anterior a la caminata aleatoria $\widehat{Z}_n = -Z_n$, $n \geq 0$, veremos que se tiene una probabilidad positiva de, a partir de un cierto instante aleatorio, la caminata Z no alcance un nuevo infimo; en este caso se dice que la caminata aleatoria Z tiende hacia ∞. De nuevo esto refuerza la idea de que se tiene mayor probabilidad de ir hacia los valores positivos que hacia los negativos cuando $p > q$. □

1.1. Caminatas aleatorias y procesos de Galton Watson

Los procesos de ramificación permiten modelar la evolución de poblaciones de humanos, células, neutrones en un reactor, o una epidemia en una población. Los primeros en estudiar esta clase de procesos fueron Galton y Watson, sus motivaciones era saber si los nombres de la realeza Inglesa desaparecerían o no en un momento dado.

Supongamos que una población evoluciona, en generaciones, y sea $\{Z_n, n \geq 0\}$ una sucesión de variables independientes tal que Z_n denota el número de individuos de la n-ésima generación de la población. Cada miembro de la población muere y da nacimiento a una familia (se ramifica), posiblemente vacía, de miembros de la n+1-ésima generación; el tamaño de la familia es una variable aleatoria. Supondremos que los tamaños de las familias satisfacen los siguientes supuestos:

1. los tamaños de las familias de cada individuo del proceso forman una familia de variables aleatorias independientes,

2. todos los tamaõs de las familias tienen la misma dsitribución de probabilidad, digamos f, y denotaremos por G, su función generadora:

$$G(s) = \sum_{n=0}^{\infty} s^n f(n), \qquad |s| < R,$$

donde R es el radio de convergencia de la función generadora G.

Sean $Y_k, k \geq 0$ v.a.i.i.d. y con ley común f. Observemos que

$$Z_{n+1} = Y_1 + \cdots + Y_{Z_n},$$

puesto que el i-ésimo individuo de la $n$ésima generación, da nacimiento a Y_i nuevos individuos. Además, por los supuestos del modelo sabemos que los Y_i, son independientes de Z_n. Se sigue de esto que $\{Z_n, n \geq 0\}$ es una cadena de Markov y su matriz de transición está dada por

$$\mathbf{P}_{0,k} = 1_{\{k=0\}}, \qquad \mathbf{P}_{i,k} = f^{*i}(k), \qquad i, k \geq 0,$$

con f^{*i} la convolución de f consigo misma i-veces, esto se debe a que

$$f^{*i}(k) = \mathbb{P}(Y_1 + \cdots + Y_i = k);$$

y la suma de dos variables aleatorias independientes sigue la ley determinada por la convolución de sus leyes. (Recordemos que la convolución de dos sucesiónes de reales $\{a_n, n \geq 0\}$, $\{b_n, n \geq 0\}$, se define por

$$c_n = a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \cdots + a_n b_0, \qquad n \geq 0.)$$

Recordemos los siguientes resultados sobre funciones generadoras: Sea Y una variable aleatoria Y que toma valores en $\{0, 1, 2, \ldots\}$, y $G_Y : \mathbb{R} \to \mathbb{R}$ la función generadora de Y,

$$G_Y(s) = \mathbb{E}(s^Y) = \sum_{n=0}^{\infty} s^n \, \mathbb{P}(Y = n),$$

para aquellos s para los cuales $\mathbb{E}(s^Y)$ es finita.

Convergencia Existe un $R > 0$ tal que la suma que define a G, es absolutamente convergente si $|s| < R$, y divergente si $|s| > R$. La suma es uniformemente convergente sobre conjuntos de la forma $\{s : |s| < R'\}$, con $R' < R$.

Derivabilidad G_Y puede ser derivada o integrada termino a termino cualquier número de veces, dentro del radio de convergencia, i.e. siempre que $|s| < R$.

Unicidad Si $G_Y(s) = G_Y(s)$, para todo $s < R' \leq R$, entonces $\mathbb{P}(Y = n) = \mathbb{P}(Y = n)$, para todo $n \geq 0$. Además,

$$\mathbb{P}(Y = n) = \frac{1}{n!} G_Y^{(n)}(0), \qquad n \geq 0.$$

Momentos $E(Y) = G'(1)$, y para $n \geq 2$,

$$G^{(n)}(1) = \mathbb{E}(Y(Y-1)(Y-2)\cdots Y-n+1).$$

En el caso en que el radio de convergencia sea $R = 1$, las expresiones anteriores se toman como

$$\lim_{s\uparrow 1} G^{(n)}(s).$$

En particular, se tiene que la varianza de Y esta dada por

$$Var(Y) = \mathbb{E}\left((Y - \mathbb{E}(Y))\right) = G^{(2)}(1) + G'(1) - \left(G'(1)\right)^2.$$

En el siguiente resultado veremos que una herramienta importante en el estudio de los procesos de ramificación es la función generadora de f.

En el siguiente resultado veremos que una herramienta importante en el estudio de los procesos de ramificación es la función generadora de f.

Teorema 6. *Para $n \geq 0$, sea $G_n(s)$ la función generadora de Z_n. Se tiene que G_n es la n-ésima composición de G consigo misma.*

Demostración. Condicionando con respecto al valor de Z_n se obtiene la siguiente expresión para G_{n+1},

$$\begin{aligned}
G_{n+1}(s) &= \mathbb{E}(s^{Z_{n+1}}) \\
&= \mathbb{E}\left(\left(s^{Z_{n+1}} | Z_n\right)\right) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(Z_n = k)\, \mathbb{E}(s^{X_1+\cdots+X_{Z_n}} | Z_n = k) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(Z_n = k)\, \mathbb{E}(s^{X_1+\cdots+X_k}) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(Z_n = k)\, \left(G(s)\right)^k \\
&= G_n(G(s)).
\end{aligned}$$

El resultado se obtiene mediante un razonamiento inductivo. □

Veamos una consecuencia de estos hechos en el siguiente Lema, en el cual suponemos que $Z_0 = 1$.

