Universidad Nacional Autónoma de México

Facultad de Ciencias

Modelos probabilísticos de selección natural y la conjetura de Muller

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

PRESENTA:

LUIS ENRIQUE IGNACIO GÓMEZ ORDOÑEZ

TUTOR:

DR. ADRIÁN GONZÁLEZ-CASANOVA SOBERÓN

Ciudad de México, 2019

Hoja de datos del jurado:

1. Datos del alumno.
    Gómez
    Ordoñez
    Luis Enrique Ignacio
    7224693356
    Universidad Nacional Autónoma de México
    Facultad de Ciencias
    Actuaría
    415031660

2. Datos del tutor:
    Dr.
    Adrian
    González
    Casanova Soberón

3. Datos del sinodal 1
    Dra.
    Sandra
    Palau
    Calderón

4. Datos del sinodal 2
    Dra.
    Verónica de la Santisima Faz
    Miró
    Pina

5. Datos del sinodal 3
    Dra.
    María Clara
    Fittipaldi

6. Datos del sinodal 4
    M. en C.
    Lizbeth
    Peñaloza
    Velasco

7. Datos de la tesis.
    Título: Modelos probabilisticos en selección natural y la conjetura de Muller
    83 p
    2019

Índice general

1. Introducción **1**

2. Marco Teórico **3**

2.1. Cadenas de Markov a tiempo continuo 3

2.2. Difusiones unidimensionales 8

2.3. Simulación estocástica 18

2.3.1. Números aleatorios 18

2.3.2. Simulación de variables aleatorias y procesos estocásticos 19

2.3.2.1. Distribuciones discretas 19

2.3.2.2. Distribuciones absolutamente continuas 20

2.3.2.3. Cadenas de Markov 21

2.3.3. Difusiones y ecuaciones diferenciales estocásticas 23

2.3.3.1. Método de Euler-Maruyama 23

2.3.3.2. Método de Milstein 24

2.3.4. Estimadores de Monte Carlo 24

2.4. Gráficas 26

2.4.1. Gráficas aleatorias 27

3. Modelos de deriva génica **31**

3.1. Modelo de Wright-Fisher 31

3.1.1. El modelo de Wright-Fisher es una martingala 33

3.1.2. Simulación del modelo de Wright-Fisher 34

3.2. Modelo de Moran 36

4. Selección y mutación **39**

4.1. Selección en el modelo de Wright-Fisher 39

4.2. Selección en el modelo de Moran 43

4.2.1. Convergencia a la difusión de Wright-Fisher 45

4.3. Modelo de González Casanova-Spanò 46

4.3.1. Gráfica aleatoria de Wright-Fisher 49

4.3.2. Frecuencias de los alelos 50

5. Matraca de Muller **53**

5.1. Matraca de Muller en un modelo de gráficas aleatorias 54

5.1.1. La tasa de la matraca de Muller 56

5.1.2. Posibles variaciones 62

6. Discusión **63**

A. Apéndice **65**

A.1. Matraca de Muller usando el enfoque de Haigh 65

A.1.1. El modelo 65

A.1.2. Análisis de la matraca de Muller 67

A.2. Códigos de simulación 69

A.2.1. Análisis del Modelo de Wright Fisher 69
A.2.2. Modelo de selección y mutación en gráficas aleatorias 70
A.2.3. Matraca de Muller en el modelo de selección y mutación en gráficas aleatorias 72

Capítulo 1

Introducción

El presente trabajo tiene como objetivo brindar al lector una introducción a los modelos clásicos de deriva génica, el papel de la selección natural y la mutación en dicho mecanismo, el análisis de la matraca de Muller, así como presentar algoritmos para simular uno de los modelos más recientes del área.

La deriva génica (o deriva genética) es un fenómeno estocástico que altera la composición de una población, pues las frecuencias alélicas (número de individuos de un tipo como proporción del total de individuos en la población) de una población cambian a lo largo de las generaciones por efecto del azar.

Después de algunas generaciones y debido a la deriva génica puede que se presente la perdida de algunos alelos y la fijación de otros (es decir, el aumento al 100 % de su frecuencia).

En la teoría evolutiva clásica se ha visto que la deriva genética puede tener efectos importantes cuando una población pierde tamaño de forma considerable por un desastre natural, o cuando un grupo se separa de la población principal y crean una nueva población, además. Se ha observado que en poblaciones con pocos individuos, los efectos de la deriva génica pueden ser irreversibles.

A diferencia de la selección natural, la deriva génica no depende de los efectos beneficiosos

o perjudiciales de un alelo; pues depende únicamente del azar, al seleccionar subconjuntos aleatorios de individuos para producir la siguiente generación.

Sin embargo, a los modelos clásicos se le puede agregar selección natural e incluso mutación, con el objetivo de tener un mecanismo más robusto.

Tomando como base el modelo teórico de González Casanova-Spanò [GS18] y agregando teoría de gráficas aleatorias, se introduce un nuevo modelo con el objetivo de estudiar un fenómeno conocido como la matraca de Muller, en pocas palabras, la matraca de Muller consiste en la acumulación de mutaciones desfavorables en una población que causa la extinción de alguna clase (tipo) de los individuos, cuando esto pasa se dice que la matraca hace *click*.

La relevancia de este trabajo es que para la matraca de Muller aún no se ha tenido una formula cerrada para su tasa, que es el número de generaciones tras las cuales habrá desaparecido la mejor clase de la población (por ejemplo, la clase sin ninguna mutación). En este trabajo se abordó el estudio de dicha tasa en el modelo propuesto (una de las contribuciones de la tesis), usando un enfoque computacional para tener una idea de como afectan los parámetros de selección y mutación a la tasa de la matraca de Muller.

Capítulo 2

Marco Teórico

2.1. Cadenas de Markov a tiempo continuo

Se dará una breve introducción de Cadenas de Markov a tiempo continuo, basada en los libros de Norris [Nor97], Rincón [Rin12] y Ross [Ros96].

Definición 2.1.1. *(Cadena de Markov a tiempo continuo)*

Un proceso estocástico $(X_t)_{t\geq 0}$ *es una cadena de Markov a tiempo continuo con espacio de estados* E*, donde* E *es un conjunto discreto, si para todo* $0 \leq t$*,* $0 \leq u \leq s$ *así como* $i, j, k \in E$*, se cumple:*

$$\mathbb{P}(X_{t+s} = j | X_s = i, X_u = k) = \mathbb{P}(X_{t+s} = j | X_s = i)$$

con $t > 0$ *para una cantidad finita de tiempo.*

La interpretación de la propiedad anterior es que la distribución condicional del estado del proceso al tiempo futuro $t + s$ dado el estado a tiempo presente s y los estados a tiempo pasado, depende únicamente del estado presente y es independiente de los estados pasados. Además, se suele definir la *función de transición* del proceso como:

Definición 2.1.2. *Función de transición*

$$P_{i,j}^{(t)} := \mathbb{P}(X_{t+s} = j | X_s = i), \quad t > s > 0$$

$$P_{i,j}^{0} := \delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} .$$

Definición 2.1.3. *(Homogeneidad)*

Si

$$P_{i,j}^{(t)} = \mathbb{P}(X_{t+s} = j \mid X_s = i) = \mathbb{P}(X_t = j \mid X_0 = i), \quad t > s > 0$$

se dice que la cadena de Markov es homogénea.

Observación. *[Ros96] Una Cadena de Markov a tiempo continuo es un proceso estocástico que tiene las siguientes propiedades, cada vez que entra al estado i:*

- *La cantidad de tiempo que pasa en ese estado antes de hacer una transición a un estado diferente se distribuye exponencial con media λ_i, donde $\lambda_i = \sum_{j \in E} \lambda_{i,j}$.*
- *Cuando el proceso deja el estado i, entra al estado j con probabilidad $P_{i,j}$, donde $P_{i,j}$ satisface:*

$$P_{i,i} = 0, \quad \forall\, i$$

$$\sum_{j \in E} P_{i,j} = 1, \quad \forall\, i$$

.

La intuición detrás de la observación anterior es que una Cadena de Markov a tiempo continuo se mueve de estado a estado con el mismo mecanismo que una Cadena de Markov a tiempo discreto, pero es tal que la cantidad de tiempo que pasa en cada estado, antes de

pasar al siguiente estado, se distribuye exponencialmente; además, la cantidad de tiempo que el proceso pasa en el estado i, y el siguiente estado visitado, deben ser variables aleatorias independientes pues de lo contrario no se cumpliría la propiedad de Markov.

Teorema 2.1.4. *(Probabilidades de transición en t unidades de tiempo)*

Sean i y j dos estados, para cualquier $t \geq 0$. (Prop. 5.1 de [Rin12]).

$$P_{i,j}^{(t)} = \delta_{i,j} e^{-\lambda_i t} + \lambda_i e^{-\lambda_i t} \int_0^t e^{\lambda_i u} \Big(\sum_{k \neq i} P_{i,k} P_{k,j}^{(u)} \Big) du.$$

Prueba Si $i \in E$ no es un estado absorbente, $T_i \sim \exp(\lambda)$ es el tiempo de vida en el estado i y X_u es el momento de cambio de estado, entonces:

$$\begin{aligned} P_{i,j}^{(t)} &= \mathbb{P}(X_t = j \mid X_s = i) \\ &= \mathbb{P}(X_t = j,\ T_i > t \mid X_s = i) + \mathbb{P}(X_t = j,\ T_i \leq t \mid X_s = i) \\ &= \delta_{i,j}\ \mathrm{e}^{\lambda_i t} + \int_s^t f_{X_t,\, T_i \mid X_s}\ (j, u \mid i)\ du \\ &= \delta_{i,j}\ \mathrm{e}^{\lambda_i t} + \int_s^t \sum_{k \neq i} f_{X_t,\, X_u\, T_i \mid X_s}\ (j, k, u \mid i)\ du, \end{aligned}$$

usando independencia y la propiedad de Markov:

$$\begin{aligned} f_{X_t,\, X_u\, ,T_i \mid X_s}\ (j, k, u \mid i) &= f_{X_t \mid\ X_u\ ,T_i\ X_s}\ (j \mid k, u, i) f_{X_u \mid T_i\ ,X_0}\ (k \mid u, i) f_{T_i\ X_s}(u \mid i) \\ &= P_{k,j}^{(t-u)} P_{i,k}\ \lambda_i \mathrm{e}^{-\lambda_i u}, \end{aligned}$$

por lo tanto:

$$P_{i,j}^{(t)} = \delta_{i,j}\ \mathrm{e}^{-\lambda_i t} + \int_0^t -\lambda_i\ \mathrm{e}^{-\lambda_i u} \Big(\sum_{k \neq i} P_{i,k} P_{k,j}^{(t-u)} \Big)\ du.$$

Teorema 2.1.5. *(Ecuación de Chapman-Kolmogorov)*

Para cualquier par de estados i y j, y para cualquier $t \geq 0$ y $s \geq 0$,

$$P_{i,j}^{(t+s)} = \sum_k P_{i,k}^{(t)} P_{k,j}^{(s)}.$$

Prueba Por la propiedad de Markov:

$$\begin{aligned} P_{i,j}^{(t+s)} &= \mathbb{P}(X_{t+s} = j | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{t+s} = j,\ X_t = k \mid X_0 = i) \\ &= \sum_k \mathbb{P}(X_{t+s} = j \mid X_t = k)\, \mathbb{P}(X_t = k \mid X_0 = i) \\ &= \sum_k P_{i,k}^{(t)} P_{k,j}^{(s)}. \end{aligned}$$

Observación. *Del teorema 2.1.4 se sigue que $P_{i,j}^{(t)}$ es una función de t continua y diferenciable, con derivada:*

$$\frac{d}{dt} P_{i,j}^{(t)} = -\lambda_i P_{i,j}^{(t)} + \lambda_i \sum_{k \neq i} P_{i,k} P_{k,j}^{(t)}.$$

Si se calcula el limite cuando t tiende a 0 se tiene que

$$\lim_{t \to 0} \frac{d}{dt} P_{i,j}^{(t)} = \frac{d}{dt} P_{i,j}^{(0)} = -\lambda_i \delta_{i,j} + \lambda_i P_{i,j}.$$

Definición 2.1.6. *(Generador infinitesimal)*

La matriz $\{Q = q_{i,j},\ i, j \in E\}$ se conoce como el generador infinitesimal de una cadena de Markov a tiempo continuo, donde

$$q_{i,j} = P'_{i,j} = \begin{cases} -\lambda_i, & i = j \\ \lambda_i P_{i,j} = \lambda_{i,j}, & i \neq j \end{cases},$$

así como la matriz de probabilidades de transición caracteriza a una cadena de Markov a tiempo discreto el generador infinitesimal hace lo propio con una cadena de Markov a tiempo continuo; además, la matriz G tiene las siguientes propiedades:

- $q_{i,j} \geq 0$, si $i \neq j$

- $q_{i,i} \leq 0$
- $\sum_j q_{i,j} = 0.$

Ejemplo 2.1.7. *(Proceso Poisson) Un Proceso Poisson es una Cadena de Markov a tiempo continuo, tal que* $\mathbb{P}(X_0 = 0) = 1$*, los tiempos exponenciales de estancia en cada estado tienen parámetro* λ *y las probabilidades de saltos de un estado a otro son:*

$$P_{i,j} = \begin{cases} 1, & j = i+1 \\ 0, & j \neq i+1 \end{cases}$$

Además, el generador infinitesimal G *del proceso de Poisson de parámetro* λ *es:*

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Observación. *Un proceso de Markov a tiempo continuo puede definirse a través del comportamiento de las probabilidades de transición* $P_{i,j}^{(t)}$ *cuando* $t \to 0$*. A estas probabilidades se les conoce como probabilidades infinitesimales y pueden expresarse en términos de los parámetros infinitesimales:*

Si $t \to 0$*:*

1. $P_{i,i}^{(t)} = 1 + t\, q_{i,i} + o(t)$
2. $P_{i,j}^{(t)} = t\, q_{i,j} + o(t), \quad i \neq j,$

recordando que $E = o(t) \iff \frac{E}{t} \to 0$.

El resultado anterior se sigue del desarrollo de la serie de Taylor de la función $p_{i,j}(t)$ alrededor de cero hasta el término lineal.

2.2. Difusiones unidimensionales

Esta sección esta basada en los textos de Karlin y Taylor [KT81], y Knight [Kni00].

De acuerdo a Etheridge [Eth11], un proceso estocástico que cumpla la propiedad fuerte de Markov, y cuyas trayectorias $X(t)$ son funciones continuas de t suele llamarse un proceso de difusión.

Observación. *(Propiedad fuerte de Markov)*

Sea X_n una cadena de Markov y sea τ un tiempo de paro respecto de este proceso. Condicionado al evento $(\tau < \infty)$, el proceso $X_{\tau+n}$ es una cadena de Markov, es decir, la probabilidad $\mathbb{P}(X_{\tau+n+1} = j | X_0 = x_0, ..., X_{\tau+n-1} = x_{n-1}, X_{\tau+n} = i)$ es igual a $\mathbb{P}(X_{\tau+n+1} = j | X_{\tau+n} = i)$.

La utilidad de las difusiones recae en que pueden ser usadas con diversos propósitos:

- Fenómenos económicos, sociales y biológicos; como por ejemplo, fluctuaciones de precios de valores o variaciones en el tamaño de la población.
- Algunos funcionales como, por ejemplo, probabilidades de primer paso o distribuciones estacionarias, pueden ser calculadas explícitamente para las difusiones.

