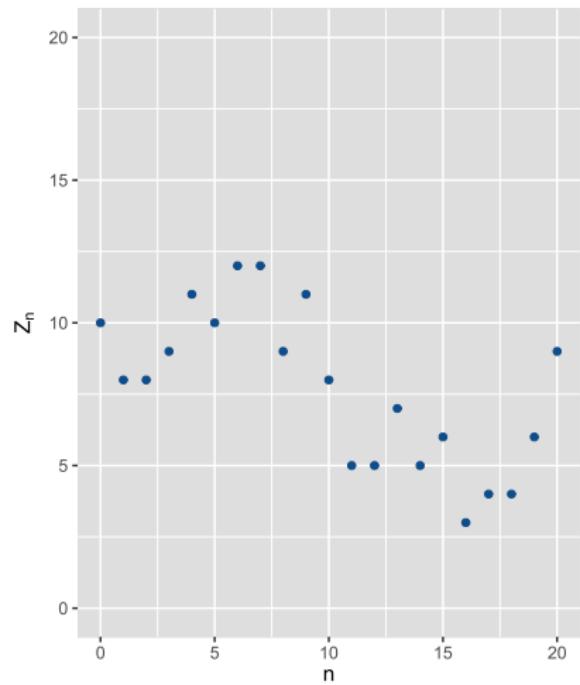


Procesos de Markov y su aplicación al modelo de Wright-Fisher

Gerardo Martínez

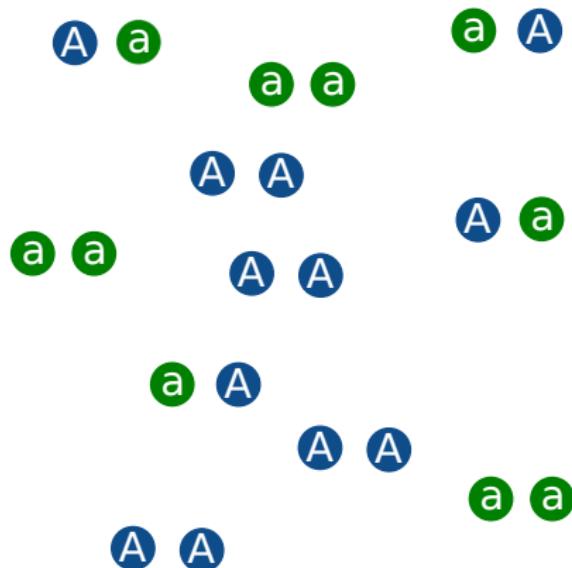
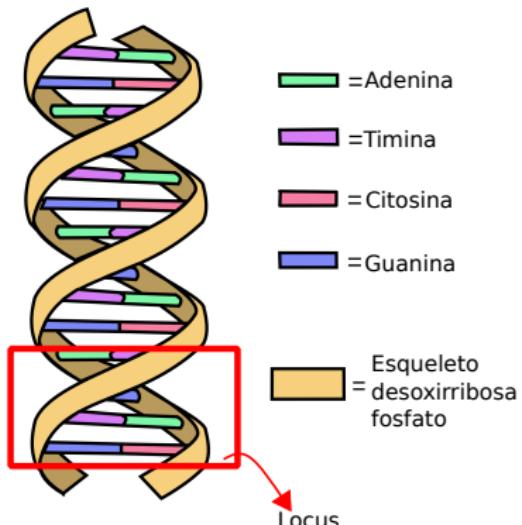
31 de octubre de 2019

Facultad de Ciencias Económicas y Administración — Facultad de Ingeniería - UdeLaR