Lema 5. *Con la notación $\mu = \mathbb{E}(X_1)$, $Var(X_1) = \sigma^2$, se tiene que*

$$\mathbb{E}(Z_n) = \mu^n, \qquad n \geq 1,$$

y

$$Var(Z_n) = \begin{cases} n\sigma^2 & si \quad \mu = 1, \\ \sigma^2 \mu^{n-1} \frac{\mu^n - 1}{\mu - 1} & si \quad \mu \neq 1. \end{cases}$$

Demostración. Para probar la primera afirmación basta con derivar $G_n(s) = G(G_n(s))$, en $s = 1$ para ver que

$$G_n'(1) = \mathbb{E}(Z_n) = \mu\,\mathbb{E}(Z_{n-1}) = \mu G_{n-1}'(1),$$

la afirmación se sigue de esto por iteración. Diferenciando dos veces vemos que

$$G_n^{(2)}(1) = G^{(2)}(1)(G_{n-1}'(1))^2 + G'(1)G_{n-1}^{(2)}(1),$$

y usando la formular para la varianza que acabamos de mencionar arriba, se prueba el segundo resultado. En efecto, si $\mu = 1$, tenemos que

$$G^{(2)}(1) = \sigma^2,$$

y

$$G_n^{(2)}(1) = \sigma^2 + G_{n-1}^{(2)}(1),$$

y la solución a esta ecuación en recurrencias es

$$G_n^{(2)}(1) = \sigma^2 n, \qquad n \geq 0.$$

ya que en la primera generación hay solamente un individuo,

$$G_1^{(2)}(1) = G^{(2)}(1) = \sigma^2.$$

En el caso, $\mu \neq 1$, se tiene que

$$\begin{aligned}
G_n^{(2)}(1) &= G^{(2)}(1)(G_{n-1}'(1))^2 + G'(1)G_{n-1}^{(2)}(1) \\
&= G^{(2)}(1)\mu^{2n-2} + \mu G_{n-1}^{(2)}(1) \\
&= G^{(2)}(1)\mu^{2n-2} + \mu\left(G^{(2)}(1)\mu^{2n-4} + \mu G_{n-2}^{(2)}(1)\right) \\
&= G^{(2)}(1)\left(\mu^{2n-2} + \mu^{2n-3}\right) + \mu^2 G_{n-2}^{(2)}(1) \\
&= G^{(2)}(1)\left(\mu^{2n-2} + \mu^{2n-3} + \mu^{2n-4}\right) + \mu^3 G_{n-3}^{(2)}(1) \\
&\vdots \\
&= G^{(2)}(1)\left(\mu^{2n-2} + \mu^{2n-3} + \cdots + \mu^{2n-(k+1)}\right) + \mu^k G_{n-k}^{(2)}(1) \\
&\vdots \\
&= G^{(2)}(1)\left(\mu^{2n-2} + \mu^{2n-3} + \cdots + \mu^{n-1)}\right) \\
&= G^{(2)}(1)\mu^{n-1}\frac{\mu^n - 1}{\mu - 1}.
\end{aligned}$$

Podemos concluir que

$$\begin{aligned}
Var(Z_n) &= \sigma^2\mu^{n-1}\frac{\mu^n - 1}{\mu - 1} - (\mu - \mu^2)\mu^{n-1}\frac{\mu^n - 1}{\mu - 1} + G_n' - (G_n')^2 \\
&= \sigma^2\mu^{n-1}\frac{\mu^n - 1}{\mu - 1} + \mu^{2n} - \mu^n + \mu^n - \mu^{2n}.
\end{aligned}$$

□

Recordemos algunas otras propiedades de la función generadora de f. Supondremos que $f(0)+f(1)<1$, para evitar trivialidades.

1. *G es estrictamente convexa y creciente en* $[0,1]$. Esto es una consecuencia inmediata del hecho que G es una serie de potencias.

2. $G(0)=\mathbb{P}(Y=0)$, $G_n(0)=\mathbb{P}(Z_n=0)$, $G(1)=1$.

3. *Si $0<\mu\leq 1$ entonces $G(t)>t$, para todo $t\in$* $[0,1[$. Para probarlo, definimos una función $g(t)=G(t)-t$, esta función satisface que $g(0)=G(0)$, $g(1)=0$ y es decreciente puesto que su derivada $g'(t)=G'(t)-1$ es negativa, y esto se debe al hecho que G' es estrictamente creciente y $G'(1)=\mu\leq 1$. Entonces, $g(t)>0$, para $0\leq t<1$. En particular, la ecuación $G(t)=t$, no tiene raíces en $]0,1[$.

4. *Si $Var(Y)=0$, entonces $G(s)=s$.*

5. *Si $\mu>1$, entonces la ecuación $G(t)=t$ tiene una única sol ución en* $[0,1[$. Esto implica que $\lim_{t\uparrow t} G'(t)=G'(1)=\mu>1$. Por continuidad existe un $t_0<1$, tal que $G'(t)>1$ para todo $t_0<t\leq 1$, por el teorema del valor intermedio vemos que

$$\frac{G(1)-G(t_0)}{1-t_0}=\frac{1-G(t_0)}{1-t_0}=G'(\widetilde{t_0})>1, \qquad \text{para algún } \widetilde{t_0}\in]t_0,1[.$$

De donde que $g(t_0)=G(t_0)-t_0<0$, y puesto que g es continua y $g(0)=\mathbb{P}(Y=0)$, podemos afirmar que existe un $0<\eta<1$ donde $g(\eta)=0$. Por la convexidad estricta de G es claro que g no puede tener ninguna otra raíz en $]\eta,1[$, ni en $]0,\eta[$.