Para dar una intuición antes de la definición formal de una difusión se usará el desarrollo de Durrett [Dur10], se puede empezar recordando que una cadena de Markov a tiempo continuo, X_t se define dadas las tasas $q_{i,j}$ con las cuales la cadena salta de i a j. Esto es, si se usa P_i

como la distribución empezando en i entonces

$$P_i(X_s = j) = q_{i,j}s + o(s) \tag{2.1}$$

donde $o(s)$ representa un término de orden menor a s. Si se define $q_{i,i} = -\sum_{j\neq i} q_{i,j}$ entonces las filas de la matriz suman 0 y

$$P_i(X_s = i) = 1 + q_{i,i}s + o(s). \tag{2.2}$$

Al combinar las últimas dos expresiones, se sigue que si f es una función acotada entonces

$$E_i f(X_s) = (1 + q_{i,i}s)f(i) + \sum_{j\neq i} q_{i,j}sf(j) + o(s). \tag{2.3}$$

Al hacer álgebra se tiene un cambio en el índice de la suma, y tras restar $f(i)$ y dividir entre s se tiene

$$\frac{E_i f(X_s) - f(i)}{s} = \sum_j q_{i,j}f(j) + o(1) \tag{2.4}$$

Haciendo tender $s \to 0$,

$$\frac{d}{ds}E_i f(X_s)\Big|_{s=0} = Qf(i). \tag{2.5}$$

En la expresión anterior, el lado derecho es la i-ésima componente de la matriz $Q = q_{i,j}$ y el vector $f(j)$. A la matriz Q se le llamará el generador infinitesimal de X_s.

A continuación se harán algunos cálculos para entender los generadores infinitesimales heuristicamente.

Considérense ahora los siguientes generadores de procesos de Markov, más adelante se definirán μ y σ^2, sea

$$L^N f(i) = \sigma^2\Big(\frac{N^2}{2}\Big[f(i + \frac{1}{N}) - f(i)\Big] + \frac{N^2}{2}\Big[f(i - \frac{1}{N}) - f(i)\Big]\Big) \tag{2.6}$$

el generador del proceso S_{N^2t} que es una caminata aleatoria; además se tiene que $\lim_{N\to\infty} \frac{S_{N^2t}}{N} = B_t$ donde B_t es el movimiento Browniano estándar (más adelante se da una definición más precisa).

Observación. *(Caminata aleatoria)*

Una caminata aletoria S_n es un proceso estocástico de la forma:

$$S_n = \sum_{i=1}^{n} Z_i$$

donde Z_i es una variable aletoria independiente que toma los valores 1 y -1 con igual probabilidad.

En la figura 2.1 se observa como por medio de simulaciones se puede obtener una trayectoria que se parece cada vez más (conforme N aumenta) a la de un movimiento Browniano estándar tomando una caminata aleatoria simple con pasos cada vez más pequeños y más rápidos.

Aplicando la serie de Taylor al término $f(i + \frac{1}{N})$ de 2.6 se tiene

$$f(i + \frac{1}{N}) = f(i) + \frac{1}{N}f'(i) + \frac{1}{N^2}\frac{f''(i)}{2} + o(\frac{1}{N^2}). \tag{2.7}$$

Sustituyendo esta expresión en la original se tiene

$$\begin{aligned} L^N f(i) = \sigma^2(&\frac{N^2}{2}[f(i) + \frac{1}{N}f'(i) + \frac{1}{N^2}\frac{f''(i)}{2} + o(\frac{1}{N^2}) - f(i)] \\ &+ \frac{N^2}{2}[f(i) + \frac{-1}{N}f'(i) + \frac{1}{N^2}\frac{f''(i)}{2} + o(\frac{1}{N^2}) - f(i)]). \end{aligned}$$

Si se hace tender $N \to \infty$ se tiene

$$\begin{aligned} \lim_{N\to\infty} L^N f(i) &= \sigma^2(\frac{N^2}{2}[\frac{1}{N}f'(i) + \frac{1}{N^2}\frac{f''(i)}{2}] + \frac{N^2}{2}[\frac{-1}{N}f'(i) + \frac{1}{N^2}\frac{f''(i)}{2}]) \\ &= \sigma^2\frac{N^2}{2}[\frac{f''(i)}{N^2}] \\ &= \sigma^2\frac{f''(i)}{2}. \end{aligned} \tag{2.8}$$

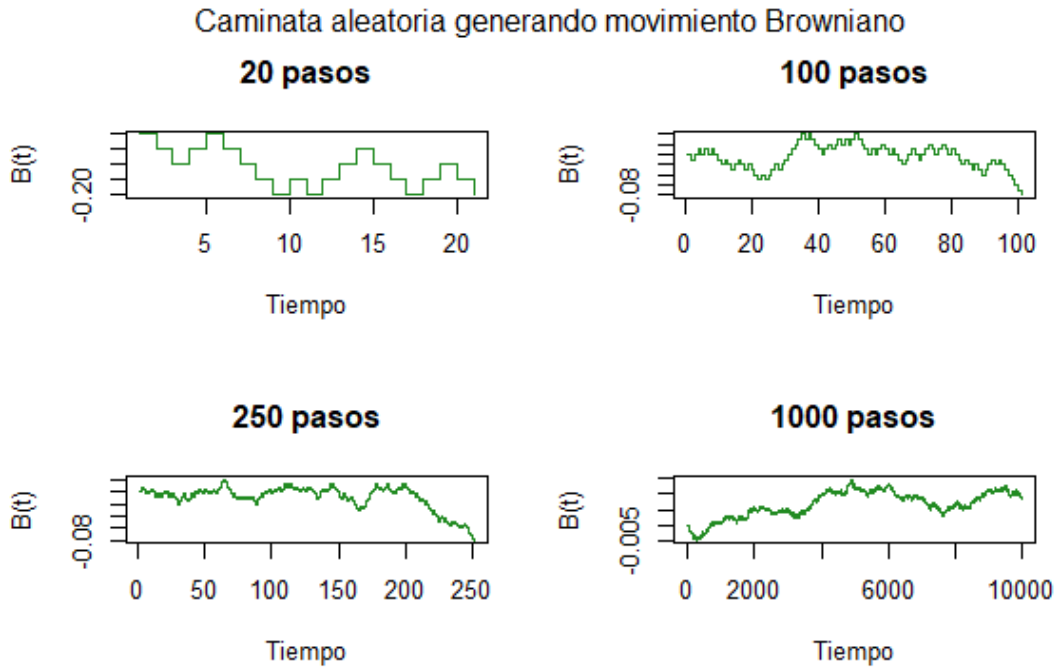


Figura 2.1: Trayectorias de caminatas aleatorias obtenidas mediante simulaciones en R

Así, el generador del movimiento Browniano B_t es $\sigma^2 \frac{f''(i)}{2}$.

Por otro lado, sea:

$$L^N f(i) = \mu(i) N([f(i + \frac{1}{N}) - f(i)] \tag{2.9}$$

el generador de un proceso Poisson $\frac{W_{N_t}}{N}$. Aplicando nuevamente el desarrollo de Taylor al término $f(i + \frac{1}{N})$ de 2.9 se tiene

$$f(i + \frac{1}{N}) = N[f(i) + \frac{1}{N} f'(i) + o(N^2) - f(i)].$$

Si $N \to \infty$ se tiene

$$\lim_{N \to \infty} L^N f(i) = \mu(i) f'(i), \tag{2.10}$$

donde $\lim_{N\to\infty} L^N f(i) = \mu(i)f'(i)$ es el generador de la ecuación diferencial $\frac{dg(t)}{d(t)} = \mu(t)$.

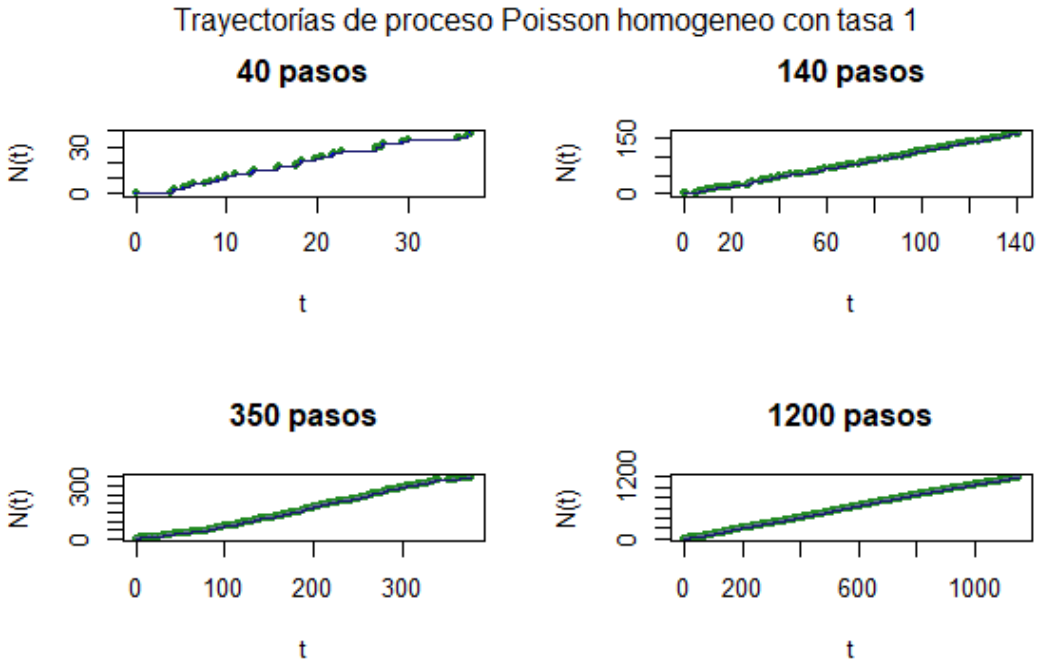


Figura 2.2: Trayectorias de proceso Poisson homogéneo obtenidas mediante simulaciones en R

En la figura 2.2 se tiene los resultados de simulaciones de trayectorias de proceso Poisson homogéneo (ver definición en 2.1.7) de tasa 1, se puede ver que a medida que se amplia la ventana de observación la trayectoria se aproxima más a la recta $y = x$, por lo que en el límite tendría naturaleza determinista.

Definición 2.2.1. *(Difusión unidimensional)*

Una difusión unidimensional $(X_t)_{t\geq 0}$ *es un proceso continuo de Markov con generador*

infinitesimal

$$Lf(x) = \frac{1}{2}\sigma^2(x)\frac{d^2f}{dx^2}(x) + \mu(x)\frac{df}{dx}(x) \tag{2.11}$$

donde a $\mu(x)$ *se denomina media infinitesimal y* $\sigma^2(x)$ *se denomina varianza infinitesimal.*

Para la definición anterior se deben tomar ciertos supuestos, en primer lugar, se puede probar que esta definición es equivalente a la que da Etheridge [Eth11] y que se encuentra al inicio de esta sección 2.2.

Una observación importante tras presentar la definición de una difusión es que el primer sumando del lado derecho de la expresión anterior es igual al obtenido en 2.8 mientras que el segundo sumando es igual al obtenido en 2.10 si se toma $t = x$. Si se vuelve a analizar los desarrollos que se hicieron para obtener 2.8 y 2.10 podrá notar que se desarrolló el comportamiento límite de una caminata aleatoria (que se sabe que es un movimiento Browniano) y de un proceso Poisson homogéneo (si $\mu \equiv 1$, que convergió a una línea recta). Es decir, se tiene un componente estocástico (ecuación diferencial estocástica) y un componente determinista (ecuación diferencial ordinaria).

En cualquier momento del tiempo, X_t es una variable aleatoria absolutamente continua, pero también cualquier realización de la difusión es una función continua del tiempo. El rango no necesariamente debe ser $\mathbb{R}$, y por el momento bastará con usar el intervalo (a, b) posiblemente con unión de uno o dos de los extremos a o b.

Además,se asume que los coeficientes $\mu(x)$ y $\sigma^2(x)$ son funciones continuas de $x \in (a, b)$ para tener existencia y unicidad en la ecuación diferencial estocástica.

Para una mejor interpretación de los coeficientes, se puede notar que si $f(x) = x$ es claro que $f'(x) = 1$ y $f''(x) = 0$, entonces

$$\frac{d}{dt}E_x X_t\Big|_{t=0} = \mu(x).$$

Mientras que si se fija x y se define $f(y) = (y - x)^2$ entonces $f'(x) = 0$ y $f''(y) = 2$, por lo que:

$$\frac{d}{dt} E_x(X_t - x)^2 \Big|_{t=0} = \sigma^2(x).$$

Por esta razón se conoce a $\mu(x)$ como media infinitesimal y a $\sigma^2(x)$ como varianza infinitesimal.

Ejemplo 2.2.2. *(Movimiento determinista)*

Supóngase que $X_0 = x$ y $\frac{dX_t}{dt} = \mu(X_t)$. Realizando algunas cuentas se tiene que

$$f(X_t) - f(X_0) = \int_0^t \frac{d}{ds} f(X_s) ds = \int_0^t f'(X_s) \frac{dX_s}{ds} ds = \int_0^t f'(X_s)\mu(X_s) ds$$

por lo que si f' y μ son continuos se tiene que

$$\frac{f(X_t) - f(x)}{t} \to f'(x)\mu(x),$$

por lo tanto $Lf = \mu(x)f'(x)$. Entonces, si $\sigma^2(x) = 0$, una difusión se reduce a una ecuación diferencial y de manera inversa, las difusiones generalizan a las ecuaciones diferenciales.

Ejemplo 2.2.3. *(Movimiento Browniano)*

Supóngase $B(0) = x$ y para $0 = t_0 < t_1 < ... < t_n$, $B(t_1) - B(t_0)$, los incrementos $B(t_2) - B(t_1), ..., B(t_n) - B(t_{n-1})$ son independientes entre si, con $B(t_i) - B(t_{i-1})$ distribuidos normalmente con media $\mu = 0$ y varianza $\sigma^2(t_i - t_{i-1})$. Si se hace el desarrollo de Taylor, al tomar una t pequeña se tiene:

$$f(B_t) - f(B_0) \approx f'(B_0)(B_t - B_0) + \frac{1}{2} f''(B_0)(B_t - B_0)^2$$

y al tomar valores esperados:

$$E_x[f(B_t) - f(x)] \approx \frac{1}{2} f''(x)\sigma^2 t$$

Por lo que $Lf(x) = (\frac{\sigma^2}{2})f''(x)$.

Las difusiones unidimensionales son útiles pues pueden usarse para calcular cantidades explícitamente, ya que (casi siempre) las difusiones unidimensionales pueden ser transformadas en *Movimiento Browniano*, primero con una transformación del espacio, y después con un ajuste en la escala del tiempo.

Definición 2.2.4. *(Función de escala)*

Para una difusión X_t *en un intervalo* (a, b) *con deriva* μ *y varianza* σ^2*, la función de escala se define como:*

$$S(x) = \int_{x_0}^{x} \exp\Big(- \int_{\nu}^{y} \frac{2\mu(z)}{\sigma^2(z)} dz dy \Big) \tag{2.12}$$

donde x_0 *y* ν *son puntos fijados arbitrariamente en* (a, b).

Definición 2.2.5. *(Escala natural)*

Se dice que una difusión está en escala natural si $S(x)$ *puede tomarse como una función lineal.*

El cambio de escala $S(X_t) = X_t$ resulta en un movimiento Browniano con cambio de tiempo en $(S(A), S(B))$. El cambio de tiempo requerido para transformar esto en un movimiento Browniano estándar se logra con la siguiente función.

Definición 2.2.6. *(Medida de velocidad)*

La función $m(t) = \dfrac{1}{\sigma^2(t)S'(t)}$ *es la densidad de la velocidad del procesos* X_t*. La medida de velocidad* M *esta dada por:*

$$M(x) = \int_{x_0}^{x} m(t)dt.$$

Teorema 2.2.7. *Denotando a la función de escala con* S *y a la medida de velocidad con* M *se tiene que el generador infinitesimal* L *de cierto proceso de difusión puede verse como:*

$$Lf = \frac{1}{2}\frac{1}{dM/dS}\frac{d^2f}{dS^2} = \frac{1}{2}\frac{d}{dM}\frac{df}{dS}$$

Prueba

$$\begin{aligned}
\frac{1}{2}\frac{d}{dM}\frac{df}{dS} &= \frac{1}{2}\frac{1}{dM/dx}\frac{d}{dxdx}\frac{1}{dS/dx}\frac{df}{dxdx} \\
&= \frac{1}{2}\sigma^2(x)S'(x)\frac{d}{dx}\frac{1}{S'(x)}\frac{df}{dx} \\
&= \frac{1}{2}\sigma^2(x)\frac{d^2f}{dx^2} - \frac{1}{2}\sigma^2(x)S'(x)\frac{S''(x)}{(S'(x))^2}\frac{df}{dx} \\
&= \frac{1}{2}\sigma^2(x)\frac{d^2f}{dx^2} + \mu(x)\frac{df}{dx}
\end{aligned}$$

pues S resuelve $LS = 0$.