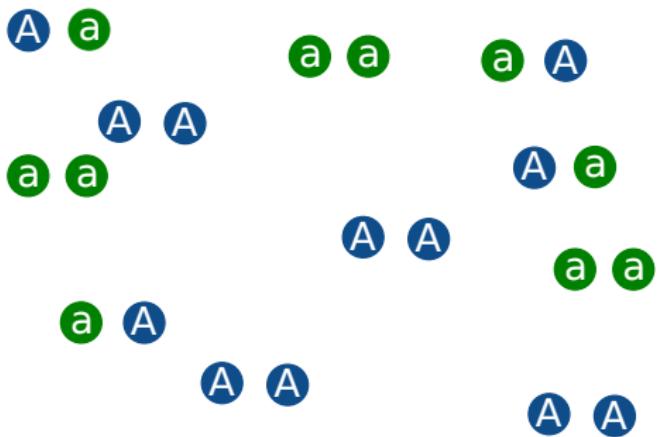


Introducción

- Población diploide de tamaño N .
- Locus en el genoma con dos alelos A y a .
- $2N$ copias de los alelos.

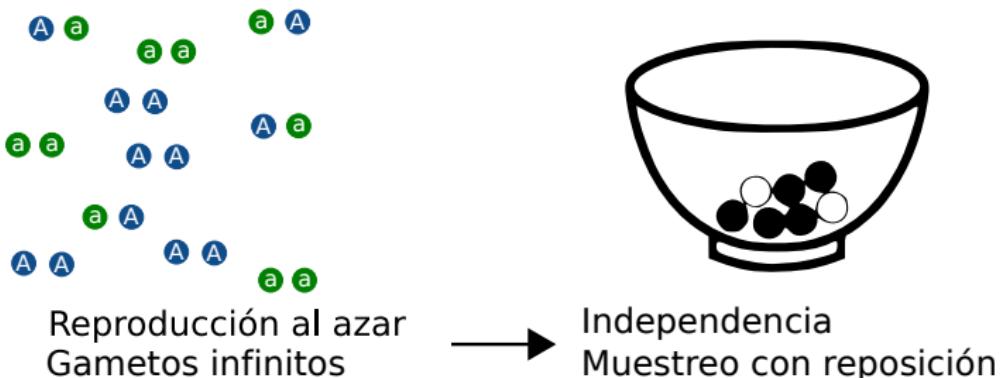


Supongamos que en la generación $n = 0$ la frecuencia del alelo A es $\frac{i}{2N}$ para $i \in \{0, 1, \dots, 2N\}$ y la frecuencia del alelo a es $1 - \frac{i}{2N}$



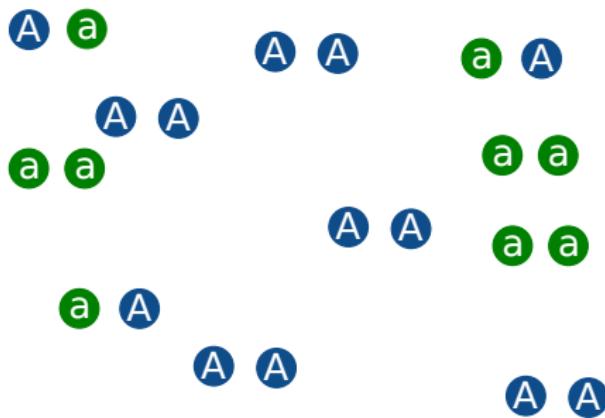
Frecuencia del A : $\frac{12}{22}$, frecuencia del a : $\frac{10}{22}$

Queremos construir ahora una población de tamaño $2N$ que será la generación $n = 1$.

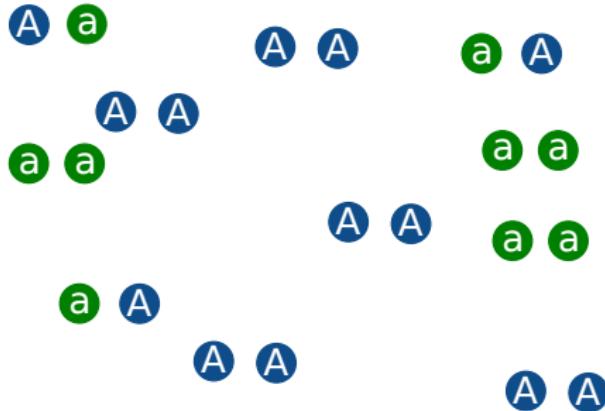


Con esta forma de reproducción, estamos interesados en conocer cuál es la frecuencia del alelo A en la generación $n = 1$ dada la configuración poblacional en $n = 0$.