Sea η la raíz más pequeña a la ecuación $G(t)=t$, en $[0,1]$. Los hechos anteriores implican que una tal solución existe, y además: *si $\mu\leq 1$, entonces $\eta=1$; si $\mu>1$, entonces $\eta<1$.*

Ejercicio 2 (Un ejemplo importante)**.** Supongamos que el tamaño de las familias se distribuye según una ley de la forma, geométrica con parámetro q,

$$\mathbb{P}(Y=k)=qp^k=f(k), \qquad k\geq 0, \qquad \text{para algún } p\in]0,1[.$$

Es fácil de calcular la función generadora G,

$$G(s)=\sum_{n=0}^{\infty}qs^np^n=\frac{q}{1-ps}, \qquad |s|<p^{-1}.$$

La media vale $\mu=\frac{p}{q}$. Se puede verificar usando un argumento de inducción que la n-ésima composición de G consigo misma puede ser escrita como

$$G_n(s)=\begin{cases}\dfrac{n-(n-1)s}{n+1-ns} & \text{si } p=q=1/2,\\ \dfrac{q\left[p^n-q^n-ps(p^{n-1}-q^{n-1})\right]}{p^{n+1}-q^{n+1}-ps(p^n-q^n)} & \text{si } p\neq q.\end{cases}$$

Una pregunta natural en el contexto de los procesos de ramificación y en particular para el ejemplo que nos ocupa actualmente, es: ¿Cual es el comportamiento asintótico del proceso después de varias generaciones? La población se extingue en algun momento dado? o la población nunca alcanza el tamaño cero? Usaremos la forma explícita de G_n para responder a esta pregunta. Recordemos que

$$\mathbb{P}(Z_n = 0) = G_n(0) = \begin{cases} \dfrac{n}{n+1} & \text{si } p = q = 1/2, \\ \dfrac{q\,[p^n - q^n)]}{p^{n+1} - q^{n+1})} & \text{si } p \neq q. \end{cases}$$

Por lo que al hacer tender $n \to \infty$, vemos que

$$\lim_{n\to\infty} \mathbb{P}(Z_n = 0) = \begin{cases} 1 & \text{si } \ p \leq q \\ \frac{q}{p}, & \text{si } \ p > q. \end{cases}$$

Observemos que si para algún $n \geq 1$ $Z_n = 0$ entonces $Z_{n+k} = 0$, para toda $k \geq 0$. Es decir que la población se extingue en algún tiempo anterior o igual a n. Tenemos que

$$\begin{aligned} \mathbb{P}(\text{extinción en un tiempo finito}) &= \lim_{n\to\infty} \mathbb{P}(\text{extinción antes del instante } n) \\ &= \lim_{n\to\infty} \mathbb{P}(Z_{n+k}, k \geq 0) \\ &= \lim_{n\to\infty} \mathbb{P}(Z_n = 0) \\ &= \begin{cases} 1 & \text{si } \ p \leq q \\ \frac{q}{p}, & \text{si } \ p > q. \end{cases} \end{aligned}$$

Conclusión: La extinsión ocurre con probabilidad 1, solamente en el caso en que $p/q = \mu = \mathbb{E}(Z_1) \leq 1$. La cual es una condición bastante natural, puesto que $\mathbb{E}(Z_n) = \mathbb{E}(Z_1)^n \leq 1$, y es entonces de esperarse que $Z_n = 0$ tarde o temprano.

Veremos que el resultado del ejercicio es consecuencia de un resultado mas general.

Teorema 7. *Tenemos que*

$$\lim_{n\to\infty} \mathbb{P}(Z_n = 0) = \mathbb{P}(\textit{extinción en un tiempo finito}) = \eta,$$

donde η es la solución más pequeña a la ecuación, $t = G(t)$. Además, $\eta = 1$, si $\mu < 1$, (el caso subcrítico) y $\eta < 1$, si $\mu > 1$ (caso super-crítico), mientras que en el caso en que $\mu = 1$, (el caso crítico) $\eta = 1$ si el tamaño de las familias tiene varianza estrictamente positiva.

Demostración. Sea $\eta_n = \mathbb{P}(Z_n = 0)$. Tenemos que

$$\eta_n = G_n(0) = G(G_{n-1}(0)) = G(\eta_{n-1}).$$

Es claro que $\{Z_n = 0\} \subset \{Z_{n+1}\}$, para todo $n \geq 1$, entonces η_n es una sucesión creciente y acotada; por la continuidad el limite de η_n existe y debe de satisfacer

$$\eta = \lim_{n\to\infty} \eta_n \leq 1, \qquad \eta = G(\eta).$$

Veamos ahora que si e es otra raíz positiva de la ecuación, entonces $\eta \leq e$. Dado que G es una función estrictamente creciente, tenemos que

$$\eta_1 = G(0) \leq G(e) = e,$$

y se sigue que

$$\eta_2 = G(\eta_1) \leq G(e) = e,$$

y por inducción se ve que $\eta_n \leq e$, para todo $n \geq 1$, y por lo tanto que $\eta \leq e$. Por lo tanto, η es la solución mas pequeña a la ecuación $t = G(t)$. Ya vimos que si $\mu > 1$ entonces la ecuación $t = G(t)$, tiene una única solución en $[0, 1[$, y de hecho otra solución a la ecuación es $t = 1$. La solución más pequeña es $\eta < 1$. Por otro lado, en el caso en que $\mu < 1$, vimos que $G(t) > t$ para todo $t \in [0, 1[$, y es claro que $G(1) = 1$, por lo tanto la soución positiva y más pequeña a la ecuación $G(t) = t$ es $\eta = 1$. En el caso especial en que $\mu = 1$, el caso crítico, necesitamos distinguir entre el caso en que $\sigma^2 = 0$, en este caso $G(s) = s$ y por lo tanto $\eta = 0$, y el caso $\sigma^2 > 0$, $G(s) > s$, $s \in [0, 1[$ y por lo tanto $\eta = 1$. □

Veamos ahora como construir un proceso de Galton Watson a partir de una caminata aleatoria.