Una pregunta que surge en varias aplicaciones de las difusiones es: "¿Cuál es la probabilidad de que la difusión toque 0 antes que 1?". Primero se verá el caso del Movimiento Browniano.

Teorema 2.2.8. *Sea* $\{W_t\}_{t\geq 0}$ *un movimiento Browniano estándar, para cada* $y \in \mathbb{R}$*, sea* T_y *el tiempo aleatorio en el cual se llega al estado* y *por primera vez, entonces para* $a < x < b$ *se tiene que:*

$$\mathbb{P}(T_a < T_b | W_0 = x) = \frac{b-x}{b-a}.$$

Prueba *(Heurística)* Sea $u(x) = \mathbb{P}(T_a < T_b | W_0 = x)$ y se asumirá que $\mathbb{P}(\text{mín}(T_a, T_b) < h | W_0 = x) = o(h)$ conforme $h \to 0$. Si además se supone que u es suficientemente suave entoces se puede usar la propiedad de Markov

$$\begin{aligned} u(x) &= \mathbb{E}(u(W_h)|W_0 = x) + o(h) \\ &= \mathbb{E}(u(x) + (W_h - x)u'(x) + \frac{1}{2}(W_h - x)^2 u''(x)) + o(h) \\ &= u(x) + \frac{1}{2}hu''(x) + o(h). \end{aligned}$$

Se tiene entonces que $\frac{1}{2}hu''(x) + o(h) = 0$, si se divide entre h y se hace tender h a 0 se obtiene que $u''(x) = 0$. Además se tienen las condiciones en la frontera $u(a) = 1$ y $u(b) = 0$, de modo que se obtiene $u(x) = \dfrac{b-x}{b-a}$. $\square$

Pasando al caso general, se tiene el siguiente resultado.

Teorema 2.2.9. *Sea $\{X_{t_{\geq}0}\}$ una difusión unidimensional en (a, b) con media infinitesimal $\mu(x)$ y varianza $\sigma^2(x)$. Si $a < a_0 < x < b_0 < b$ entonces escribiendo T_y como el primer momento en el cual $X_t = y$ se tiene que:*

$$\mathbb{P}(T_{a_0} < T_{b_0} | X_0 = x) = \frac{S(b_0) - S(x)}{S(b_0) - S(a_0)}$$

donde S es la función de escala de la difusión.

Prueba Es suficiente considerar los correspondientes tiempos de arribo de probabilidades del proceso $Z_t = S(X_t)$, donde S es la función de escala. El proceso Z_t es un movimiento Browniano con tiempo cambiado, pero como solo es relevante (para este teorema) la posición y no el tiempo en el que toca $(S(a_0), S(b_0))$, entonces se necesita determinar únicamente las probabilidades de arribo para el movimiento Browniano y el resultado se sigue del teorema anterior.

2.3. Simulación estocástica

Esta sección esta basada en el libro de Gentle [Gen10], así como en el curso de Simulación Estocástica impartido por Sergio López en la Facultad de Ciencias durante el semestre 2018-1 [LB].

La simulación estocástica se ha convertido en una técnica muy popular en los últimos años, pues puede ser una excelente aproximación para analizar problemas complejos. De forma intuitiva, el objetivo de la simulación es reproducir de forma artificial un fenómeno real. En el contexto de este trabajo, con la ayuda de una computadora se recrearán condiciones aleatorias para estudiar la dinámica de los modelos que se presentarán en el siguiente capítulo, y se usarán los resultados obtenidos para dar ejemplos de los resultados teóricos.

2.3.1. Números aleatorios

La simulación estocástica parte de la generación de números pseudoaleatorios (se agrega el prefijo *pseudo* pues aunque los resultados de las simulaciones son aproximaciones a una verdadera colección de números provenientes de una distribución uniforme, el proceso para obtenerlos sigue un método determinista). Esto puede hacerse con distintos métodos (principalmente generadores congruenciales lineales). Una vez que se tiene un generador de números aleatorios deben realizarse pruebas estadísticas de bondad de ajuste (para garantizar la uniformidad) y de aleatoriedad (para garantizar la independencia). Como ejemplo, si se realizan dichas pruebas al generador incluido en Excel 2016, se puede ver que no se cumple la aleatoriedad. La generación de números que se comporten como uniformes independientes es vital, pues se toman como base para la simulación de variables aleatorias. En este trabajo se utilizará (para las secciones de simulación) el generador de números aleatorios que viene

incluido en R.

2.3.2. Simulación de variables aleatorias y procesos estocásticos

Casi todos los algoritmos para generar números de distribuciones específicas se basan en la generación de números pseudoaleatorios.

2.3.2.1. Distribuciones discretas

Se asumirá que se quieren generar valores de una variable aleatoria X que sigue una distribución dada por $f_r = \mathbb{P}(X = r), r = 1, 2, ...$ y $F(r) = \mathbb{P}(X \leq r)$. Cualquier distribución en un conjunto numerable puede ser reducido a la forma anterior reetiquetando los puntos. Para la mayoría de distribuciones discretas, el método por excelencia es el de *inversión*: para una distribución discreta se tiene $F^{-}(u) = \text{mín}\{x|F(x) \geq u\} = i$ donde $F_{i-1} < u \leq F_i$, por lo que el método de inversión equivale a buscar en una tabla de la distribución F_i para un índice i adecuado.

Algorithm 1 Distribuciones discretas

Require: Función de distribución F.

Ensure: Valor X de distribución F.

1: Generar $U \sim U(0, 1)$, Sea $i = 1$.
2: **while** $F(i) \leq U$ **do**
3: $\quad i = i + 1$
4: **end while**
5: **Return** $X = i$.

En el paso 2, el número esperado de comparaciones es $\mathbb{E}(X)$, ya que i comparaciones son hechas para $X = i$. El algoritmo puede volverse más rápido reordenando las (f_r) en

orden descendente, esto reduce la $\mathbb{E}(X)$ lo máximo posible, y la distribución original se puede recuperar. Una mejor manera de reducir el número de comparaciones es empezar la búsqueda en un valor más adecuado, si f_r es unimodal se puede buscar a la izquierda o a la derecha de la moda.

2.3.2.2. Distribuciones absolutamente continuas

De estadística no paramétrica se sabe que si X tiene una función de distribución F continua entonces $F(X) \approx U(0,1)$. De aquí se puede tomar la idea de obtener números de F por $X = F^{-1}(U)$ si se prueba que la inversa existe.

Teorema 2.3.1. *Se define F^- como:*

$$F^-(u) = \text{mín}\{x|F(x) \geq u\}.$$

Entonces, si $U \approx U(0,1)$ se tiene que $X = F^-(U)$ es una muestra de F.

Prueba El mínimo se alcanza porque F es continua por la derecha, entonces

$$F(F^{-1}(u)) \geq u$$

y

$$F^-(F(x)) = \text{mín}\{y|F(y) \geq F(x)\} \leq x.$$

Por lo tanto

$$\{u, x|F^-(u) \leq x\} = \{(u,x)|u \leq F(x)\}$$

y

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^-(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

Con esto, se tiene el siguiente algoritmo:

Algorithm 2 Método de la función inversa

Require: Función de distribución F.

Ensure: Valor X de distribución F.

1: Generar un número aleatorio $U \sim U(0,1)$.

2: Tomar $X = F^{-}(U)$.

3: **Return** X.

2.3.2.3. Cadenas de Markov

Para generar una trayectoria de una cadena de Markov a tiempo discreto basta tener una distribución inicial $\pi^{(0)}$ y una matriz de transición P. Primero se obtiene el valor de X_0 desde $\pi^{(0)}$, entonces, dado $X_0 = x_0$, se genera X_1 con la distribución condicional de X_1 dado $X_0 = x_0$ se genera X_2 con la fila x_1-ésima de P, y así se continua. El algoritmo se detalla a continuación:

Algorithm 3 Simulación de una trayectoria de una Cadena de Markov

Require: Matriz de transición P, distribución inicial π_0.

Ensure: Vector X con la trayectoria de la cadena de Markov.

1: Generar X_0 de la distribución inicial $\pi^{(0)}$,

2: Hacer $n = 0$.

3: Generar X_{n+1} de la distribución correspondiente a la X_t-ésima fila de P.

4: Hacer $t = t + 1$ y regresar al paso 3.

5: **Return** X.

Para las cadenas de Markov a tiempo continuo también es posible simular trayectorias. Supóngase que $X = \{X_t : t \geq 0\}$ es una cadena de Markov a tiempo continuo con probabilidades de transición $\{P_{ij}(t)\}$ (probabilidad de que dentro de t unidades de tiempo la

cadena este en el estado j, dado que en este momento está en el estado i). La trayectoria de una cadena de Markov a tiempo continuo puede describirse por una matriz de transición $P = (P_{ij})$ que describe como la cadena cambia de estado en cada paso en el que ocurre una transición, junto con un conjunto de tasas $\{\lambda_i : i \in E\}$) que son los tiempos de espera. Cada vez que el estado $i \in E$ es visitado, la cadena pasa, en promedio, $\frac{1}{\lambda_i}$ unidades de tiempo antes de volver a moverse. Con base en lo anterior se tiene el siguiente algoritmo para generar una trayectoria a tiempo $t = T$:

Algorithm 4 Simulación de una trayectoria de una Cadena de Markov

Require: Matriz de transición P, distribución inicial π_0, tasas λ_i.

Ensure: Vector X con la trayectoria de la cadena de Markov a tiempo continuio.

1: Se escoge un valor inicial, $X_0 = i_0$. Se hace $n = 0$ y $t = t_0 = 0$.
2: Se genera $H_{i_0} \sim \exp(\lambda_{i_0})$. Se hace $t = t_1 = H_{i_0}$.
3: **if** $t < T$ **then**
4: Se hace $i = X_n$, se genera Y_i, se hace $n = n+1$, $i = Y_i$, $X_n = i$, se genera $H_i \sim \exp(\lambda_i)$ y se hace $t = t + H_i$, $t_n = t$
5: **else**
6: **Return** X
7: **end if**
8: Se regresa al paso 4.

Haciendo $N(t) = \text{máx}(n : t_n \leq T)$ se tiene el número de transiciones durante $(0, T]$, el algoritmo genera todos los valores de X_n, $0 \leq n \leq N(T)$, y los correspondientes tiempos $t_1, t_2, ..., t_n$ en los cuales la cadena hace cada transición.

2.3.3. Difusiones y ecuaciones diferenciales estocásticas

Para la simulación de difusiones se requiere simular ecuaciones diferenciales estocásticas, pues las difusiones pueden verse como soluciones a las ecuaciones antes mencionadas. Tomando como base el trabajo de [KP95], se tienen dos métodos principales para lograr dicha simulación.

2.3.3.1. Método de Euler-Maruyama

El método de Euler-Maruyama es una simple generalización del método de Euler para ecuaciones diferenciales ordinarias, considerando una ecuación diferencial estocástica dada por

$$dX_t = \mu(X_t)d_t + \sigma(X_t)dW_t \tag{2.13}$$

con condición inicial $X_0 = x_0$, donde W_t es un movimiento Browniano estándar, y asumiendo que se quiere resolver esta ecuación en un intervalo de tiempo $[0, T]$.

El método consiste en dividir el intervalo $[0, T]$ en n subintervalos de igual longitud $\Delta t = I > 0$: $0 = t_0 < t_1 < ... < t_n = T$ y $tk = k\Delta t$, posteriormente se define $Y_0 = x_0$ y se prosigue de manera recursiva para obtener los siguientes valores de Y_n como:

$$Y_{n+1} = Y_n + \mu(Y_n)\Delta t + \sigma(Y_n)\Delta W_n$$

donde

$$\Delta W_n = W_{t_{n+1}} - W_{t_n}.$$

Por propiedades del movimiento Browniano se sabe que las variables aleatorias ΔW_n son independientes e idénticamente distribuidas normal con media 0 y varianza Δt.

2.3.3.2. Método de Milstein

Considerando la misma ecuación dada en 2.13. La aproximación de Milstein a la solución real X_t es el proceso Y_t que se construye de la siguiente forma. Como en el método de Euler, se divide el intervalo $[0, T]$ en n intervalos de igual longitud $\Delta t > 0$: $0 = t_0 < t_1 < ... < t_n = T$ con $t_n = n\Delta t$ y $\Delta t = \frac{T}{n} > 0$, posteriormente se define $Y_0 = x_0$ y se definen los valores de Y_t para $1 \leq s \leq n$ como:

$$Y_{s+1} = Y_s + a(Y_s)\Delta t + b(Y_s)\Delta W_s + \frac{1}{2}b(Y_s)b'(Y_s)((\Delta W_s)^2 - \Delta_t)$$

donde b' denota la derivada de $b(x)$ respecto a x y $\Delta W_n = W_{t_{n+1}} - W_{t_n}$ donde las variables aleatorias ΔW_n son independientes e idénticamente distribuidas normal con media 0 y varianza Δt. Entonces Y_n aproximará a X_{t_n} para $0 \leq s \leq n$ y si se incrementa n se tendrá una mejor aproximación.

Observación. *Cuando* $b'(Y_s) = 0$ *este método es equivalente a el método de Euler Maruyama.*

2.3.4. Estimadores de Monte Carlo

Una vez que se tienen herramientas para simular variables y procesos aleatorios, se buscará encontrar una forma de encontrar estimadores para dichas variables.

Supóngase que se tiene la muestra $X_1, ..., X_n$ como resultado de n simulaciones de un experimento independientes e idénticamente distribuidas de acuerdo a alguna función de densidad f (conocida o no conocida). Esa muestra puede obtenerse al ejecutar n veces la rutina de la simulación, produciendo como resultado a X_i para la $i-$ésima ejecución. Además, el objetivo de la simulación es estimar la esperanza $\mu = \mathbb{E}(X)$, donde $X \sim f$. Si $|\mu| < \infty$, un

estimador insesgado por μ es la *media muestral* de $\{X_i\}$, esta es:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.14}$$

Si la varianza σ^2 de X es finita, entonces $\bar{X}$ tiene aproximadamente una distribución $\mathrm{N}(\mu, \sigma^2/n)$ para un valor de n grande (esto es una consecuencia del Teorema del límite central), por otro lado, si σ^2 es desconocida siempre es posible estimarla sin sesgo usando la *varianza muestral* de $\{X_i\}$.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}[X_i - (\bar{X})]^2 \tag{2.15}$$

que tiende a σ^2 conforme $n \to \infty$. Así, se tiene un *intervalo de confianza* aproximado para μ:

$$\left(\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \quad \bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) \tag{2.16}$$

donde z_γ denota al cuantil γ de una distribución $\mathrm{N}(0,1)$, aunque en vez de dar un intervalo de confianza, se suele reportar el estimador del error estándar: $\dfrac{S}{\sqrt{n}}$, o el estimador del error relativo: $\dfrac{S}{\bar{X}\sqrt{n}}$.

Definición 2.3.2. *Estimador de Monte Carlo Crudo*

Dada una colección de resultados $X_1, ..., X_n$ de una simulación de una distribución X, el estimador de Monte Carlo crudo se obtiene por:

$$\bar{\theta} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2.17}$$

Y el algoritmo relacionado es el siguiente:

Algorithm 5 Estimador de Monte Carlo crudo

Require: Valor final n, generador de distribución X.

Ensure: Valor de $\bar{\theta}$

1: Se generan $x_1, ..., x_n$ de forma independiente $\sim X$.

2: Se hace $\bar{\theta} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

A veces se tiene que X sea función de algún vector aleatorio o proceso estocástico, es decir, que $X = g(\mathbf{Y})$, donde g es una función real y $\mathbf{Y}$ es un vector o proceso aleatorio. La utilidad del estimador de Monte Carlo es que 2.16 se sigue cumpliendo sin importar la dimensión de $\mathbf{Y}$.

2.4. Gráficas

Esta sección esta basada en el capítulo 1 del texto de Bondy y Morty [BM10]. Aunque a primera instancia parece tener poca relación con el tema, fue incluida debido a un modelo reciente que se consideró relevante. Primero, se empezará por definir que es una gráfica, en matemáticas discretas.