Supongamos que se obtiene la siguiente población en la generación $n = 1$



Las frecuencias de los alelos A y a son $\frac{13}{22}$ y $\frac{9}{22}$, respectivamente.
¿Cuál es la probabilidad de obtener estas frecuencias a partir de la generación $n = 0$?



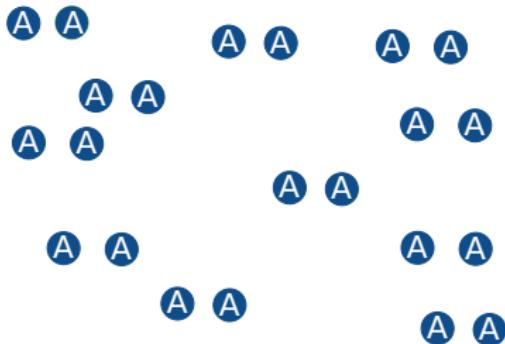
¿Cuál es la probabilidad de obtener estas frecuencias a partir de la generación $n = 0$?

$$\begin{aligned} P(\text{generación 1} | \text{generación 0}) &= \binom{22}{13} \left(\frac{12}{22} \right)^{13} \left(\frac{10}{22} \right)^9 \\ &\approx 0,1709505 \end{aligned}$$

- $Z_n : \Omega \rightarrow \{0, 1, \dots, 2N\}$ el número de alelos de tipo A en la generación n .
- $\{Z_n\}_{n \in \mathbb{N}}$ es una cadena de Márkov con espacio de estados $E = \{0, 1, \dots, 2N\}$.
- La matriz de transición $\mathbb{P} = ((p_{ij}))_{i,j=0,1,\dots,2N}$ es tal que

$$p_{ij} = \mathbb{P}(Z_{n+1} = j | Z_n = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

El proceso presenta dos estados absorbentes, a saber, 0 y $2N$. Esto se obtiene debido a la forma en que la generación en tiempo $n + 1$ a partir de la generación n .



Cuando ocurre esto decimos que un alelo *se fija*.

Esto nos conduce a algunas preguntas:

- ¿Cuál es la probabilidad de que se alcancen los estados absorbentes?
- ¿Cuál es el tiempo esperado para que esto ocurra?
- ¿Existe una distribución estacionaria de esta cadena?

Esto nos conduce a algunas preguntas:

- ¿Cuál es la probabilidad de que se alcancen los estados absorbentes?
- ¿Cuál es el tiempo esperado para que esto ocurra?
- ¿Existe una distribución estacionaria de esta cadena?

Esto nos conduce a algunas preguntas:

- ¿Cuál es la probabilidad de que se alcancen los estados absorbentes?
- ¿Cuál es el tiempo esperado para que esto ocurra?
- ¿Existe una distribución estacionaria de esta cadena?

- Presentar algunos conceptos de la teoría de procesos markovianos y de difusiones.
- Responder a las preguntas planteadas a través de la aproximación del modelo de Wright-Fisher discreto a través de una difusión límite.
 - Agregar nuevos parámetros al modelo y mostrar cómo esto modifica la difusión límite.

Procesos markovianos y ecuaciones diferenciales estocásticas

Cadenas de Markov: breve repaso

Definición: Cadena de Markov discreta

- (Ω, \mathcal{A}, P) un espacio de probabilidad
- $X = \{X_n\}_{n \in \mathbb{N}}$ proceso estocástico que toma variables en E numerable.

X es una **cadena de Markov**, si

$$\begin{aligned} P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) = \\ P(X_{n+1} = i_{n+1} | X_n = i_n), \end{aligned}$$

para todo $n \in \mathbb{N}$ e $i_0, i_1, \dots, i_{n-1}, i_n, i_{n+1} \in E$.