1.1.1. El proceso de Galton Watson como una caminata aleatoria cambiada de tiempo

Sea $\{Y_i, i \geq 0\}$ una sucesión de variables aleatorias independientes e identicamente distribuidas, que toman valores en $\{-1, 0, 1, 2, \ldots\}$ y con función de densidad de probabilidad

$$f(k) = \mathbb{P}(Y_1 = k), \qquad k \in \{-1, 0, 1, 2, \ldots\}.$$

Sea $\{S_n, n \geq 0\}$ la caminata aleatoria asociada a las Y_i's con estado inicial S_0, que es independiente de la sucesión $\{Y_i, i \geq 0\}$, es decir

$$S_n = S_0 + \sum_{i=1}^{n} Y_i, \qquad n \geq 0.$$

Diremos que S es una caminata aleatoria sin saltos negativos. Supondremos que el estado inicial es $S_0 > 0$.

Sea $\nu_0 = 0$ y $\{\nu_n, n \geq 1\}$ la sucesión de variables aleatorias definidas de manera inductiva mediante

$$\nu_{n+1} = \nu_n + S_{\nu_n}, \qquad n \geq 0.$$

Lema 6. *Sea $\nu_0 = 0$ y $\{\nu_n, n \geq 1\}$ la sucesión de variables aleatorias definidas de manera inductiva mediante*

$$\nu_{n+1} = \nu_n + S_{\nu_n}, \qquad n \geq 0.$$

Se tiene que ν_n es una sucesión no-decreciente de tiempos para la caminata aleatoria $\{S_n, n \geq 0\}$. Esta cumple además que $\nu_n \leq T_0 = \text{ínf}\{n > 0 : S_n = 0\}$, y de hecho

$$\lim_{n \to \infty} \nu_n = T_0.$$

Demostración. La variable aleatoria ν_0 es por supuesto tiempo de paro ya que es una constante. Si $n = 1$ entonces $\nu_1 = S_0 > 0$, ahora como la caminata aleatoria, cuando decrece, lo hace de exactamente una unidad, entonces es claro que el primer tiempo de llegada a 0 ocurre al menos S_0 unidades de tiempo, $\nu_1 = S_0 \leq T_0$, y además $S_{\nu_1} \geq 0$. Por lo tanto para $k \geq 1$ se tiene que

$$\{\nu_1 = k\} = \{S_0 = k\}$$

que claramente es un evento de $\{S_0, S_1, \ldots, S_k\}$. Sea $\nu_2 = \nu_1 + S_{\nu_1}$ tenemos que para $k \geq 0$,

$$\begin{aligned} \{\nu_2 = k\} &= \bigcup_{r=0}^{\infty} \{\nu_1 = r, S_r = k - r\} \\ &= \bigcup_{r=0}^{k} \{\nu_1 = r, S_r = k - r\}, \end{aligned} \tag{1.1}$$

donde la segunda igualdad se debe al hecho mencionado arriba que $S_{\nu_1} \geq 0$. Esto implica también que $\nu_2 = \nu_1 + S_{\nu_1} \geq \nu_1$. Para verificar que $T_0 \geq \nu_2$ utilizamos de nuevo que como la caminata aleatoria desciende de una unidad, entonces partiendo de S_{ν_1}, al tiempo ν_1, se requieren al menos S_{ν_1}, unidades de tiempo más para llegar a 0 y por lo tanto

$$\nu_2 = \nu_1 + S_{\nu_1} \leq T_0,$$

y por lo tanto $S_{\nu_2} \geq 0$. Guardemos en mente que cada vez ν_{n+1} representa lo mínimo de tiempo en lo que se puede llegar a 0, conociendo lo ocurrido en el tiempo ν_n. Probemos ahora que si

$$0 \leq \nu_{n-1} \leq \nu_n \leq T_0, \quad S_{\nu_n} \geq 0, \text{ y } \nu_n \text{ es tiempo de paro,}$$

entonces $\nu_{n+1} = \nu_n + S_{\nu_n}$ también cumple estas propiedades. Por el argumento anterior es claro que $\nu_{n+1} \leq T_0$, y en consecuencia $S_{\nu_{n+1}} \geq 0$; por la definición se tiene que $\nu_{n+1} = \nu_n + S_{\nu_n} \geq \nu_n$; y finalmente para $k \geq 0$, se tiene que

$$\begin{aligned} \{\nu_{n+1} = k\} &= \bigcup_{r=0}^{\infty} \{\nu_n = r, S_r = k - r\} \\ &= \bigcup_{r=0}^{k} \{\nu_n = r, S_r = k - r\}, \end{aligned} \tag{1.2}$$

como ν_n es tiempo de paro se tiene que para cada $0 \leq r \leq k$ $\{\nu_n = r\}$ es un evento que puede ser determinado solamente usando $(S_0, S_1, \ldots, S_r)$ y por lo tanto su unión es un evento que puede ser determinado solamente usando a $(S_0, S_1, \ldots, S_k)$.

Faltaría demostrar que en efecto $\lim_{n\to\infty} \nu_n = T_0$. Primero si $\lim_{n\to\infty} \nu_n = \infty$, entonces $T_0 = \infty$ y por lo tanto se tiene la igualdad. Ahora, si $\lim_{n\to\infty} \nu_n = \nu_\infty < \infty$, entonces en el evento $\{\nu_\infty = j\}$, con $j \geq 0$, se tendría que existe un n_0 tal que $\nu_n = j$ para todo $n \geq n_0$. Esto implica que para un $n \geq n_0$, se tendría que

$$j = \nu_{n+1} = \nu_n + S_{\nu_n} = j + S_{\nu_n} \Rightarrow \; S_{\nu_n} = 0,$$

y por lo tanto $\nu_n \geq T_0$, ya que T_0 es el primer instante al cual la caminata aleatoria alcanza el nivel 0. Pero demostramos anteriormente que $\nu_n \leq T_0$, por lo que $\nu_n = j = T_0$. Como esto se vale para todo $j \geq 0$, se tiene entonces que $\nu_n = T_0$. $\square$