Definición 2.4.1. *(Gráfica simple)*

Una gráfica simple es un par ordenado $(V(G), E(G))$ conformado por un conjunto $V(G)$ de vértices y un conjunto $E(G)$, disjunto de $V(G)$, de aristas de G.

Se dice que una arista $\{v, w\}$ está *unida* a los vértices v y w y suele usarse la notación vw. En cualquier gráfica simple hay a lo más una arista uniendo un par dado de vértices. Al número de vértices y aristas de G suele representarse por $\#v(G)$ y $e(G)$ respectivamente; estos parámetros se denominan *orden* y *tamaño* de G respectivamente.

Las gráficas, en este contexto, reciben su nombre porque pueden ser representadas en un diagrama que suele ayudar a entender sus propiedades: cada vértice es representado por un punto o nodo, y cada arista es representada por una línea que une los puntos que corresponden a los extremos de la arista.

Además se tienen ciertas convenciones, se dice que los extremos de un arista son *incidentes* con la arista y que la arista es *incidente* con sus extremos. Dos vértices que son incidentes con una arista en común son *adyacentes*, al igual que dos aristas que sean incidentes con un vértice en común, dos vértices adyacentes que sean vecinos se denominan *vecinos* y al conjunto de vecinos de un vértice v suele representarse por $N_G(v)$.

Una arista con extremos idénticos se llama *bucle*, mientras que si son distintos se denomina *enlace*. Dos o más enlaces con el mismo par de extremos se denominan *aristas paralelas*.

2.4.1. Gráficas aleatorias

A continuación, se presentan dos definiciones para introducir el concepto de *gráfica aleatoria*.

En el primero, se tiene un enfoque de muestreo que estipula que una gráfica aleatoria es una gráfica que es obtenida tras ser elegida de forma aleatoria de una colección de gráficas. Esta colección puede caracterizarse por ciertos parámetros de la gráfica que tienen valores fijos.

Definición 2.4.2. *Gráfica aleatoria (enfoque de muestreo)*

$G(n, m)$ es una gráfica obtenida luego de elegir uniformemente de todas las gráficas de n vértices y m aristas.

La probabilidad de elegir a la gráfica $G(n, m)$ requiere saber el tamaño del conjunto de todas las gráficas resultantes, calculando todas las posibles combinaciones de m aristas de todos los posibles pares de vértices n. El número total de gráficas aleatorias posibles dados n vértices y m aristas es:

$$|\Omega| = \frac{1}{\binom{\binom{n}{2}}{m}}.$$

En el segundo enfoque, se tiene un enfoque probabilístico.

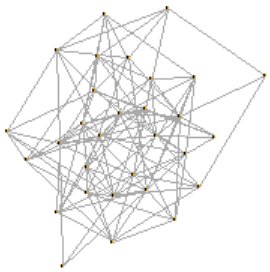
Definición 2.4.3. *Gráfica aleatoria (enfoque constructivo)*

Para el enfoque constructivo, se dice que una gráfica aleatoria es aquella donde la presencia y colocación de sus aristas sigue una distribución aleatoria.

Se considera un conjunto de vértices $V = \{1, 2, 3, ..., n\}$ para la construcción de una gráfica aleatoria y se continua eligiendo uniformemente de forma aleatoria un borde del conjunto de aristas que no hayan sido elegidas aún, repitiendo este proceso m veces.

Ejemplo 2.4.4. *La gráfica aleatoria de Erdös-Renyi $G(n, p)$ es una gráfica aleatoria obtenida iniciando con el conjunto de vértices $V = \{1, 2, 3, ..., n\}$, haciendo $0 \leq p \leq 1$ y conectando cada par de vértices $\{i, j\}$ mediante una arista con probabilidad p.*

Con la función **erdos.renyi.game** del paquete **igraph** de R se obtuvo la siguiente gráfica aleatoria correspondiente a un modelo de Erdös-Renyi$(30, 0.3)$.



Gráfica aleatoria con 30 nodos y probabilidad de arista p=0.3

Figura 2.3: Gráfica aleatoria de Erdös-Renyi

Ejemplo 2.4.5. *Gráfica aleatoria de Watts-Strogatz*

La gráfica de Watts-Strogatz representa el fenómeno de mundo pequeño, también conocido como seis grados de separación. Dos individuos en la Tierra estan a sólo 6 *personas de distancia uno de otro. La gráfica tiene una estructura inicial de vértices con aristas a sus* k *vecinos más cercanos. Cada arista tiene probabilidad* p *de que sea realmente colocada.*

Con la función **watts.strogatz.game** del paquete **igraph** de R se obtuvo la siguiente gráfica aleatoria correspondiente a un modelo de Watts-Strogatz($k = 500, p = 0.35$).

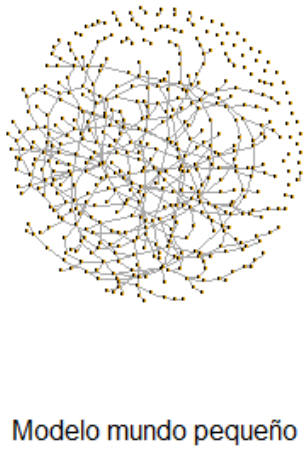


Figura 2.4: Gráfica aleatoria de Watts-Strogatz, $k = 500$, $p = 0.35$

Capítulo 3

Modelos de deriva génica

En este capítulo se introducirán los modelos clásicos de deriva génica, se podrá ver que son muy similares y que, de hecho, coinciden en cierto sentido. Las siguientes secciones se basan en los textos de Durrett [Dur10] (capítulo 1) y Etheridge [Eth11] (capítulo 1).

3.1. Modelo de Wright-Fisher

En el modelo clásico de Wright-Fisher, se asume la existencia de una población de alelos de tamaño $2N$, en la población sólo hay dos tipos de alelos: A y a. Esta población tiene como característica que no se traslapan las generaciones. En la generación n, de los $2N$ alelos que se tienen hay i que son de tipo A y $2N - i$ que son de tipo a. Para construir la generación $n + 1$ se usa la generación anterior n, para lo que se realizan $2N$ muestreos con reemplazo.

Definición 3.1.1. *(Proceso de frecuencias asociado al Modelo de Wright-Fisher)*

El proceso de frecuencias está definido por: $Y_n = \dfrac{X_n}{2N}$, *donde* X_n *es el número de alelos de tipo* A *en la generación* n; Y_n *es una cadena de Markov con espacio de estados* $E =$

$\{0, \frac{1}{2N}, ..., \frac{2N-1}{2N}, 1\}$ *y con probabilidades de transición:*

$$P_{\frac{i}{[2N]},\frac{k}{[2N]}} = \bar{P}_{i,k} = \binom{2N}{k} \bar{P}_i^k((1-\bar{P}_i)^{2N-k}) \tag{3.1}$$

con P la matriz de transición asociada a Y_n y $\bar{P}$ la matriz de transición asociada a X_n.

En 3.1, se tiene que $\frac{i}{[2N]}, \frac{k}{[2N]} \in E$, $P_i = \dfrac{i}{2N}$ es la probabilidad de elegir un alelo de tipo A en la población en un intento, mientras que $\binom{2N}{k} = \dfrac{(2N)!}{k!(2N-k)!}$ es el número de formas de elegir k alelos de un tipo, de un total de $2N$. Es decir, sigue una distribución binomial.

La cadena de Markov asociada al proceso de frecuencias presenta dos estados absorbentes: 0 y 1, por lo que si existe una generación τ en la cual Y_τ llegue a alguno de esos estados nunca saldrá de ahí. Biológicamente esto representa la extinción del alelo A (si $Y_\tau = 0$) o del alelo a (si $Y_\tau = 1$).

Sea $\tau = \text{mín}\{n : Y_n = 0 \text{ o } Y_n = 1\} = \text{mín}\{n : X_n = 0 \text{ o } X_n = 2N\}$ el *tiempo de fijación*, es decir, el primer tiempo en el cual la población se conforma de alelos de únicamente un tipo. La notación que se usa es: $\mathbb{P}_i$ como la función de probabilidad del proceso Y_n empezando desde $Y_0 = i$, y $\mathbb{E}_i$ denotará al valor esperado con respecto a $\mathbb{P}_i$. La probabilidad de fijación puede calcularse (de acuerdo a Durrett [Dur10]) considerando el proceso de conteo de un tipo de alelo ($X_t = \#$ de alelos de tipo A en la generación n).

Teorema 3.1.2. *En el modelo de Wright-Fisher, la probabilidad de fijación en todos los estados de A, es*

$$\mathbb{P}_i(X_\tau = 2N) = \frac{i}{2N} \tag{3.2}$$

Prueba Como el número de individuos es finito, y siempre es posible elegir entre todos los alelos de tipo A, o todos los alelos de tipo a la fijación ocurrirá casi seguramente. Sea X_n el

número de alelos de tipo A a tiempo n. Como la esperanza de la binomial en 3.1 es $2Np$, se sigue que

$$\mathbb{E}(X_{n+1}|X_n = i) = 2N\Big(\frac{i}{2N}\Big) = i = X_n.$$

Tomando valor esperado, se tiene que $\mathbb{E}_i(X_\tau) = \mathbb{E}_i(X_n)$, (esto se sigue de que X_τ es acotada, así como el teorema del paro opcional) es decir, el valor promedio de X_n permanece constante en el tiempo, pues la cadena es homogénea en el tiempo.

La propiedad anterior implica que

$$i = \mathbb{E}_i(X_\tau) = 2N\mathbb{P}_i(X_\tau = 2N).$$

De donde se puede despejar $\mathbb{P}_i(X_\tau = 2N)$ para obtener la formula deseada. Para probar esto hay que notar que como $X_n = X_\tau$ c.s cuando $n > \tau$,

$$i = \mathbb{E}_i(X_n) = \mathbb{E}_i(X_\tau; \tau \leq n) + \mathbb{E}_i(X_n; \tau > n),$$

dónde $\mathbb{E}(X; S)$ es la esperanza de X sobre el conjunto S. Haciendo tender $n \to \infty$ y usando el hecho de que $|X_n| \leq 2N$ se puede concluir que el primer termino converge a $\mathbb{E}_i(X_\tau)$, y el segundo a 0.

3.1.1. El modelo de Wright-Fisher es una martingala

Otro aspecto interesante del modelo de Wright-Fisher es que presenta la propiedad de martingala, recordando, un proceso estocástico Y_n cumple la propiedad de martingala si se preservan las siguientes tres condiciones:

- Es integrable.
- Es adaptado.

- Cumple que $E(Y_{n+1}|Y_n, Y_n - 1, ..., Y_0) = Y_n$.

En el caso del proceso de frecuencias asociado al modelo de Wright-Fisher la integrabilidad se tiene pues, primero, siempre se tiene que $Y_n \geq 0$ ya que el espacio de estados tiene valores no negativos y entonces $|Y_n| = Y_n$. Luego,

$$\begin{aligned} \mathbb{E}(Y_n) = \mathbb{E}(\mathbb{E}(Y_n|Y_{n-1})) = \mathbb{E}(2N\frac{Y_{n-1}}{2N}) \\ = \mathbb{E}(Y_{n-1}) = \mathbb{E}(\mathbb{E}(Y_{n-1}|Y_{n-2})) = \mathbb{E}(2N\frac{Y_{n-2}}{2N}) \\ = \mathbb{E}(Y_{n-2}) = ... = Y_0. \end{aligned}$$

Es decir, el valor esperado del modelo de Wright-Fisher en cualquier punto del tiempo es únicamente la frecuencia de alelos de tipo A en la primera generación.

Por otro lado, la adaptabilidad del modelo se tiene pues el proceso es una variable aleatoria con distribución conocida.

La condición técnica se cumple ya que, por definición $(Y_{n+1}|Y_n, ..., Y_0) = (Y_{n+1}|Y_n) \sim$ Binomial$(1, Y_n)$. Por lo tanto, la esperanza de $(Y_{n+1}|Y_n)$ es conocida (el producto de los parámetros, ya que sigue una distribución binomial) y se tiene que $\mathbb{E}(Y_{n+1}|Y_n) = Y_n$.

3.1.2. Simulación del modelo de Wright-Fisher

En el sitio del profesor Marcus Stephens [Ste16] de la Universidad de Chicago se encuentra un excelente resumen del modelo de Wright-Fisher, así como una forma de implementarlo en R, cuyo código se encuentra en los anexos.

En la figura 3.1 se tienen varios casos de trayectorias del modelo de Wright-Fisher, las columnas indican la frecuencia inicial de alelos, donde se tomaron valores igual a $0.1, 0.3, 0.6$ y 0.98, mientras que las filas indican el tamaño de las poblaciones.

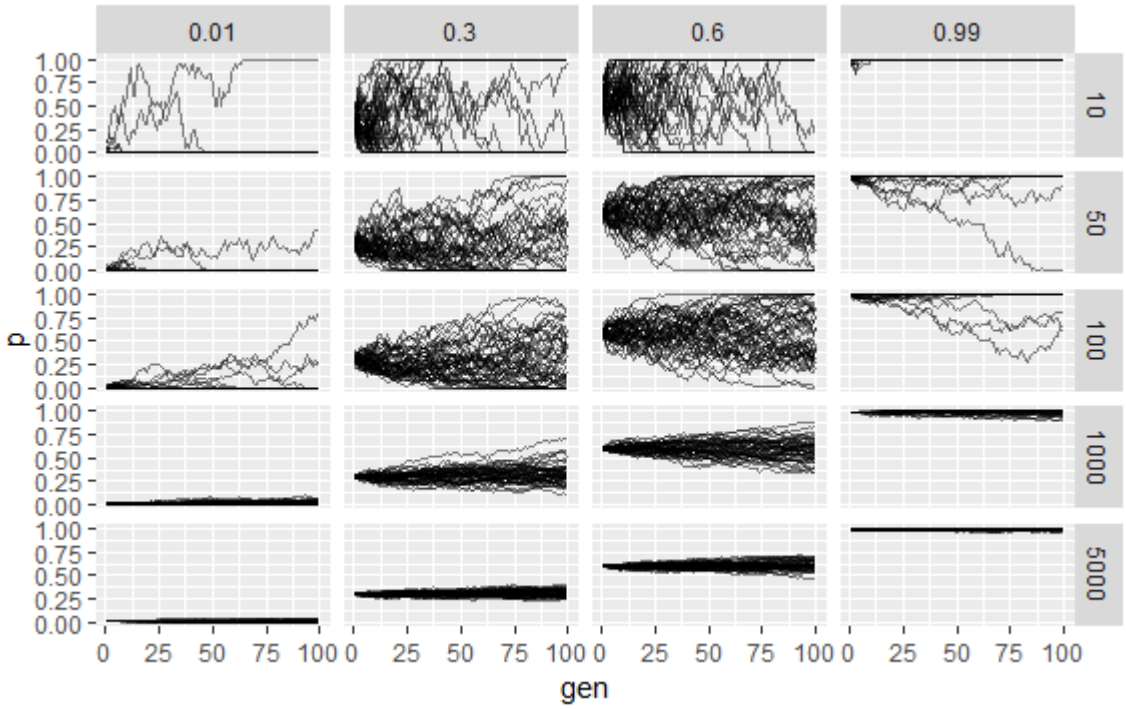


Figura 3.1: Modelo de Wright-Fisher, simulación obtenida con R

Cada simulación tiene 100 generaciones y se hizo 50 veces, se puede ver que, cuando se tiene un número de individuos alto, no ocurre una fijación de un tipo de alelo durante el tiempo en el que se ejecutó la simulación, por otro lado, cuando se tiene una frecuencia inicial muy cargada hacía un tipo de alelo, la fijación podría llegar muy rápido.

En los casos de frecuencias de alelos iniciales no cargadas se tienen comportamientos similares de deriva génica, pero estos comportamientos van suavizándose conforme va aumentando el tamaño de la población.