X es **homogénea** si

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i),$$

para todo $n \in \mathbb{N}$.

Definición: Matriz de transición

- X cadena de Markov homogénea definida en $E = \{1, 2, \dots, n\}$.

La matriz $\mathbb{P} = ((p_{ij})) \in \mathbb{R}^{n \times n}$ es la **matriz de transición** de X si

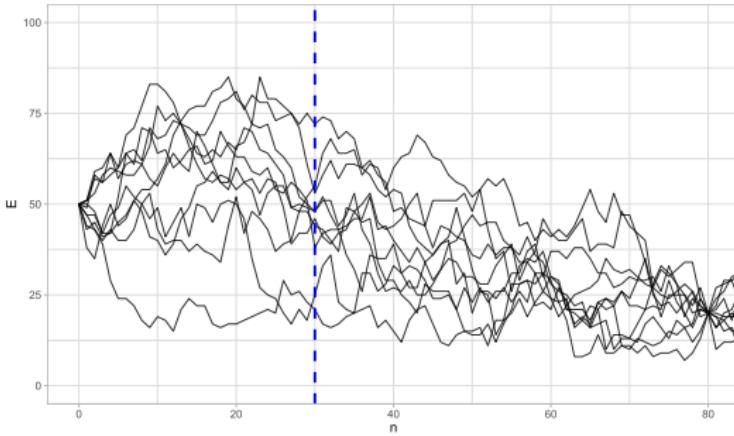
$$\mathbb{P} = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{pmatrix}$$

con $p_{ij} = P(X_1 = j | X_0 = i)$.

Proposición: Ecuación de Chapman-Kolmogorov

Dada una cadena de Márkov $X = \{X_n\}_{n \in \mathbb{N}}$ homogénea definida en un espacio de estados numerable E se cumple que

$$p_{ij}^{(n+m)} = \sum_{k \in E} p_{ik}^{(n)} p_{kj}^{(m)}$$



¿Cómo definir el análogo de cadena de Markov para un proceso $X = \{X_t\}_{t \geq 0}$ que toma valores en un espacio métrico (E, d) ?

Definición: Filtración

- (Ω, \mathcal{A}, P) un espacio de probabilidad.
- $\mathcal{F} = \{\mathcal{F}_t\}_{t \geq 0}$ es una **filtración** si
 1. \mathcal{F}_t es una sub- σ -álgebra de \mathcal{A} , para todo $t \in [0, +\infty)$, y
 2. dados $t, s \in [0, +\infty)$ se cumple que $\mathcal{F}_t \subset \mathcal{F}_{t+s}$.

Definición: Filtración natural

- $X = \{X_t\}_{t \geq 0}$ un proceso estocástico sobre un espacio de estados (E, \mathcal{E}) .
- $\mathcal{F}^X = \{\mathcal{F}_t\}_{t \geq 0}$ es la **filtración generada por X** o la **filtración natural de X** si

$$\mathcal{F}_t = \sigma\{X_s^{-1}(B) : s \leq t, B \in \mathcal{E}\}, \quad (1)$$

Definición: Proceso de Markov

- $(\Omega, \mathcal{A}, \mathbb{P})$ espacio de probabilidad.
- $X = \{X_t\}_{t \geq 0}$ un proceso estocástico que toma valores en (E, d) un espacio métrico.
- $\mathcal{F}^X = \{\mathcal{F}_t^X\}_{t \geq 0}$ la filtración natural de X .

X es un proceso **de Markov o markoviano** si

$$\mathbb{P}(X_{t+s} \in B | \mathcal{F}_t^X) = \mathbb{P}(X_{t+s} \in B | X_t),$$

para todos $s, t \geq 0$ y $B \in \mathcal{B}(E)$.

Un proceso de Markov es **homogéneo en el tiempo** si

$$\mathbb{P}(X_{t+h} \in B | X_t) = \mathbb{P}(X_h \in B | X_0)$$

para todo $B \in \mathcal{B}(E)$ y todo $t, h \geq 0$.