Definamos un nuevo proceso estocástico $\{Z_n = S_{\nu_n}, n \geq 0\}$. Verifiquemos que es una cadena de Markov. Recordemos que la propiedad de incrementos independientes de la caminata aleatoria se vale para todo tiempo de paro, ver Lema **??**. Esto implica que condicionalmente al evento $\{S_{\nu_n} = y, \nu_n = k, S_{\nu_i} = x_i, 0 \leq i \leq n-1\}$ que es un evento que dependen solamente de $(S_0, \ldots, S_k)$ se tenga que

$$S_{\nu_{n+1}} = S_{(\nu_{n+1}-\nu_n)+\nu_n} - S_{\nu_n} + S_{\nu_n} = y + \sum_{i=1}^{y} Y_{k+i} \stackrel{\text{Ley}}{=} y + \sum_{i=1}^{y} Y_i,$$

y además el termino $y + \sum_{i=1}^{y} Y_{k+i}$ sea independiente de $(S_0, \ldots, S_k)$. Usando esto se tiene que

$$\begin{aligned}
&\mathbb{P}(S_{\nu_{n+1}} = x | S_{\nu_n} = y, S_{\nu_i} = x_i, 0 \leq i \leq n-1) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(S_{\nu_{n+1}} = x | S_{\nu_n} = y, \nu_n = k, S_{\nu_i} = x_i, 0 \leq i \leq n-1) \\
&\qquad \times \mathbb{P}(\nu_n = k | S_{\nu_n} = y, S_{\nu_i} = x_i, 0 \leq i \leq n-1) \\
&= \sum_{k=0}^{\infty} \mathbb{P}\left(y + \sum_{i=1}^{y} Y_i = x\right) \mathbb{P}(\nu_n = k | S_{\nu_n} = y, S_{\nu_i} = x_i, 0 \leq i \leq n-1) \\
&= \mathbb{P}\left(y + \sum_{i=1}^{y} Y_i = x\right) \sum_{k=0}^{\infty} \mathbb{P}(\nu_n = k | S_{\nu_n} = y, S_{\nu_i} = x_i, 0 \leq i \leq n-1) \\
&= \mathbb{P}\left(y + \sum_{i=1}^{y} Y_i = x\right) \mathbb{P}(\nu_n < \infty | S_{\nu_n} = y, S_{\nu_i} = x_i, 0 \leq i \leq n-1) \\
&= \mathbb{P}\left(y + \sum_{i=1}^{y} Y_i = x\right).
\end{aligned}$$

De manera análoga se prueba que

$$\mathbb{P}(S_{\nu_{n+1}} = x | S_{\nu_n} = y) = \mathbb{P}\left(y + \sum_{i=1}^{y} Y_i = x\right).$$

Por lo tanto el proceso $(Z_n, n \geq 0)$ es una cadena de Markov, con matriz de transición

$$\mathbb{P}(S_{\nu_{n+1}} = x | S_{\nu_n} = y) = \mathbb{P}\left(y + \sum_{i=1}^{y} Y_i = x\right) = \mathbb{P}\left(\sum_{i=1}^{y} (1 + Y_i) = x\right) = g^{*i}(x), \qquad \forall x, y,$$

donde entendemos por g^{*0}, como la delta de Dirac en 0, $g^{*0}(k) = \delta_0(k)$, $k \geq 0$; y

$$g(x) = f(x-1), \qquad x \geq 1,$$

es decir que Z es un proceso de Galton Watson con ley de reproducción g.

Existen otras formas muy elegantes de relacionar caminatas aleatorias y procesos de Galton Watson y más general arboles de Galton Watson. La versión continua de esta transformación, es decir usando procesos de Lévy sin saltos negativos y cambios de tiempo en lugar de caminatas aleatorias, fue propuesto por Lamperti en los años setentas pero no dio una prueba formal. Sin embargo, hasta muy recientemente la comunidad trabajando en el tema lo tomó como un hecho, razón por la que Caballero, Lambert, Uribe [?] establecieron una, de ehcho dos pruebas. Referimos a ese articulo para mayor información al respecto.

Capítulo 2

El proceso de Wright-Fisher sin mutación ni selección

Consideremos un *locus* (posición de un gen a lo largo de un cromosoma) con dos tipos de alelos $A1$ y $A2$ en una población haploide de tamaño constante N con generaciones que no se traslapan y reproducción aleatoria.

Podemos representar el estado de la población en la generación n con una urna que contiene N bolas: i del tipo $A1$ y $N-i$ del tipo $A2$, para construir la generación $n+1$ escogemos al azar de la urna N bolas con reemplazo. Así si la población está compuesta de i genes del tipo $A1$ y $N-i$ del tipo $A2$, cada ensayo resulta en un gen del tipo $A1$ o del tipo $A2$ con probabilidades

$$p_i = \frac{i}{N}, \qquad q_i = \frac{N-i}{N},$$

respectivamente. Sea $\{X_n, n \geq 0\}$ el proceso estocástico tal que X_n denota el numero de $A1$-genes en la n-ésima generación.

Ejercicio 3. Según esta descripción, ¿se puede afirmar que $\{X_n, n \geq 0\}$ es una cadena de Markov con espacio de estados $E = \{0, 1, 2, \ldots, N\}$ y su matriz de transición $\mathbf{P}$, tiene la forma:

$$\mathbf{P}_{i,j} = \mathbb{P}(X_{n+1} = j | X_n = i) = \binom{N}{j} p_i^j q_i^{N-j} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j},$$

para $i, j \in \{0, 1, \ldots, N\}$? (Justificar su respuesta)

El objetivo de los siguientes ejercicios es, suponiendo que la respuesta a la pregunta anterior es afirmativa y como ejercicio de aplicación de la teoría estudiada, analizar diversas propiedades de esta cadena de Markov. La cual recibe el nombre de proceso de Wright-Fisher sin mutación ni selección.

Ejercicio 4. Dé una clasificación de los estados para esta cadena de Markov. (Incluyendo: clases de comunicación, estados transitorios, recurrentes, absorbentes).