3.2. Modelo de Moran

El Modelo de Moran es otro de los modelos clásicos en la genética de poblaciones. Al igual que en el modeo de Wright-Fisher, se asume una población de alelos de tamaño $2N$, donde, en el caso base, no existen migraciones ni selección natural, y en la cual los alelos se dividen en aquellos de tipo A o a. En esta población las generaciones se pueden traslapar, a tiempo t se tiene que de los $2N$ alelos que hay, i son de tipo A y $2N-i$ son de tipo a. La población va cambiando del siguiente modo: cada individuo es reemplazado a tasa exponencial con parámetro 1. Es decir, el individuo x vive por un tiempo distribuido de forma exponencial con media 1 y entonces es "reemplazado". Para "reemplazar" al individuo x, se escoge un individuo de forma aleatoria de la población (incluyendo al propio x) para ser el padre del nuevo individuo.

Se puede definir el proceso de conteo X_t como el número de copias del alelo A a tiempo t. Este proceso es de utilidad para definir el proceso de frecuencias $Y_t = \dfrac{X_t}{2N}$.

Definición 3.2.1. *(El proceso de frecuencias del modelo de Moran)*

El proceso de frecuencias $Y_t = \dfrac{X_t}{2N}$, (donde X_t es el número de copias del alelo A a tiempo t) es una cadena de Markov con espacio de estados $E = \left\{0, \dfrac{1}{2N}, ..., \dfrac{2N-1}{2N}, 1\right\}$ y probabilidades de transición dadas por:

$$\frac{i}{2N} \to \frac{i+1}{2N}, \ \textit{a tasa} \quad b_i = \frac{(2N-i)\cdot i}{2N}$$
$$\frac{i}{2N} \to \frac{i-1}{2N}, \ \textit{a tasa} \quad d_i = \frac{i\cdot(2N-i)}{2N}$$

donde b representa nacimientos y d representa muertes de los individuos de tipo A.

De manera intuitiva, las copias de alelos de tipo a son elegidas por reemplazo a tasa total de $2N-i$. El número de A se incrementará si un individuo de tipo A fue elegido para ser

el ancestro del nuevo individuo, un evento con probabilidad $\frac{i}{2N}$. El razonamiento para el segundo caso es análogo.

La cadena Y_t presenta dos estados absorbentes: 0 y 1. Por lo que existe un τ tal que una vez que $Y_\tau = 0$ o $Y_\tau = 1$, la cadena se quedará en ese estado, pues todos los individuos del universo serán del mismo tipo y al realizar el proceso de elección de padre no podrán cambiar de tipo.

Sea $\tau = \text{mín}\{t : X_t = 0 \quad \text{o} \quad X_t = 2N\} = \text{mín}\{t : Y_t = 0 \quad \text{o} \quad Y_t = 1\}$ el *tiempo de fijación* del proceso.

Teorema 3.2.2. *En el modelo de Moran, la probabilidad de que el tipo de alelo A sea fijado cuando inicialmente hay i copias es* $\frac{i}{2N}$.

Prueba Como las tasas para que aumenten o disminuyan las copias de alelos de tipo A son las mismas se sigue que $\frac{d}{dt}\mathbb{E}_i(X_t) = 0$. Además, se puede ver que

$$i = \mathbb{E}_i(X_t) = 2N\mathbb{P}(X_\tau = 2N, \tau < t) + \mathbb{E}_i(X_t; \tau > T).$$

Haciendo tender $t \to \infty$ se tiene que $\mathbb{P}_i(\tau > t) \to 0$, además, como $|X_t| \leq 2N$ se tiene que

$$i = \mathbb{E}_i(X_\tau) = 2N\mathbb{P}_i(X_\tau = 2N).$$

Al despejar en la última ecuación se llega a $\mathbb{P}(X_\tau = 2N) = \frac{i}{2N}$.

Capítulo 4

Selección y mutación

En esta sección se introducirá el concepto de selección a los modelos de deriva génica. Se dice que un alelo es selectivamente ventajoso (desventajoso) si es más (menos) propenso a ser elegido durante la reproducción.

4.1. Selección en el modelo de Wright-Fisher

En el modelo de Wright Fisher con selección se considera una población con tamaño constante $2N$, donde los alelos están divididos en dos tipos: a y A, si la generación n tiene i alelos de tipo A y $2N - i$ de tipo a, entonces, la generación $n + 1$ es formada mediante un muestreo independiente con reemplazo con las siguientes probabilidades.

$$\mathbb{P}(\text{ Elegir a } A) = \frac{i(1+s)}{i(1+s) + 2N - i}$$
$$\mathbb{P}(\text{ Elegir a } a) = \frac{2N - i}{i(1+s) + 2N - i}.$$

El parámetro s se denomina "coeficiente de selección".

- Si $s > 0$ se dice que A es un alelo "ventajoso".

- Si $s < 0$ se dice que A es un alelo "no ventajoso".

Además, se puede agregar mutación al modelo. Si se supone que durante el evento reproductivo, cada alelo de tipo a de la población muta a A con probabilidad μ_2 y cada individuo de tipo A muta a a con probabilidad μ_1. Entonces, la proporción de descendencia potencial de los alelos de tipo A tras la presencia de selección y mutación es:

$$\Psi_i = \frac{i(1+s)(1-\mu_1)}{i(1+s)+2N-i} + \frac{(2N-i)\mu_2}{i(1+s)+2N-i}.$$

Definición 4.1.1. *(Modelo de Wright-Fisher con selección y mutación)*

Si existen i alelos de tipo A en la generación n (y por complemento, $2N-i$ de tipo a), entonces bajo el modelo de Wright-Fisher con selección y mutación, el número de individuos de tipo A en la generación $n+1$ esta dado por una distribución $Bin(2N, \Psi_i)$.

Calcular explícitamente propiedades del modelo de Wright-Fisher con la definición anterior puede llegar a ser complicado, por lo que se suele usar una aproximación con la teoría de las difusiones. El tiempo se medirá en unidades de N generaciones y se considerarán las proporciones del alelo de tipo A en la población. Además se hacen los siguientes supuestos:

- $\alpha = 2Ns$
- $\nu_1 = 2N\mu_1$
- $\nu_2 = 2N\mu_2$.

Teorema 4.1.2. *Si se hace tender $2N \to \infty$, el modelo re-escalado de Wright-Fisher converge a la difusión unidimensional con media infinitesimal:*

$$\mu(p) = \alpha p(1-p) - \nu_1 p + \nu_2(1-p)$$

y varianza infinitesimal:

$$\sigma^2(p) = p(1-p).$$

Prueba Sea $\delta_t = \frac{1}{N}$ el tiempo entre dos generaciones, se quiere identificar la media infinitesimal así como la varianza, si la proporción actual de alelos tipo A es p, entonces el número actual de individuos de tipo a es $k \equiv 2Np$. Así se tiene:

$$\mathbb{E}[(p_{1/2N} - p | p_0 = p)] = \frac{1}{2N}(2N\Psi_i - i)$$

Sustituyendo:

$$\begin{aligned}
2N\Psi_i - i &= \frac{2Ni(1+\frac{\alpha}{2N})}{2(N+\frac{\alpha i}{2N})}(1-\frac{\nu_1}{2N}) + \frac{2N(2N-i)}{2(N+\frac{\alpha i}{2N})}\frac{\nu_2}{2N} - i \\
&= \frac{1}{2N+\frac{\alpha i}{2N}}\left[2Ni(1+\frac{\alpha}{2N})(1-\frac{\nu_1}{2N}) + (2N-i)\nu_2 - i2N - \frac{\alpha i^2}{2N}\right] \\
&= \frac{2N}{2N+\frac{\alpha i}{2N}}\left(\frac{\alpha i}{2N} - \frac{\nu_1 i}{2N} + \nu_2 - \frac{\nu_2 i}{2N} - \alpha\frac{i^2}{(2N)^2} - \alpha\frac{\nu_1 i}{(2N)^2}\right) \\
&= \alpha p - \nu_1 p + \nu_2 - \nu_2 p^2 + o(\frac{1}{2N}), \\
&= \alpha p(1-p) - \nu_1 p + \nu_2(1-p) + o(\frac{1}{2N}).
\end{aligned} \tag{4.1}$$

Para la varianza infinitesimal, como $\Psi_i = \frac{i}{2N} + o(\frac{1}{2N})$ se tiene que

$$\begin{aligned}
\mathbb{E}[(p_{\frac{1}{2N}} - p)^2 | p_0 = p] &= \frac{1}{(2N)^2}2N\Psi_i(1-\Psi_i) + o\Big(\frac{1}{(2N)^2}\Big) \\
&= \frac{1}{2N}p\left(1 - p + o\Big(\frac{1}{(2N)^2}\Big)\right).
\end{aligned} \tag{4.2}$$

Definición 4.1.3. *(Límite débil de selección)*

Al la difusión del teorema anterior se le conoce como límite de selección débil del modelo de Wright-Fisher con selección y mutación

Teorema 4.1.4. *(Probabilidades de fijación)*

Supóngase que no se tiene mutación ($\nu_1 = \nu_2 = 0$). Si la proporción inicial de alelos de tipo A es p_0, la probabilidad $p_{fix}(p_0) = \mathbb{P}(\{A \text{ sea fijado eventualmente en la población}\})$ (es decir, la difusión es absorbida en $p = 1$) es:

$$p_{fix}(p_0) = \begin{cases} \dfrac{1 - exp(-2\alpha p_0)}{1 - \exp(-2\alpha)}, & si\ \alpha \neq 0 \\ p_0, & si\ \alpha = 0 \end{cases}$$

Prueba Usando el teorema 2.2.9 para $\alpha \neq 0$.

$$\begin{aligned} p_{fix}(p_0) &= \frac{S(p_0) - S(0)}{S(1) - S(0)} \\ &= \frac{\int_0^{p_0} \exp(-\int_\eta^y \frac{2\mu(z)}{\sigma^2(z)} dz) dy}{\int_0^1 \exp(-\int_\eta^y \frac{2\mu(z)}{\sigma^2(z)} dz) dy} \\ &= \frac{\int_0^{p_0} \exp(-2\alpha y) dy}{\int_0^1 \exp(-2\alpha y) dy} \\ &= \frac{1 - \exp(-2\alpha p_0)}{1 - \exp(-2\alpha)}. \end{aligned} \tag{4.3}$$

Para $\alpha = 0$ la difusión se encuentra ya en la escala natural y el resultado se sigue del teorema 2.2.8.

Con los últimos dos teoremas, A. Etheridge considera algunos casos especiales considerando $p_0 = \frac{1}{2N}$.

1. Alelos no ventajosos: $s < 0$, si $|s| \ll 1$ ($a \ll b \to \dfrac{a}{b} \approx 0$) y $N_{|s|} \gg 1$, $p_{fix}(\frac{1}{2N} \approx 2|s| \exp(-2(2N)|s|)$. La probabilidad de fijación de un alelo no ventajoso es exponencialmente pequeña y decrece conforme se incrementa el tamaño de la población.

2. Alelos ventajoso: $s > 0$, $s \ll 1$, $N_s \gg 1$, entonces $p_{fix}(1/2N) \approx 2s$, casi independiente del tamaño de la población.

3. Alelos cercanos a ser neutrales. Si $N_{|s|} \ll 1$, entonces A es *casi* neutral y $p_{fix}(1/2N) \approx \frac{1}{2N}$

En resumen, la mayoría de los alelos (aquellos que inician con $p_0 = \frac{1}{2N}$) se pierden y el resto de fija.

4.2. Selección en el modelo de Moran

Considerando el modelo de Moran presentado en la sección anterior, se tienen $2N$ alelos, de los cuales i son de tipo A y $2N - i$ son de tipo a. Considerando nuevamente el proceso de frecuencias y asignando ventajas evolutivas relativas de los alelos A y a como 1 y $1 - s$ (por el momento se tomarán valores de s entre 0 y 1). Se pueden formular las siguientes probabilidades de transición para el Modelo de Moran con selección.

$$\begin{aligned} i \to i+1, \text{ a tasa} \quad b_i &= \frac{(2N-i)i}{2N} \\ i \to i-1, \text{ a tasa} \quad d_i &= \frac{i(2N-i)(1-s)}{2N}. \end{aligned}$$

En palabras, las copias de alelo de tipo a son elegidas por reemplazo a tasa total de $2N-i$, mientras que el número de copias de alelos de tipo A va a incrementarse si un alelo A es elegido como padre del nuevo individuo, esto sucede con probabilidad $\frac{i}{2N}$. Para la segunda tasa (d_i) se toma en cuenta que el reemplazo ocurre a tasa $1 - s$ dada que la desventaja evolutiva relativa de a es $1 - s$.

Como no hay mutación en el modelo, el alelo A se perderá o se fijará en la población. Sea $T_y = \text{mín}\{t : X_t = y\}$ el primer tiempo en el que la cadena llega a y. Sea $h(i) = \mathbb{P}_i(T_{2N} < T_0)$

la probabilidad de que el alelo A sea fijado cuando hay inicialmente i individuos de ese tipo, en el caso neutral se tiene que $h(i) = \dfrac{i}{2N}$.

Teorema 4.2.1. *En el modelo de Moran con selección $s > 0$.*

$$\mathbb{P}_i(T_{2N} < T_0) = \frac{1-(1-s)^i}{1-(1-s)^{2N}}. \tag{4.4}$$

Cuando $i = 1$, el numerador es únicamente s. Si la selección es fuerte ($2N$ es grande) entonces $(1-s)^{2N} \approx 0$ y la probabilidad de fijación es únicamente s. Si s es pequeña entonces $(1-s) \approx \exp^{-s}$ por lo que el resultado anterior puede escribirse como:

$$\mathbb{P}_i(T_{2N} < T_0) \cong \frac{1-\exp^{-is}}{1-\exp^{-2Ns}}. \tag{4.5}$$

Prueba Sea $\tau = \text{mín}(T_0, T_{2N})$. Cuando $s = 0$ se sabe que $\mathbb{E}(X_i)$ es constante en el tiempo, por lo que se tiene

$$\mathbb{E}(X_i) = 2N \cdot \mathbb{P}_i(X_\tau = 2N) + 0 \cdot \mathbb{P}_i(X_\tau = 0) = i$$

Despejando se tiene que $\mathbb{P}_i(X_\tau = 2N) = \frac{i}{2N}$. Cuando $s > 0$, $\dfrac{b_i}{b_i + d_i} = \dfrac{1}{2-s}$.

Haciendo un poco de álgebra se tiene que:

$$(1-s)^{i+1}\frac{1}{2-s} + (1-s)^{i-1}\frac{1-s}{2-s} = (1-s)^i\frac{1-s}{2-s} + (1-s)^i\frac{1}{2-s} = (1-s)^i, \tag{4.6}$$

por lo que, en este caso, el valor de $\mathbb{E}[(1-s)^{X_t}]$ permanece constante en el tiempo, por lo tanto es una martingala. Usando el mismo argumento que antes:

$$(1-s)^i = (1-s)^{2N}\mathbb{P}_i(X_\tau = 2N) + 1(1-\mathbb{P}_i(X_\tau = 2N)) \tag{4.7}$$

Despejando se tiene:

$$\mathbb{P}_i(X_\tau = 2N) = \frac{1-(1-s)^i}{1-(1-s)^{2N}}. \tag{4.8}$$

En el trabajo de Durrett (6.1.2) [Dur10] también se da razón del tiempo antes de la fijación de algún alelo, el siguiente resultado ilustra esta idea.

Observación. *En el modelo de Moran con selección $s > 0$, conforme $N \to \infty$ se tiene que*

$$\mathbb{E}_1(\tau | T_{2N} < T_0) \sim \frac{2}{s} log N \tag{4.9}$$

donde $a_N \sim b_N$ significa que $\frac{a_N}{b_N} \to 1$ conforme $N \to \infty$.

4.2.1. Convergencia a la difusión de Wright-Fisher

Considérese el modelo de Moran donde cada alelo de tipo A tiene una competitividad (fitness) de 1 mientras que la competitividad del tipo a es $1 - s$. Las mutaciones $a \to A$ ocurren a tasa μ_1 mientras que $A \to a$ ocurren a tasa μ_2. Por simplicidad se asume que las mutaciones ocurren durante la vida del individuo (es decir, no en el nacimiento), por lo que agregando las tasas de mutación a las tasas de transición antes mencionadas se tiene:

$$\begin{aligned} i \to i+1, \text{ a tasa} \quad b_i &= (2N-i)(\frac{i}{2N} + \mu_1) \\ i \to i-1, \text{ a tasa} \quad d_i &= i((2N-i)/(2N)(1-s) + \mu_2). \end{aligned} \tag{4.10}$$

Teorema 4.2.2. *(Convergencia del modelo de Moran a la difusión de Wright-Fisher)*

El proceso de frecuencia Y_t asociado al modelo de Moran con selección y mutación con tasas 4.10 converge a una difusión con $\mu(x) = (1-x)\beta_1 - x\beta_2 + x(1-x)\gamma$ y $\sigma^2(x) = x(1-x)$, donde $x = \frac{1}{2N}$, $\beta_i = N\mu_i$ y $\gamma = Ns$ y el proceso converge en el sentido de la convergencia de los generadores.

Sea Y_t la proporción de individuos de el alelo A, para obtener la aproximación a la difusión primero hay que notar que si $\frac{i}{2N} = x$ entonces:

$$\frac{d}{dt}\mathbb{E}(Y_t) = \frac{1}{2N}\left[(2N-i)(\frac{i}{2N} + \mu_1) - i(\frac{2N-i}{2N}(1-s) + \mu_2)\right].$$

Haciendo $\beta_i = N\mu_i$, y $\gamma = Ns$ se tiene que el coeficiente de media para el proceso que corre a tasa N es:

$$\mu(x) = (1-x)\beta_1 - x\beta_2 + x(1-x)\gamma. \tag{4.11}$$

Para calcular el término de segundo orden (la varianza infinitesimal) hay que notar que después de cada salto o bajada, $(X_t - x)^2 = \left(\frac{1}{2N}\right)^2$, entonces:

$$\frac{d}{dt}\mathbb{E}(X_t - x)^2 = \frac{1}{(2N)^2}[(2N-i)(\frac{i}{2N} + \mu_1) + i(\frac{2N-i}{2N}(1-s) + \mu_2)].$$

Como μ_1, μ_2, $s \to 0$ y $\frac{i}{2N} = x$ se tiene:

$$\frac{d}{dt} = \mathbb{E}(X_t - x)^2 = \frac{1}{2N}[2x(1-x) + o(1)].$$

Por lo tanto, para el proceso que corre a tasa N el coeficiente de difusión es $\sigma^2(x) = x(1-x)$.

Al combinar los cálculos se tiene nuevamente que el proceso de frecuencia converge a una difusión cuyo generador infinitesimal es:

$$\mathcal{L}f = \frac{1}{2}x(1-x)\frac{d^2}{dx^2}f + [\gamma x(1-x) + \beta_1(1-x) - \beta_2 x]\frac{d}{dx}f. \tag{4.12}$$

4.3. Modelo de González Casanova-Spanò

Para concluir el capítulo de selección, se presentará un modelo reciente (desarrollado por González-Casanova y Spanò [GS18]) que incorpora una novedosa manera de entender la selección, así como liga el comportamiento de los modelos probabilísticos de selección con la teoría de gráficas aleatorias.

Se comenzará estableciendo un universo de $2N$ individuos (para tener consistencia con los modelos anteriores), los cuales podrán ser alelos de tipo a o A, a los alelos de tipo A

se les denominará alelos ventajosos selectivamente. Para tener noción sobre la "fuerza"de la selección se empleará una distribución geométrica de parámetro q (aunque en [GS18] se generaliza a cualquier distribución que este en $\mathbb{N} \cup \infty$). Las generaciones no se traslaparán y el tamaño de la población sera constante a lo largo del tiempo (las unidades de tiempo son las generaciones $g \in \mathbb{Z}$). La reproducción de los individuos de la población se lleva a cabo de acuerdo al siguiente mecanismo:

1. Elección de padres potenciales: En cada generación g, cada individuo $i \in \{0, ..., 2N\}$ escoge, de forma independiente un número aleatorio $K_{(g,i)}$ de *padres potenciales* de entre los $2N$ individuos de la generación anterior $(g-1)$, donde $K_{(g,i)} \sim Geo(q)$. La elección se hace de modo que dado $K_{(g,i)} = k$, el individuo i elegirá a sus k padres potenciales asignando k *etiquetas* independiente y uniformemente de modo aleatorio entre $\{1, ..., 2N\}$. Sin embargo, la ancestría del proceso únicamente retendrá la información de los distintos padres potenciales elegidos por el individuo, es decir, los tipos de alelo de los padres potenciales que hayan sido elegidos más de una vez serán incluidos en la ancestría una única vez.

2. Elección de tipo de alelo: Al inicio, (en la generación $g = 0$, por ejemplo) los tipos de alelo a y A son asignados de forma arbitraria, sea j el número de alelos de tipo a y $2N - j$ el número de alelos de tipo A. Para cada generación subsecuente, cada individuo será de tipo a si todos sus padres potenciales son de tipo a, si al menos uno de sus padres potenciales es de tipo A entonces será de tipo A (pues A es el tipo de alelo selectivamente ventajoso).

3. Padres potenciales contra padre actual: Se dirá que el *padre actual* del individuo i es el individuo con la primer *etiqueta* observada de todos los padres potenciales del mismo tipo de i, el resto de padres potenciales se considerarán *padres virtuales.*

La dinámica de este modelo resulta ser muy sencilla, pues la ventaja selectiva de los alelos tipo A ocasiona que sea suficiente con que un individuo presente entre sus padres potenciales a un alelo selectivamente ventajoso para que herede su tipo, la naturaleza estocástica del modelo viene dada por dos factores: el número de padres potenciales (mediante una distribución geométrica) así como, dado el número de padres potenciales, la elección de estos mismos (mediante un muestreo con reemplazo).

Con esta información se pueden realizar simulaciones, fijando valores de N. A continuación se presenta un algoritmo que permite crear g generaciones futuras dando como valores iniciales el tamaño de la población, la proporción de alelos, así como el parámetro de la variable aleatoria $k(g, i)$. Este algoritmo no se dio en [GS18], por lo que es una de las aportaciones de esta tesis.

Algorithm 6 Modelo de GC-S

Require: Valor de N, j, q y g

Ensure: Matriz M donde cada columna será una generación del modelo.

1: Generar una matríz de dimensiones $[g, 2N]$
2: En la columna 1, asignar j entradas con a y $2N - j$ entradas con A de forma aleatoria.
3: **for** $i \in 2 : g$ **do**
4: **for** $l \in 1 : 2N$ **do**
5: k <-variable aleatoria geométrica de parámetro q.
6: Generar k posiciones sin reemplazo de $M[i-1, 2N]$.
7: Elegir k individuos de los [1:2N] de la columna anterior.
8: **if** Uno de los k individuos es A **then**
9: La entrada l de la i columna va a ser A

10: **else**
11: La entrada l de la i columna va a ser a
12: **end if**
13: **end for**
14: **end for**
15: **Return** M.

4.3.1. Gráfica aleatoria de Wright-Fisher

El número 3 del mecanismo de reproducción es la liga que conecta este modelo con la teoría de gráficas. Es relevante tener una representación gráfica del modelo, pues se tienen en el mismo espacio de probabilidad el proceso de frecuencias futuras de alelos, así como el proceso de ancestría de la población.

Considérese el conjunto de vértices:

$$V_{2N} := \mathbb{Z} \times \{1, ..., 2N\}.$$

Para cada vértice $v = (g, i) \in V_{2N}$ (cada vértice será un par ordenado donde la coordenada x corresponde a la generación ($g(v) = g$) y la coordenada y a la etiqueta del individuo $i(v) = i$). Una arista se dibuja desde $(g(v) - 1, l)$ hacia v en caso de que l sea un padre potencial de v. El conjunto E_{2N} de todas las aristas tendrá entonces una naturaleza aleatoria y depende de las siguientes variables aleatorias.

1. La colección $K = (K_v : v \in V_{2N})$ de variables aleatorias independientes e idénticamente distribuidas Q_{2N}, que indican el número de padres potenciales que elegirá cada individuo i en la generación g.

2. La colección $L(K) = \{L_{(v,1)}, L_{(v,2)}, ..., L_{(v,K_v)}\}_{v \in V_{2N}}$, de variables aleatorias independientes e idénticamente distribuidas de forma uniforme sobre $\{1, 2, ..., 2N\}$, donde para cada vértice $v = (g, i)$, L_v lista las etiquetas de todos los padres potenciales seleccionados por el individuo i en la generación g.

Para cada vértice v, sea $J_v \leq K_v$ el número de padres distintos que aparecen en L_v; además, sean $\tilde{L}_v = (\tilde{L}_{v,1}, ..., \tilde{L}_{v,J_v})$ las etiquetas correspondientes.

Definición 4.3.1. *Gráfica aleatoria de Wright-Fisher*

Para todo $N \in \mathbb{N}$ y Q_{2N}, la gráfica aleatoria de Wright-Fisher con selección Q_{2N} dependiente de la frecuencia es la gráfica con conjunto de vértices V_{2N} y conjunto de aristas:

$$E_{2N} := \{\{(g(v) - 1, \tilde{L}_{v,j}), (g, i)\} : j = 1, ..., J_v, v = (g, i) \in V_{2N}\}. \tag{4.13}$$

donde $\tilde{L}_{v,j}$ y J_v son las que se definieron anteriormente.

4.3.2. Frecuencias de los alelos

Sin perdida de generalidad, se asigna a cada vértice de una generación inicial fija $g = 0$ ya sea el tipo a o el tipo A arbitrariamente. Se tiene entonces interés en la evolución de las frecuencias de los alelos tipo a. Sea

$$\xi(v) = \begin{cases} 0, & \text{si } v \text{ es de tipo } a \\ 1, & \text{si } v \text{ es de tipo } A. \end{cases}$$

Y con X_g se denotará a la frecuencia de los individuos tipo a en la generación $g = 0, 1, ...$, donde X_g esta dada por:

$$X_g^{2N} = 1 - \frac{1}{2N} \sum_{\{v : g(v) = g\}} \xi(v). \tag{4.14}$$

Observación. *(Función generadora de probabilidad) Si X es una variable aleatoria discreta tomando valores en los enteros no negativos, entonces la función generadora de probabilidad de X esta definida como:*

$$G(z) = \mathbb{E}(z^X) = \sum_{x=0}^{\infty} p(x) z^x,$$

donde p es al función de masa de probabilidad de X.

Sea $G_{Q_{2N}}(x)$ la función generadora de probabilidad de Q_{2N}. Considérese la función dada por:

$$\mu_{2N}(x) = \mathbb{P}(\xi(v) = 0 | X^{2N}_{g(v)-1} = x). \tag{4.15}$$

En la ecuación anterior se tiene la probabilidad de que un vértice v sea de tipo a si la frecuencia de los individuos de tipo a en la generación anterior es x. De acuerdo al paso 2 del mecanismo de reproducción (4.3), el evento de interés ocurre únicamente si todos los padres potenciales de v son de tipo a. Como (K_v) es una secuencia de variables aleatorias independientes e idénticamente distribuidas, la probabilidad en 4.15 no depende de v, y como los valores de las coordenadas en L_v no dependen de K_v entonces se tiene:

$$\begin{aligned} \mu_{2N}(x) &= \sum_{k=1}^{\infty} \mathbb{P}(\xi(g(v)-1, L_{(v,j)}) = 0 \text{ para todo } j \leq K_v, K_v = x | X^{2N}_{g(v)-1} = x) \\ &= \sum_{k=1}^{\infty} = x^k Q_{2N}(K_v = k) = G_{Q_{2N}}(x). \end{aligned} \tag{4.16}$$

Con base en lo anterior, y usando el hecho de que la elección de padres potenciales por parte de los individuos se realiza de forma independiente se tiene el siguiente resultado.

Observación. *En una gráfica aleatoria de Wright-Fisher con parámetro de selección Q_{2N}, el proceso de frecuencias de los alelos tipo a $(X^{2N}_g : g \in \mathbb{N})$ evoluciona como una cadena de*

Markov homogénea con espacio de estados $\frac{[2N]}{2N} := \{0, \frac{1}{2N}, \frac{2}{2N}, ..., \frac{2N-1}{2N}, 1\}$ *con probabilidades de transición dadas por:*

$$\mathbb{P}(Y_g^{2N} = \frac{m}{2N} | Y_{g-1}^{2N} = x) = \binom{2N}{m} G_{Q_{2N}}(x)^m (1 - G_{Q_{2N}}(x))^{2N-m},$$
$$m = 0, 1, 2, ..., 2N, \text{ para cada } x \in \frac{|2N|}{2N}.$$

Ejemplo 4.3.2. *(Q_{2N} con distribución geométrica.)*

Con la elección

$$Q_{2N}(K_v > m) = S_{2N}^{m-1}, m = 1, 2, ...$$

para algún $S_{2N} > 0$ (distribución geométrica), la función generadora de probabilidad μ_{2N} se convierte en:

$$\mu_{2N}(x) = \frac{x(1 - s_{2N})}{1 - x + x(1 - s_{2N})}.$$

Así se tiene que el proceso de frecuencias de la gráfica (V_{2N}, E_{2N}, Q_{2N}) coincide con el modelo clásico de Wright-Fisher con coeficiente de selección débil S_{2N}.

Capítulo 5

Matraca de Muller

En este capítulo se estudiará un fenómeno que se presenta en poblaciones asexuales por efectos de la mutación. Las siguientes secciones están basadas en Etheridge, Pfaffelhuber y Wakolbinger [EPW09].

La ausencia de recombinación en los organismos que tienen una reproducción asexual suele presentar una acumulación de mutaciones negativas (aquellas que, de algún modo, empeoran el tipo de los alelos) en los individuos. Estas mutaciones pueden llegar a un punto donde la clase (o tipo) de los individuos originales se pierde (en otras palabras, el tipo de los individuos mutados es fijado). Una vez que el individuo menos mutado acumula al menos una mala mutación, no habrá individuos con menos mutaciones. A este fenómeno se le conoce como matraca de Muller (*Muller ratchet* en inglés), en honor a Hermann Joseph Mullerr.

Desde 1932 [Mul32], Muller hablaba sobre las ventajas de la reproducción sexual sobre la asexual y en 1964 [Mul64] acuñó el término *ratchet* a este fenómeno por su similitud con el funcionamiento de una matraca, pues una vez que se presenta el fenómeno, es decir, se pierde la clase original de los individuos, ya no se podrá recuperar ese tipo. Entonces se dice que la matraca hace *click*.

Una de las preguntas naturales que surgen luego de presentar el Muller's Ratchet es:

¿Cuántas generaciones le tomará a una población perder su mejor tipo actual? O lo que es equivalente, ¿cuál es la tasa de la matraca de Muller?

Se va a estudiar este fenómeno usando un modelo de gráficas aleatorias utilizando mutación.

Observación. *Tradicionalmente se ha estudiado este fenómeno usando el modelo de John Haigh. Para más detalles de este modelo se puede consultar el anexo A.1.*

5.1. Matraca de Muller en un modelo de gráficas aleatorias

Para esta sección se introducirá un nuevo modelo, en el cual se usará la selección como en el modelo presentado en [GS18] y se agregará mutación con el propósito de estudiar la tasa de la matraca de Muller. Se comenzará primero dando una explicación sobre como se realizará la mutación y posteriormente se dará una definición formal.

Considérese una población constante de N individuos, las generaciones no se traslapan y en la generación inicial (dígase, $g = 0$) se tiene que todos los individuos son del mismo tipo 0 (en este modelo será más útil denotar los tipos por números). Cada individuo nuevamente será un nodo (aunque también puede pensarse como un elemento de una matriz, donde cada columna será una generación) de la gráfica aleatoria, y se tendrán aristas con los padres potenciales.

Para determinar el tipo de los individuos de la siguiente generación (por ejemplo, $g = 1$), cada individuo escogerá un número aleatorio k de padres potenciales, donde k sigue una distribución geométrica $(1 - s)$ y s es el parámetro asociado a la selección. Después, de entre todos los individuos de la generación anterior elegirán k individuos específicos de manera

uniformemente aleatoria y tomará el tipo del (o los) padre(s) que tenga(n) el menor tipo. Una vez que se tiene el "tipo inicial" del individuo en la nueva generación, este podrá mutar, es decir, su "tipo definitivo" estará determinado como el tipo inicial más una variable aleatoría Bernoulli de parámetro μ, donde μ se llamará *parámetro de mutación*.