Solución. Las clases de comunicación de este proceso estocástico son $\{0\}, \{1, 2, \ldots, N-1\}, \{N\}$. Los estados $\{0\}$ y $\{N\}$ son absorbentes y por lo tanto recurrentes. Los estados $\{1, 2, \ldots, N-1\}$

son transitorios ya que si la población inicia en el estado 1 tiene probabilidad

$$\mathbf{P}_{1,0} = \mathbb{P}(X_{n+1} = 0 | X_n = 1) = \binom{N}{0} p_1^0 q_1^N = \binom{N}{0} \left(\frac{1}{N}\right)^0 \left(\frac{N-1}{N}\right)^N > 0,$$

de alcanzar el estado 0 del cual no regresará por ser este un estado absorbente. Como el ser recurrente o transitorio es una propiedad de clase se sigue que los estados $\{2, 3, \ldots, N-1\}$ son también transitorios. $\square$

Ejercicio 5. Demuestre que

$$\mathbb{E}_i(X_1) = \sum_{z \in E} z \mathbf{P}_{i,z} = i, \qquad \forall i \in E.$$

Pruebe usando inducción que para todo $n \geq 1$,

$$\mathbb{E}_i(X_n) = \sum_{z \in E} z \mathbf{P}_{i,z}^{(n)} = i, \qquad \forall i \in E,$$

es decir que $\{X_n, n \geq 0\}$ cumple la *propiedad de martingala.*

Solución. Por la forma de la matriz de transición se sigue para cada $i \in \{0, 1, \ldots, N\}$ que la ley de X_{n+1} dado $X_n = i$, es una ley Binomial de paramétros (p_i, N) y por propiedades de la ley Binomial se sigue que

$$\mathbb{E}(X_{n+1} | X_n = i) = \mathbb{E}(X_1 | X_0 = i) = N p_i = i.$$

Veamos ahora por inducción que la segunda afirmación se vale. Como ya hemos demostrado que la propiedad es valida para $n = 0$, supongamos que el resultado se vale para $n - 1$ y demostremos para n. En efecto, se obtiene usando la ecuación de Chapman Kolmogorov, un cambio en el orden de las sumas y la hipótesis de inducción que

$$\mathbb{E}_i(X_{n+1}) = \sum_{z \in E} z \mathbf{P}_{i,z}^{(n+1)} = \sum_{z \in E} z \left(\sum_{j \in E} \mathbf{P}_{i,j} \, \mathbf{P}_{j,z}^{(n)} \right) = \sum_{j \in E} \mathbf{P}_{i,j} \left(\sum_{z \in E} z \, \mathbf{P}_{j,z}^{(n)} \right) = \sum_{j \in E} j \, \mathbf{P}_{i,j} = i,$$

para todo $i \in E$. $\square$

Ejercicio 6. Usando la expresión anterior demuestre que

$$\lim_{n \to \infty} \mathbf{P}_{i,N}^{(n)} = \frac{i}{N}, \quad \lim_{n \to \infty} \mathbf{P}_{i,0}^{(n)} = 1 - \frac{i}{N} \qquad \forall i \in E.$$

(Indicación: recuerde que implicaciones tiene que que un estado sea transitorio.)

Solución. Del inciso anterior concluimos que se tiene la igualdad

$$i = \sum_{z \in E} z \mathbf{P}_{i,z}^{(n)} = \sum_{z=1}^{N} z \mathbf{P}_{i,z}^{(n)} = N \mathbf{P}_{i,N}^{(n)} + \sum_{z=1}^{N-1} z \mathbf{P}_{i,z}^{(n)}, \qquad n \geq 0.$$

Tomando límites en ambos lados de la igualdad y usando que se tiene una suma finita para intercambiar el límite y la suma obtenemos que

$$i = N \lim_{n\to\infty} \mathbf{P}_{i,N}^{(n)} + \lim_{n\to\infty} \sum_{z=1}^{N-1} z\mathbf{P}_{i,z}^{(n)} = N \lim_{n\to\infty} \mathbf{P}_{i,N}^{(n)} + \sum_{z=1}^{N-1} z \lim_{n\to\infty} \mathbf{P}_{i,z}^{(n)}.$$

Recordemos que los estados $\{1, 2, \ldots, N-1\}$ son transitorios y por la Proposición **??** se tiene que

$$\lim_{n\to\infty} \mathbf{P}_{i,z}^{(n)} = 0, \qquad z \in \{1, 2, \ldots, N-1\}.$$

De lo anterior se sigue que

$$\lim_{n\to\infty} \mathbf{P}_{i,N}^{(n)} = \frac{i}{N}.$$

Tomando complementos se tiene que

$$\mathbf{P}_{i,0}^{(n)} = 1 - \mathbf{P}_{i,N}^{(n)} - \sum_{z=1}^{N-1} \mathbf{P}_{i,z}^{(n)}, \qquad n \geq 1.$$

Tomando límites en ambos lados de la igualdad y argumentando como en las lineas anteriores se concluye que

$$\lim_{n\to\infty} \mathbf{P}_{i,0}^{(n)} = \lim_{n\to\infty} (1 - \mathbf{P}_{i,N}^{(n)}) - \lim_{n\to\infty} \sum_{z=1}^{N-1} \mathbf{P}_{i,z}^{(n)} = 1 - \frac{i}{N} - 0 = \frac{N-i}{N}.$$

□

Ejercicio 7. Interprete el resultado anterior en términos del modelo, demostrando y usando los siguientes hechos.

- Se tiene la siguiente igualdad de conjuntos

  $$\{H_N \leq n\} = \{X_n = N\}$$

  de donde que

  $$\mathbb{P}_i(H_N \leq n) = \mathbb{P}_i(X_n = N) = \mathbf{P}_{i,N}^{(n)}, \qquad i \in E.$$

- Se tiene el siguiente límite

  $$\mathbb{P}_i(H_N < \infty) = \lim_{n\to\infty} \mathbb{P}_i(H_N \leq n) = \frac{i}{N}.$$

- Se tiene la siguiente igualdad:

  $$\mathbb{P}_i(H_N < \infty) = \mathbb{P}_i(H_N < H_0) = \frac{i}{N}, \qquad i \in E.$$

Solución. Observemos que puesto que N es absorbente se tiene la igualdad de eventos

$$\{H_N \leq n\} = \{X_n = N\}$$

de donde que

$$\mathbb{P}_i(H_N \leq n) = \mathbb{P}_i(X_n = N) = \mathbf{P}^{(n)}_{i,N}, \qquad i \in E.$$