Como se hizo en el modelo de GC-S, se considerarán las siguientes variables aleatorias para definir la gráfica aleatoria asociada a este modelo:

1. La colección $K = (K_v : v \in V_n)$ donde V_N es el conjunto de todos los vértices ($V_N := \mathbb{Z}X\mathbb{N}$), K_v sigue una distribución geométrica de parámetro $(1-s)$ que indica el número de padres potenciales que elegirá el individuo i en la generación g.

2. La colección $L(K) = \{L_{v,1}, L_{v,2}, ..., L_{v,K_v}\}$ de variables aleatorias independientes e idénticamente distribuidas de forma uniforme sobre $\mathbb{N}$, donde para cada vértice $v = (g, i)$, L_v almacena las etiquetas (posiciones) de todos los padres potenciales de i.

Además, para cada vértice v, $J_v \leq K_v$ es el número de padres distintos que eligió el individuo i en la generación g, mientras que $\tilde{L_v} = \{\tilde{L}_{v,1}, ..., \tilde{L}_{v,J_v}\}$ son las posiciones correspondientes.

Una forma matemática de expresar la mutación en este enfoque es considerando $\{M_v\}_{v\in V}$ variables aleatorias independientes e idénticamente distribuidas Bernoulli(μ). Entonces considerese a $\nu = \{v \in V : M_v = 1\}$. Estas son las variables asociadas a la mutación del individuo del vértice v.

Así, se puede definir a $T_v := \{$"Camino del vértice v a la generación 0".$\}$, y al tipo v como $\rho_v = \text{ínf}\{|\tau \cap \nu|\}_{\tau\in T_v}$.

Definición 5.1.1. *Modelo de selección y mutación en gráficas aleatorias*

La gráfica aleatoria del modelo de selección y mutación en gráficas aleatorias está dada por (V_N, E_N, ν_N)*, donde* V_n *es el conjunto de vértices,* E_N *es el conjunto de aristas,* ν_N *representa*

la mutación de los individuos mediante su tipo particular ρ_v. *Sea* $v = (g, i, \rho_v)$.

$$E_N := \{\{((g(v) - 1, \tilde{L}_{v,j}, \rho_{\tilde{L}_{v,j}}), v)\} : j = 1, ..., Jv, v \in V_N\} \tag{5.1}$$

Y $\tilde{L}_{v,j}$ *y* J_v *son las que se definieron anteriormente.*

Entonces, se dice que la matraca de Muller hará "click" en este modelo en aquella generación n_0 en la que no queden individuos del tipo inicial, es decir, todos los individuos de la generación n_0 habrán heredado al menos una mutación, y nuevamente a causa de la falta de recombinación o mutaciones favorables heredarán esas características a toda su descendencia.

5.1.1. La tasa de la matraca de Muller

Para tener idea de como se comportaba la matraca de Muller en este modelo se usó el algoritmo 7.

Algorithm 7 Modelo de selección y mutación en gráficas aleatorias

Require: Valor de N, j, q y g

Ensure: Valor c de número de generación en la que hizo click la matraca.

1: Generar una matriz de dimensiones $[g, 2N]$, hacer $c = 1$, $m_f = 1$, $i = 2$.
2: En la columna 1, asignar las $2N$ entradas con 1.
3: **while** $m_f < 2$ **do**
4: **for** $l \in 1 : 2N$ **do**
5: k <-variable aleatoria geométrica de parámetro $1 - s$.
6: Generar k posiciones sin reemplazo de $M[i - 1, 2N]$.
7: Elegir k individuos de los [1:2N] de la columna anterior.
8: Hacer m_0 al menor tipo de los k individuos.

9: ν<-variable aleatoria bernoulli de parámetro μ.
10: La entrada l será de tipo $m_0 + \nu$
11: **end for**
12: Hacer m_f al valor mínimo de la columna i.
13: **if** $m_f = 1$. **then**
14: $c = c + 1$, $i = i + 1$.
15: **end if**
16: **end while**
17: **Return** c.

Es posible usar el algoritmo 7 para obtener un estimador $\bar{\theta}$ de Monte Carlo crudo del tiempo en el cual hace click la matraca de Muller, en este contexto se define a X como la variable aleatoria asociada a la tasa de la matraca y a θ como su media, entonces $\bar{\theta} = \frac{\sum_{i=i}^{n} x_i}{n}$, donde las x_i se obtendrán por medio de simulaciones. En el desarrollo de este trabajo no se logró llegar a una forma analítica de θ, pero se tiene la intuición de que esta relacionada con la distribución geométrica de parámetro $f(n, s, \mu)$ dependiente del tamaño de la población, el coeficiente de selección y/o el parámetro de mutación.

Se hizo una simulación usando un tamaño de población pequeño ($n = 30$), se presentan los valores de $\bar{\theta}$ para distintas combinaciones de λ y μ en la tabla 5.1.

Tabla 5.1: $n = 30$ individuos, estimación del tiempo en el que hace click la matraca de Muller

	$s = 0.01$	$s = 0.1$	$s = 0.2$	$s = 0.3$	$s = 0.4$	$s = 0.5$	$s = 0.6$	$s = 0.7$	$s = 0.8$	$s = 0.9$	$s \approx 1$
$\mu = 0.01$	216.28	5125.5	> 20000	> 20000	> 20000	> 20000	> 20000	> 20000	> 20000	> 20000	> 20000
$\mu = 0.1$	39.54	62.91	184.75	6001.3	> 20000	> 20000	> 20000	> 20000	> 20000	> 20000	> 20000
$\mu = 0.2$	23.02	27.45	40.35	85.36	619.3	13510	> 10000	> 20000	> 20000	> 20000	> 20000
$\mu = 0.3$	16.3	17.97	22.24	30.4	49.98	354.1	19425	> 20000	> 20000	> 20000	> 20000
$\mu = 0.4$	12.33	12.36	14.49	16.92	23.81	46.79	248.33	15923	> 20000	> 20000	> 20000
$\mu = 0.5$	9.16	9.63	10.71	11.71	14.28	19.56	33.4	182.48	7194	> 20000	> 20000
$\mu = 0.6$	6.86	7.3	8.13	8.46	9.59	11.53	14.09	30.38	193.78	17319	> 20000
$\mu = 0.7$	5.64	5.92	6.14	6.39	6.88	7.72	8.34	12.53	22.53	530.67	12834
$\mu = 0.8$	4.23	4.35	4.52	4.67	4.92	5.18	5.9	6.68	8.36	21.81	555.58
$\mu = 0.9$	3.29	3.26	3.12	3.2	3.25	3.34	3.57	3.9	3.96	4.77	16.99
$\mu = 1$	1	1	1	1	1	1	1	1	1	1	1

Se pudo observar que conforme va creciendo la selección, la mutación deja de influir y la matraca tarda más tiempo en hacer click, por otro lado, hay que recordar que si $\mu = 0$ entonces la matraca nunca hará click, pues nunca habrá mutaciones. Además, el tamaño de la población también debe considerarse, ya que una mayor población le da a los individuos más opciones al momento de elegir a sus padres potenciales, aunque opera más en el sentido de un parámetro de escala, mientras que la selección y la mutación influyen como un parámetro de forma.

Luego de notar que en los casos en los que la selección es alta y la mutación baja la matraca podría tardar mucho (más de 20000 generaciones) en hacer click. Se hicieron las siguientes simulaciones para valores bajos de selección ($s = 0$ y $s = \frac{1}{3}$). Así como para valores de mutación altos: $(0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.95, 1)$.

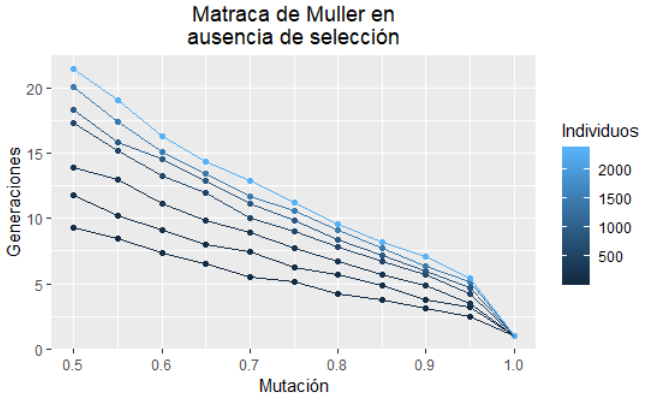


Figura 5.1: Tiempo en hacer click la matraca de Muller, $s = 0$

Con selección baja se pudo observar que conforme aumenta el tamaño de la población, aumenta también el valor del tiempo en el cual la matraca de Muller hace click, esto se debe a que al aumentar el número de individuos hay más oportunidad de elegir a un ancestro que no ha mutado, aunque, conforme aumenta la mutación va decreciendo el valor de dicho tiempo.

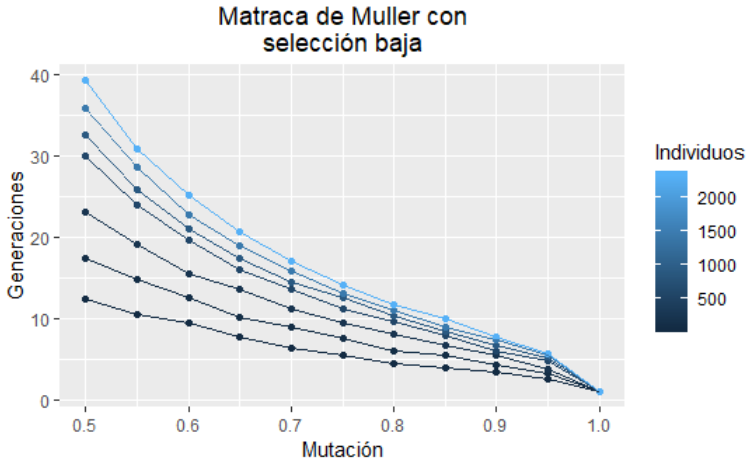


Figura 5.2: Tiempo en hacer click la matraca de Muller, $s = \frac{1}{3}$

Tras repetir el mismo ejercicio, pero ahora con selección baja (ver figura 5.2), se puede notar que las tendencias de las gráficas son concordantes con las de la primera simulación.

Si se considera el caso neutral (es decir, $s = 0$) se puede tener una intuición más para el comportamiento del tiempo en el cual hace click la matraca de Muller. En cada generación, cada individuo elegirá a un único padre potencial y dependiendo del resultado de su Bernoulli asociada podría mutar o no. Esta dinámica es prácticamente un modelo de Wright-Fisher con mutación, y por lo tanto se puede definir el siguiente proceso de conteo.

Definición 5.1.2. *(Proceso de frecuencias para el modelo de mutación en gráficas aleatorias) Si se tiene $s = 0$, y se denota a $x = \frac{k}{n}$ como la frecuencia de individuos sin mutar, entonces X_g es el proceso de frecuencia que indica el número de individuos que no han mutado en la población en la generación g y*

$$(X_g | X_{g-1} = x) \sim \frac{Binomial(N, p)}{N} \tag{5.2}$$

donde $p = x(1 - \mu)$.

La razón de dividir por N la distribución Binomial del proceso es que se busca tener un proceso de frecuencias. Por otro lado, el valor de p se obtiene porque, dado que si en la generación anterior la frecuencia de los individuos sin mutar era x, entonces se necesita elegir alguno de esos individuos y que no haya mutación $(1 - \mu)$.

El uso del proceso anterior cobra importancia al calcular su esperanza, considerando el caso en el que se está en la primera generación:

$$\mathbb{E}_x(X_1) = \frac{1}{N} N x(1 - \mu).$$

Para la segunda generación:

$$\mathbb{E}_x(X_2) = \mathbb{E}_x[\mathbb{E}_{X_1}(X_1')] = \mathbb{E}_x[X_1(1 - \mu)]$$
$$= (1 - \mu)\mathbb{E}_x(X_1) = (1 - \mu)x(1 - \mu) = x(1 - \mu)^2$$

donde X_1' es una realización independiente del proceso dado por 5.2.

Se puede intuir cual será el valor de la esperanza de $(X_g|X_0 = x)$, se procederá usando inducción:

$$\text{Supóngase que: } \mathbb{E}_x(X_g) = x(1-\mu)^g$$

Usando las ecuaciones de Chapman Kolmogorov y la propiedad de torre de la esperanza condicional:

$$\begin{aligned} \mathbb{E}_x(X_{g+1}) &= \mathbb{E}_x[\mathbb{E}_{X_g}(X_1')] \\ &= \mathbb{E}_x(X_g(1-\mu)) \\ &= (1-\mu)\mathbb{E}_x(X_g) \\ &= x(1-\mu)^{g+1}. \end{aligned} \tag{5.3}$$

Por otro lado, considérese la desigualdad de Markov:

Observación. *(Desigualdad de Markov)*

Si X es una variable aleatoria no negativa, entonces

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}. \tag{5.4}$$

Usando el valor de la esperanza 5.3 y la expresión de 5.4 se tiene lo siguiente: se considerará $x = 1$, pues en el modelo desarrollado en esta sección se empieza sin individuos que hayan mutado en la primera generación:

$$\mathbb{P}_1(X_g > 0) = \mathbb{P}_1(X_g \geq \frac{1}{N}) \leq \frac{\mathbb{E}(X_g)}{\frac{1}{N}} = (1-\mu)^g N.$$

El paso de la primera a la segunda igualdad es inmediato, ya que el espacio de estados de X_g es $E = \{0, \frac{1}{N}, \frac{2}{N}, ..., 1\}$.

Finalmente, si se toma $g = \tilde{g} = \dfrac{-\log(rN)}{\log(1-\mu)}$ y se sustituye en el resultado anterior se obtiene la siguiente cota:

$$\mathbb{P}_1(X_{\tilde{g}} > 0) \leq \exp\left(\tilde{g}\log(1-\mu)\right)N = \exp\left(-\log(Nr)\right)N = \frac{1}{r}.$$

Con lo que se concluye que el tiempo en el que la matraca de Muller hace click esta acotado por un término de orden $(\log(N))$, en el caso en el que no hay selección. Con selección el orden es mayor.

5.1.2. Posibles variaciones

Aunque en este trabajo se ha limitado el estudio de la matraca de Muller a casos donde la población es constante y las mutaciones positivas no existen, se sugiere realizar los siguientes cambios para posteriores estudios:

- Para tener un modelo en poblaciones dinámicas cada individuo actual i tendrá un número de descendientes d aleatorio (con parámetro s_i, es decir, cada individuo tendrá una diferente resistencia selectiva), donde todos heredarán el tipo de su padre y mutarán de acuerdo a una Bernoulli(μ).

- Para tener un modelo con mutaciones positivas y población constante basta redefinir la variable asociada a la mutación de modo que valga 1 con probabilidad μ y -1 con probabilidad $1-\mu$. De este modo incluso se podría mejorar el tipo original de los individuos cuando algún descendente tuviera un tipo negativo.

Otro pregunta que quedará abierta será una expresión analítica para la tasa de la matraca de Muller, pues este trabajo se limitó a definir el modelo y estudiar el fenómeno desde un punto de vista computacional.

Capítulo 6

Discusión

El objetivo de este trabajo fue estudiar la matraca de Muller, en especial el tiempo en el que tarda en hacer click la matraca, pues es una de las preguntas que siguen abiertas en el área de dinámica de poblaciones (ver [EPW09]). Para esto, primero se revisaron los modelos clásicos de deriva génica, un fenómeno poco conocido por los estudiantes de actuaría, y que puede ser útil para proporcionar una multitud de ejemplos en un curso de Procesos Estocásticos I.

El modelo propuesto 5.1.1 incluye una dinámica de deriva génica Wright-Fisher e incorpora una selección natural basada en el modelo de Gonzalez-Casanova y Spanò 4.3. Por otro lado la mutación se incorporó usando una variable aleatoria Bernoulli.

Los pasos que llevaron a proponer el modelo 5.1.1 fueron: seleccionar una dinámica de deriva génica (en este caso, se tomó una dinámica discreta), incorporar una selección basada en gráficas aleatorias, e idear una forma sencilla de representar la mutación.

Una de las complicaciones que surgieron durante el proceso de aprendizaje de los modelos de deriva génica fue, dada una nueva generación, el identificar los cambios que había sufrido y saber si se debían a la selección, a la mutación, o simplemente a la deriva génica. Esta complicación se superó al abordar el problema de forma computacional, pues hubo claridad

desde el inicio en que parámetros causan las variaciones en los valores que toma el proceso, al revisar los resultados de las simulaciones.

Al final, dicha intuición no fue suficiente para obtener una expresión analítica cerrada para el tiempo en el que la matraca hace click.

De cualquier forma, la aportación de esta tesis fue presentar el modelo 5.1.1 y su implementación computacional, que permiten tener una aproximación de como afectan al comportamiento de la matraca los distintos parámetros del modelo. Por otro lado, se hizo un pequeño cálculo que permite encontrar una cota del tiempo en que hace click la matraca de Muller para el caso neutral.

Apéndice A

Apéndice

A.1. Matraca de Muller usando el enfoque de Haigh

A continuación, se presentará el análisis de la matraca de Muller realizado por John Haigh en [Hai78].

A.1.1. El modelo

En su artículo, Haigh basó su análisis en el modelo de Wright-Fisher (que se presento en la sección 3.1), con los siguientes supuestos:

- La resistencia relativa de un individuo con k mutaciones desfavorables será $(1-s)^k$, con $s > 0$.
- Cada individuo recibirá nuevas mutaciones desfavorables a tasa λ por generación.
- El número real de nuevas mutaciones sigue una distribución Poisson con media λ.

Definición A.1.1. *Proceso de conteo del Modelo de Haigh*

Considere una población de tamaño N, donde cada individuo tendrá un tipo $i \geq 0$, si $X_k(t)$ es el número de individuos en la generación t que tienen exactamente k mutaciones y sea $\mathbf{X}(t) = (X_0(t), X_1(t), ...)$, entonces la distribución de $\mathbf{X}(t+1)$ será multinomial con parámetros N y $\{p_k(t), k = 0, 1, ...\}$ donde:

$$\begin{aligned} p_k(t) &= \sum_{j=0}^{k} X_{k-j}(t) \frac{(1-s)^{k-j}}{j! T_{k-j}(t)} \exp(-\lambda)\lambda^j \\ T_r(t) &= \sum_{i=0}^{\infty} X_i(t)(1-s)^{ir}, \quad (r = 1, 2...). \end{aligned} \tag{A.1}$$

Por otro lado, considérese la distribución $\pi = (\pi_0, \pi_1, ...)$, π_k será estacionaria si satisface:

$$\pi_k = N \sum_{j=0}^{k} \pi_{k-j}(1-s)^{k-j} \exp(-\lambda) \frac{\lambda^j}{j! T}$$

donde $T = \sum_{i=0}^{\infty} n_i(1-s)^i$. De la ecuación anterior, si se toma $k = 0$ se puede ver que $\pi_0 = \dfrac{N\pi_0 \exp(-\lambda)}{T}$, por lo tanto (y si $\pi_0 \neq 0$), $T = N \exp(-\lambda)$ y la ecuación A.1 se reduce a:

$$\pi_k = \sum_{j=0}^{k} \pi_{k-j}(1-s)^{k-j} \frac{\lambda^j}{j!}$$

para $k = 0, 1, 2, ...$ cuya única solución es $\pi_k = \dfrac{\pi_0 \theta^k}{k!}$ con $\theta = \frac{\lambda}{s}$, como $N = \sum_{k=0}^{\infty} \pi_k$, la única distribución estacionaria con $\pi_0 > 0$ es:

$$\pi_k = \frac{N \exp(-\theta)\theta^k}{k!}, \quad (k = 0, 1, 2, ...). \tag{A.2}$$

Sin embargo, Haigh establece que su modelo sugerido tiene una desventaja, pues de acuerdo a las transiciones presentadas en A.1, y a que $\mathbb{E}(\mathbf{X}(t+1)|\mathbf{X}(t)) = Np_k(t)$; pero se tiene que puede llegar a ser complicado encontrar analíticamente la distribución teórica de $\mathbf{X}(t+1)$ cuando $\mathbf{X}(t)$ no esta cerca de la distribución estacionaria π

A.1.2. Análisis de la matraca de Muller

Ahora, considérese una población que inicia en el equilibrio establecido en la ecuación A.2. Como cada clase tiene un número entero de miembros, sea $X_k(0)$ el entero más cercano a π_k, y supóngase que $\pi_0 = N\exp(-\theta) \geq 0.5$, por lo que $X_0(0) \geq 1$. Entonces se tiene que

$$\mathbb{E}[T_1(t+1)|\mathbf{X}(t)] = \sum_{k=0}^{\infty} Np_k(t)(1-s)^k = N\exp(-\lambda s)\frac{T_2(t)}{T_1(t)}.$$

De forma similar, se tiene que

$$\mathbb{V}[T_1(t+1)|\mathbf{X}(t)] = N\exp(-2\lambda s)\Big(\exp(\lambda s^2)\frac{T_3(t)}{T_1(t)} - \frac{T_2^2(t)}{T_1^2(t)}\Big).$$

Al expandir los cocientes de la diferencia final como una serie de potencias en s, se obtiene:

$$\mathbb{V}[T_1(t+1)|\mathbf{X}(t)] = N\exp(-2\lambda s)(s^2(\lambda + \sigma^2(t)) + O(s^3)$$

donde $\sigma^2(t)$ es la varianza asociada a una variable aleatoria que toma el valor k con probabilidad $\frac{X_k(t)}{N}$, para $k \in \{0, 1, ...\}$. Sin embargo, como $\mathbb{E}(X_k(t)) = \pi_k = N\exp(-\theta)\frac{\theta^k}{k!}$ se tiene que $\sigma^2(t) \simeq \theta$. Por lo tanto:

$$\mathbb{V}[T_1(t+1)|\mathbf{X}(t)] \simeq N\exp(-2\lambda s)(\lambda s(1+s) + O(s^3)).$$

Y usando un argumento similar:

$$\mathbb{E}[T_1(t+1)|\mathbf{X}(t)] \simeq N\exp(-\lambda s).$$

Estas dos ultimas expresiones muestran que, cuando se tiene un tamaño de población N grande y un coeficiente de selección s pequeño, la esperanza de $T_1(t+1)$ supera por mucho su desviación estándar, entonces $T_1(t)$ cambia lentamente, y por lo tanto permanece cercana a $N\exp(-\lambda)$ por un largo tiempo. La interpretación de $T_1(t)$ es sencilla, ya que $\dfrac{T_1(t)}{N}$ es la resistencia media de la población en la generación t.

Sea C_0 la clase cuyos individuos no tienen ninguna mutación desfavorable. Además, el tamaño del resto de clases C_1, C_2 esta también gobernado por la multinomial presentada al inicio de esta sección, pero las probabilidades $\{p_k(t)\}$ están cambiando en cada generación, y por lo tanto, el tamaño de C_i en la generación en la cual C_0 se pierde tiene una distribución difícil de calcular, aunque su media será cercana al valor original de π_i.

Supóngase que C_0 se pierde en una generación t_0, y que, para $t \geq k$, $k \geq 0$, se denota a $Y_k(t)$ como $Y_k(t) = X_{k+1}(t + t_0)$. Así, se tiene el resultado principal de Haigh:

Teorema A.1.2. *(Teorema de Haigh)*
Si $\gamma(t) = (1-s)^t$ *y* $m_k(t) = \pi_{k+1}\Big(1 - \dfrac{(1-\gamma(t)^{k+1})}{1-\exp\left(-\theta\gamma(t)\right)}\Big)$.

$$\mathbb{E}\Big(Y_k(t)|\mathbb{Y}(0) = \frac{\pi_1, \pi_2, ...}{1-\exp\left(-\theta\right)}\Big) \simeq m_k(t). \tag{A.3}$$

Los valores de π *son los mismos que se definen en A.2.*

Haigh argumenta que en el momento en el que C_0 se pierda, el tamaño medio de C_i sera cercano a π_i para cualquier i dado, y esa es su justificación para usar el teorema anterior con el propósito de examinar el patrón de comportamiento de $\{X_k(t)\}$ después de la generación t_0.

Corolario A.1.3. *Sea* τ *el entero más cercano a* $\dfrac{-\log\theta}{\log(1-s)}$. *Entonces*

1. $\mathbb{E}\Big(Y_k(\tau)|\mathbf{Y(0)} = \dfrac{\pi_{\mathbf{1}}, \pi_{\mathbf{2}}, ..}{\mathbf{1}-\exp(-\theta)}\Big) \simeq 1.6\pi_k$ *para* k *pequeña.*

2. $\mathbb{E}(U_1(\tau)) \simeq \dfrac{N\exp(-\lambda)}{1-0.4s}$ *donde* $U_1(u) = \sum_{k=0}^{\infty} Y_k(u)(1-s)^k$.

A.2. Códigos de simulación

A.2.1. Análisis del Modelo de Wright Fisher

A continuación se presenta el código de [Ste16] usado para generar la figura 3.1.

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(viridis)
# data.frame a ser llenado
wf_df <- data.frame()

# tamanios de poblacion
sizes <- c(10,50, 100, 1000, 5000)

# frecuencias iniciales
starting_p <- c(.01, .3, .6, .99)

# no. de generaciones
n_gen <- 100

# no. de repeticiones por simulacion
n_reps <- 50

# correr las simulaciones
for(N in sizes){
```

```
  for(p in starting_p){
    p0 <- p
    for(j in 1:n_gen){
      X <- rbinom(n_reps, 2*N, p)
      p <- X / (2*N)
      rows <- data.frame(replicate = 1:n_reps, N = rep(N, n_reps),
                         gen = rep(j, n_reps), p0 = rep(p0, n_reps),
                         p = p)
      wf_df <- bind_rows(wf_df, rows)
    }
  }
}

# graficando
p <- ggplot(wf_df, aes(x = gen, y = p, group = replicate)) +
  geom_path(alpha = .5) + facet_grid(N ~ p0) + guides(colour=T)
p
```

A.2.2. Modelo de selección y mutación en gráficas aleatorias

El siguiente código genera una trayectoría del modelo considerando los parámetros:

1. g: número de generaciones.
2. n: tamaño de la población.
3. s: coeficiente de selección.

4. μ: coeficiente de mutación.

```
MSG = function(g=10,n=25,s=.5,mu=0.5){
  a=matrix(0,n,g)
  l=matrix(0,n,g)
  mut=matrix(0,n,g)
  cont=1
  a[,1]=rep(1,n)
  for (k in 2:(g-0)){
    x=(rgeom(n,1-s)+1) #Padres potenciales de cada individuo.

    for (i in 1:n){
      temp=x[i]
      b=sample(as.numeric(a[,k-1]),temp,T,NULL)
      #Eleccion de los padres potenciales (sus posiciones.)
      z=min(b)
      #Eleccion del padre actual
      l[i,k]=sample(which(a[,k-1]==z),1)
      u=rbinom(1,1,mu) #Hay mutacion
      mut[i,k]=u
      a[i,k]=z+u #si TRUE, se suma.


    }
    minimo=min(as.numeric(a[,k]))
    if(minimo==1){cont=cont+1}
```

```
  }
  res=list(Matriz=a,Muller=cont)
  return(res)
}
```

A.2.3. Matraca de Muller en el modelo de selección y mutación en gráficas aleatorias

Para el análisis de la matraca de Muller se modificó el algoritmo anterior para que se detuviera cuando la matraca hiciera click por primera vez, posteriormente se realizó la función que estimó la tasa de la matraca de Muller usando Monte Carlo crudo.

```
MSG1 = function(n=25,s=.5,mu=0.5){
  cont=0
  minimo=1
  cond=0
  k=2
  uno=rep(1,n)
  dos=rep(1,n)
  cero=rep(0,n)
  a=cbind(uno,dos)
  l=cbind(cero,cero)
  mut=cbind(cero,cero)
  colnew=0

  while(cond==0){
    x=(rgeom(n,1-s)+1) #Padres potenciales de cada individuo.
```

```
    for (i in 1:n){
      temp=x[i]
      #Eleccion de los padres potenciales (sus posiciones.)
      b=sample(as.numeric(a[,k-1]),temp,T,NULL)
      z=min(b)    #Eleccion del padre actual
      u=as.numeric(rbinom(1,1,mu)) #Hay mutacion
      colnew[i]=z+u #Se suma.
    }
    a=cbind(a,colnew)
    minimo=min(as.numeric(a[,k]))
    if(minimo == 1){
      cont=cont+1
      k=k+1
      colnew=0
    }
    else{cond=1}
  }
  a=a[,-1]
  a=a[,-ncol(a)]
  colnames(a) <- NULL
  res=list(Matriz=a,Muller=cont)
  return(res)
}


MontecarloMSGMean<-function(n,s,mu,t){
```

```
  b=rep(0,t)
  for (i in 1:t){
    temp=MSG1(n,s,mu)
    b[i]=temp$Muller
  }
  return(mean(b))
}
```

Bibliografía

[And12] W. J. Anderson. *Continuous-time markov chains: an applications oriented approach*. Springer, 2012.

[BM10] J. A. Bondy y U. S. R. Murty. *Graph theory*. Springer, 2010.

[CK65] J. F. Crow y M. Kimura. "Evolution in Sexual and Asexual Populations". En: *American Naturalist* 99 (909 dic. de 1965). DOI: 10.2307/2459132. URL: http://gen.lib.rus.ec/scimag/index.php?s=10.2307/2459132.

[Dur10] R. Durrett. *Probability models for DNA sequence evolution*. Springer, 2010.

[ER59] P. Erdős y A. Rényi. "On Random Graphs I." En: *Publicationes Mathematicae (Debrecen)* 6 (1959), págs. 290-297.

[Eth11] A. Etheridge. *Some mathematical models from population genetics. Lecture notes in mathematics Vol 2012*. Springer, 2011.

[EPW09] A. M. Etheridge, P. Pfaffelhuber y A. Wakolbinger. "How often does the ratchet click? Facts, heuristics, asymptotics". En: *Trends in Stochastic Analysis*. Ed. por Jochen Blath, Peter Mörters y MichaelEditors Scheutzow. London Mathematical Society Lecture Note Series. Cambridge University Press, 2009, págs. 365-390. DOI: 10.1017/CBO9781139107020.016.

[Gen10] J. E. Gentle. *Random number generation and Monte Carlo methods*. Springer, 2010.

[GS18] A González-Casanova y D Spanò. "Duality and fixation in Ξ-Wright–Fisher processes with frequency-dependent selection". En: *The Annals of Applied Probability* 28.1 (2018), págs. 250-284. DOI: 10.1214/17-aap1305.

[Hai78] J. Haigh. "The accumulation of deleterious genes in a population—Muller's Ratchet". En: *Theoretical population biology* 14 (nov. de 1978), págs. 251-67. DOI: 10.1016/0040-5809(78)90027-8.

[KT81] S. Karlin y H. M. Taylor. *A second course in stochastic processes*. Academic Press, 1981.

[KP95] P. E. Kloeden y E. Platen. *Numerical solution of stochastic differential equations*. Stochastic Modelling and Applied Probability. Springer, 1995.

[Kni00] F. B. Knight. *Essentials of Brownian motion and diffusion*. American Mathematical Society, 2000.

[LB] S. López y F. Baltazar. *Notas de Simulación Estocástica (versión preliminar)*. URL: http://sistemas.fciencias.unam.mx/~silo/.

[Mul64] H. J. Muller. "The relation of recombination to mutational advance". En: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1.1 (1964), págs. 2-9. DOI: 10.1016/0027-5107(64)90047-8.

[Mul32] Herman Joseph Muller. "Some Genetic Aspects of Sex". En: *The American Naturalist* 66.703 (1932), págs. 118-138. DOI: 10.1086/280418.

[Nor97] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[Rin12] L. A. Rincón. *Introducción a los procesos estocásticos*. Universidad Nacional Autónoma de México, 2012.

[Ros96] S. M. Ross. *Stochastic Process*. John Wiley, 1996.

[Ste16] Marcus Stephens. Mar. de 2016. URL: https://stephens999.github.io/fiveMinuteStats/wright_fisher_model.html.