Como los eventos $B_n := \{H_N \leq n\}$, $n \geq 1$ forman una sucesión creciente de eventos con límite $\{H_N < \infty\}$, y por la continuidad por abajo de las medidas de probabilidad se tiene que

$$\mathbb{P}_i(H_N < \infty) = \lim_{n\to\infty} \mathbb{P}_i(H_N \leq n).$$

Usando el resultado del inciso anterior tenemos que

$$\mathbb{P}_i(H_N < \infty) = \frac{i}{N}.$$

Además como a la larga la cadena solamente puede estar en los estados N o 0 se tiene que $H_N < \infty$ si y solamente si la cadena alcanza el nivel N antes de alcanzar el nivel 0 por lo que

$$\mathbb{P}_i(H_N < \infty) = \mathbb{P}_i(H_N < H_0) = \frac{i}{N}, \qquad i \in E.$$

Así la interpretación del resultado del inciso anterior es que la probabilidad de que la cadena sea absorbida en el estado N es proporcional al número inicial de $A1$-genes, y por lo tanto a mayor número inicial de genes del tipo $A1$, mayor probabilidad de que la cadena sea absorbida en el estado N. □

Alternativamente al cálculo que acabamos de hacer, en la bibliografía sobre el tema se encontrará que la probabilidad de fijación

$$\mathbb{P}_i(H_N < H_0) = \frac{i}{N}, \qquad \forall i \in E,$$

es determinada usando el teorema de paro optimo para martingalas, pero como en estas notas no se presuponen conocimientos de la teoría de martingalas veremos una forma aún más directa de calcularla usando el Teorema **??** del capítulo anterior.

Ejercicio 8. Proponga un método alternativo para demostrar que

$$h(i) := \mathbb{P}_i(H_N < \infty) = \frac{i}{N}, \qquad i \in \{0, 1, \ldots, N-1\}, \qquad h(N) = 1.$$

(Indicación: recuerde que este vector es solución a algún sistema de ecuaciones lineales y utilice que

$$i = \mathbb{E}_i\left(X_1\right) = \sum_{j=0}^{N} j\, \mathbf{P}_{i,j}\ .)$$

Ejercicio 9. Verificar que la varianza de X_n dado $X_0 = i$ satisface que

$$\operatorname{Var}_i(X_n) := \mathbb{E}((X_n - i)^2 | X_0 = i) = Ni\left(1 - \frac{i}{N}\right)\left(1 - \left(1 - \frac{1}{N}\right)^n\right). \tag{2.1}$$

Calcular el límite

$$\lim_{n\to\infty} \operatorname{Var}_i(X_n). \tag{2.2}$$

Indicación: Para demostrar esta afirmación usar que si Y sigue una ley Binomial(N, p) entonces

$$\operatorname{Var}(Y) = Np(1-p),$$

(verifiquelo) y la identidad

$$\operatorname{Var}(Y) = \mathbb{E}(\operatorname{Var}(Y|Z)) + \operatorname{Var}(\mathbb{E}(Y|Z)),$$

donde Z es cualquier variable aleatoria; podrá utilizar esta identidad sin demostrarla. Usar que la ley de X_{n+1} dado $X_n = j$ es Binomial de paramétros $(N, \frac{j}{N})$, para verificar que

$$\begin{aligned}
\operatorname{Var}_i(X_n) &= \mathbb{E}_i\left(\operatorname{Var}_i(X_n|X_{n-1})\right) + \operatorname{Var}_i(\mathbb{E}_i(X_n|X_{n-1})) \\
&= \mathbb{E}_i\left(N\frac{X_{n-1}}{N}\left(1 - \frac{X_{n-1}}{N}\right)\right) + \operatorname{Var}_i(X_{n-1}) \\
&= \mathbb{E}_i\left(X_{n-1}\left(1 - \frac{X_{n-1}}{N}\right)\right) + \operatorname{Var}_i(X_{n-1}) \\
&= \mathbb{E}_i(X_{n-1}) - \frac{1}{N}\mathbb{E}_i\left(X_{n-1}^2\right) + \operatorname{Var}_i(X_{n-1}) \\
&= i - \frac{1}{N}\left(\operatorname{Var}_i(X_{n-1}) + (\mathbb{E}_i(X_{n-1}))^2\right) + \operatorname{Var}_i(X_{n-1}) \\
&= i - \frac{1}{N}\left(\operatorname{Var}_i(X_{n-1}) + i^2\right) + \operatorname{Var}_i(X_{n-1}) \\
&= \left(1 - \frac{1}{N}\right)\operatorname{Var}_i(X_{n-1}) + i\left(1 - \frac{i}{N}\right)
\end{aligned}$$

Con lo cual hemos establecido la ecuación en recurrencias

$$\operatorname{Var}_i(X_n) = \left(1 - \frac{1}{N}\right)\operatorname{Var}_i(X_{n-1}) + i\left(1 - \frac{i}{N}\right), \qquad n \geq 1,$$

con la condición inicial $\operatorname{Var}_i(X_0) = 0$. Se verifica que la solución a esta ecuación está en efecto dada por (2.1) usando el método descrito a continuación con $x_n = \operatorname{Var}_i(X_n)$, y $y_0 = -Ni(1-\frac{i}{N})$.

> *El método general para resolver una ecuación en recurrencias de la forma*
>
> $$x_{n+1} = ax_n + b, \qquad n \geq 0, x_0 \in \mathbb{R}, a \neq 1$$
>
> *es: se busca primero una solución a la ecuación $x = ax + b$, es decir $x = b/(1-a)$. En seguida, se observa que la sucesión y_n definida por $y_n = x_n - (b/(1-a))$, $n \geq 0$*

satisface la ecuación $y_{n+1} = a y_n$, $n \geq 0$ y por lo tanto $y_n = a^n y_0$, De esta manera se ve que la solución general a la recurrencia está dada por

$$x_n = y_0 a^n + (b/(1-a)), \qquad n \geq 0,$$

además el valor de y_0 está completamente determinado por x_0, a, b.

□

Ejercicio 10. La varianza calculada arriba permite calcular la probabilidad esperada de muestrear dos genes diferentes (heterocigosis) en la generación n dado que al tiempo 0 hay i genes del tipo $A1$,

$$H_n^{(0)} = \frac{X_n(N - X_n)}{N(N-1)}, \qquad B_n(i) = \mathbb{E}_i(H_n^{(0)}), \qquad n \geq 0.$$

Se afirma que

$$B_n(i) = \mathbb{E}_i(H_n^{(0)}) = B_0(i)\left(1 - \frac{1}{N}\right)^n,$$

lo cual implica un decaimiento geométrico de la probabilidad de heterocigosis. Demuestre esta igualdad.

Solución. La prueba de la igualdad es una consecuencia fácil del cálculo de la varianza como se ve en las siguientes lineas.

$$\begin{aligned}
B_n(i) &= \mathbb{E}_i\left(\frac{X_n(N-X_n)}{N(N-1)}\right) \\
&= \frac{N\,\mathbb{E}_i(X_n) - \mathbb{E}_i(X_n^2)}{N(N-1)} \\
&= \frac{Ni - i^2 - \mathrm{Var}_i(X_n)}{N(N-1)} \\
&= \frac{1}{N(N-1)}\left(i(N-i) - i(N-i)\left(1 - \left(1 - \frac{1}{N}\right)^n\right)\right) \\
&= \frac{i(N-i)}{N(N-1)}\left(1 - \frac{1}{N}\right)^n
\end{aligned} \tag{2.3}$$

□

Es posible obtener más información del proceso de Wright-Fisher usando propiedades de cadenas de Markov, sin embargo para algunos conceptos simples es necesario usar técnicas más avanzadas como lo son las aproximaciones por difusión, el cálculo de la media del tiempo de absorción es un ejemplo de esto.

Denotamos por m_i al tiempo esperado de fijación dado que $X_0 = i$, es decir,

$$m_i = \mathbb{E}(H_{\{0,N\}} | X_0 = i), \qquad H_{\{0,N\}} = \text{ínf}\{n \geq 0 : \; X_n = 0 \text{ ó } X_n = N\}.$$

Sabemos por el Teorema **??** que m_i satisface el sistema de ecuaciones (**??**)

$$m_i = 1 + \sum_{j=1}^{N-1} \mathbf{P}_{i,j}\, m_j, \qquad m_0 = 0 = m_N.$$

Resolver este sistema explicitamente es complicado pero se puede verificar usando aproximaciones por difusión que

$$m_i \approx 2(i \log(i/N)) + (N-i) \log \left(1 - \frac{i}{N}\right),$$

pero omitiremos los detalles por encontrarse fuera del alcance de estas notas. Observe que

$$m_1 \approx 2 + 2\log(N), \quad m_{N/2} \approx 1{,}4N, \quad \text{generaciones}$$

Ejercicio 11. Supongamos ahora que condicionamos al proceso de Wright-Fisher a que eventualmente el alelo del tipo $A1$ se fije, es decir que condicionamos con respecto al evento $H_N < H_0$. El objetivo de este ejercicio es verificar que la cadena de Wright-Fisher condicionada a que el alelo $A1$ se fije es una nueva cadena de Markov $\{X^*_n, n \geq 0\}$ con espacio de estados $E^* = \{1, 2, \dots, N\}$ y determinar su matriz de transición. Sea P^* la matriz cuyas entradas están dadas por

$$\mathbf{P}^*_{i,j} = \frac{\mathbb{P}(X_1 = j, X_0 = i | H_N < H_0)}{\mathbb{P}(X_0 = i | H_N < H_0)}, \qquad i, j \in E^*.$$

Demuestre que en efecto $\mathbf{P}^* = \{\mathbf{P}^*_{i,j}, i, j \in E\}$, cumple las propiedades para ser una matriz de transición. Demuestre las siguientes identidades y dé una interpretación:

$$\begin{aligned}
\mathbf{P}^*_{i,j} &= \frac{\mathbb{P}(X_1 = j, X_0 = i | H_N < H_0)}{\mathbb{P}(X_0 = i | H_N < H_0)} \\
&= \frac{\mathbb{P}(X_1 = j, X_0 = i, H_N < H_0)}{\mathbb{P}(X_0 = i, H_N < H_0)} \\
&= \frac{\mathbb{P}(H_N < H_0 | X_1 = j, X_0 = i)\, \mathbb{P}(X_1 = j, X_0 = i)}{\mathbb{P}(H_N < H_0 | X_0 = i)\, \mathbb{P}(X_0 = i)} \\
&= \frac{\mathbb{P}(H_N < H_0 | X_0 = j)}{\mathbb{P}(H_N < H_0 | X_0 = i)} \mathbf{P}_{i,j} \\
&= \frac{j}{i} \mathbf{P}_{i,j} = \binom{N-1}{j-1} \left(\frac{i}{N}\right)^{j-1} \left(\frac{N-i}{N}\right)^{N-j}
\end{aligned}$$

¿Se puede afirmar que existe una cadena de Markov con esta matriz de transición?

Esto se puede interpretar como que bajo la probabilidad de que el alelo $A1$ se fija entonces al menos un alelo $A1$ debe de ser producido, por lo que $\mathbf{P}^*_{i,j}$ denota la probabilidad de que los $N-1$ ensayos produzcan exactamente $j-1$ alelos del tipo $A1$.

Usando un argumento similar al usado en el caso sin condicionar se puede verificar que el tiempo promedio de fijación en el alelo $A1$ está dado por

$$m^*_i = \mathbb{E}^*(H_{\{N\}} | X_0 = i) \approx -2N\frac{N}{i}\left(1 - \frac{i}{N}\right) \log \left(1 - \frac{i}{N}\right),$$

y de esto se sigue que

$$m^*_1 \approx 2N, \quad m^*_{N/2} \approx 1{,}4N, \quad m^*_{N-1} \approx 2\log(N).$$