Definición: función de transición

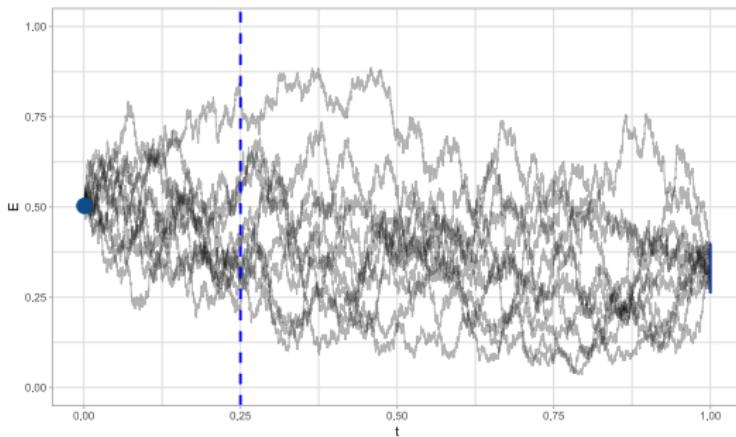
Sea E un espacio métrico y $\mathcal{B}(E)$ su σ -álgebra de Borel. Una función $P : [0, +\infty) \times E \times \mathcal{B}(E) \rightarrow \mathbb{R}$ es una **función de transición homogénea en el tiempo** si

1. fijados $x \in E$ y $t \in [0, +\infty)$, $P(t, x, \cdot)$ es una medida de probabilidad,
2. para todo $x \in E$, $P(0, x, \cdot) = \delta_x$, con δ_x la delta de Dirac en x ,
3. fijado $B \in \mathcal{B}(E)$, $P(\cdot, \cdot, B)$ es medible en $\mathcal{B}([0, +\infty) \times E)$, y

Definición: función de transición

4. para todos $s, t \geq 0, x \in E$, y $B \in \mathcal{B}(E)$ se cumple la llamada propiedad de Chapman-Kolmogorov

$$P(t+s, x, B) = \int_E P(s, y, B) P(t, x, dy).$$



Definición: función de transición asociada a un proceso de Markov
Una función de transición $P(t, x, B)$ es una función de transición de un proceso estocástico markoviano homogéneo en el tiempo X si

$$P(X_{t+s} \in B | \mathcal{F}_t^X) = P(X_{t+s} \in B | X_t) = P(s, X_t, B)$$

Proceso de Wiener

No es sencillo obtener fórmulas explícitas para las funciones de transición de procesos markovianos arbitrarios. Una excepción notable de esto es la función de transición de un movimiento browniano unidimensional:

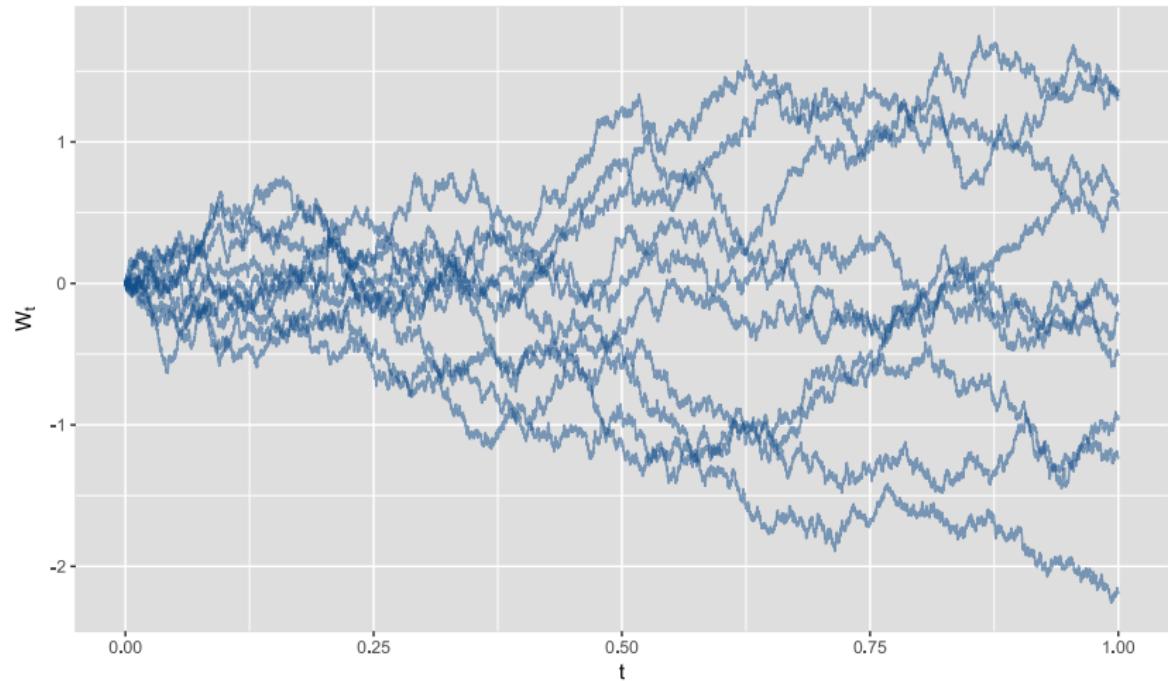
$$P(X_{t+s} \in B | X_t) = P(s, X_t, B) = \int_B \frac{1}{\sqrt{2\pi s}} e^{-\frac{(y-x)^2}{2s}} dy$$

Definición: Proceso de Wiener

Un proceso estocástico $W = \{W_t\}_{t \geq 0}$ con valores en \mathbb{R} es un **proceso de Wiener unidimensional** o un **movimiento browniano unidimensional** si

1. $W_0 = 0$ casi seguramente.
2. $P(\omega \in \Omega : W_t(\omega) \in C[0, +\infty)) = 1$.
3. El incremento $W_{t+h} - W_t$ es independiente de $\sigma(W_t)$ para todo $t, h \geq 0$.
4. $W_{t+h} - W_t \sim N(0, h)$ para todo $t, h \geq 0$.

Proceso de Wiener



Trayectorias de un proceso de Wiener con una precisión de $n = 10^5$.

Solución fuerte de una ecuación diferencial estocástica

Definición: Solución fuerte de una SDE

- $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad.
- Sean $W = \{W_t\}_{t \in [0, T]}$ un proceso de Wiener.
- ξ una variable aleatoria independiente de W .
- $\mathcal{F}_t = \sigma(\{\xi, W_s : 0 \leq s \leq t\})$ la σ -álgebra generada por ξ y el proceso de Wiener en el intervalo $[0, t]$.
- $a, b : \mathbb{R} \rightarrow \mathbb{R}$ funciones medibles.

Un proceso $X = \{X_t\}_{t \in [0, T]}$ es la **solución fuerte** de la ecuación diferencial estocástica

$$\begin{cases} dX_t = a(X_t)dt + b(X_t)dW_t \\ X_0 = \xi \end{cases}$$

si se cumple que

Solución fuerte de una ecuación diferencial estocástica

Definición: Solución fuerte de una SDE

- 1) X es un proceso con trayectorias continuas.
- 2) X es adaptado a \mathcal{F}_t .
- 3) Para todo $t \in [0, T]$ se cumple que

$$\int_0^t |a(X_s)| + |b(X_s)|^2 ds < \infty$$

- 4) Para todo $t \in [0, T]$ se cumple que

$$X_t = \xi + \int_0^t a(X_s) ds + \int_0^t b(X_s) dW_s$$

Difusiones

Definición: difusión homogénea

Sea $X = \{X_t\}_{t \in [0, T]}$ un proceso markoviano que toma valores en \mathbb{R} . El proceso X es una **difusión homogénea** si se cumplen las siguientes condiciones:

- 1) Para todo $\varepsilon > 0$, todo $t \in [0, T]$ y todo $x \in \mathbb{R}$ se cumple que

$$\lim_{h \rightarrow 0} P(|X_{t+h} - x| > \varepsilon | X_t = x) = 0$$

- 2) Existen funciones $a, b : \mathbb{R} \rightarrow \mathbb{R}$ para las cuales existe $\varepsilon > 0$ tal que para todo $t \in [0, T]$ y todo $x \in \mathbb{R}$ se cumple que

$$\lim_{h \rightarrow 0} \frac{\mathbb{E} [(X_{t+h} - x) 1_{[0, \varepsilon]} (|X_{t+h} - X_t|) | X_t = x]}{h} = a(x)$$

$$\lim_{h \rightarrow 0} \frac{\mathbb{E} [(X_{t+h} - x)^2 1_{[0, \varepsilon]} (|X_{t+h} - X_t|) | X_t = x]}{h} = b^2(x)$$

Proposición

Sean $a, b : \mathbb{R} \rightarrow \mathbb{R}$ funciones continuas. Sea X solución fuerte de

$$\begin{cases} dX_t = a(X_t)dt + b(X_t)dW_t \\ X_0 = \xi \end{cases}$$

Entonces X es una difusión con coeficiente de deriva $a(x)$ y coeficiente de difusión $b^2(x)$.

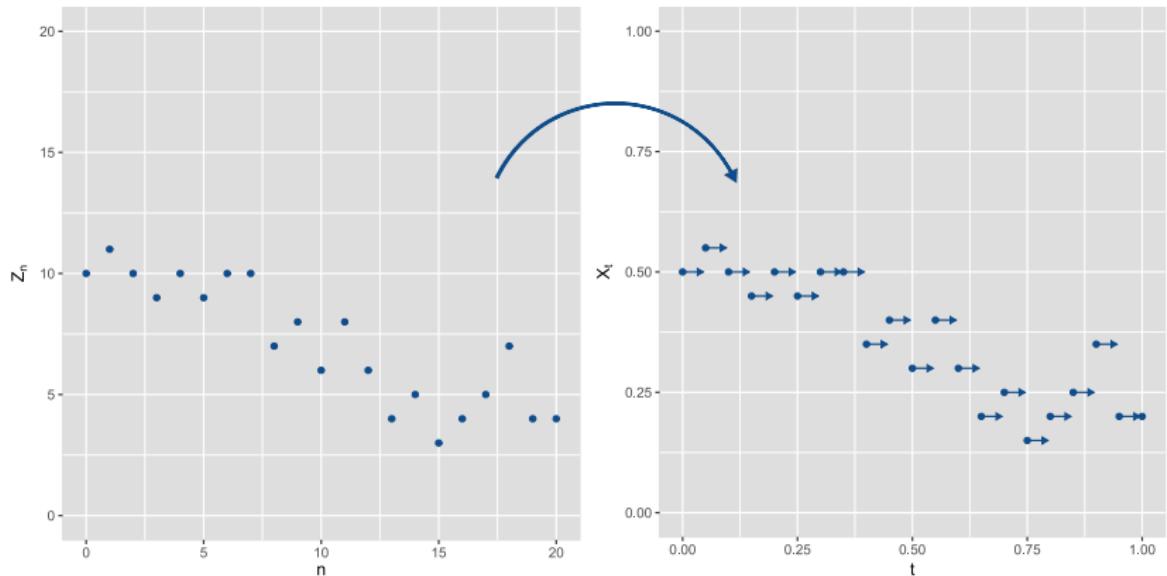
Aplicación de las difusiones al proceso de Wright-Fisher

Para cada $N \in \mathbb{N}$ se considera $Z^{(N)} = \{Z_n^{(N)}\}_{n \in \mathbb{N}}$ el proceso de Wright-Fisher de parámetros N y p_0 que toma valores en $E_N = \{0, 1, \dots, 2N\}$. Sea

$$X_t^{(N)} = \frac{1}{2N} Z_{\lfloor 2Nt \rfloor}^{(N)}$$

un reescalado de tiempo y espacio del proceso original.

Aproximación por difusión del proceso de Wright-Fisher



Reescalado de un proceso de Wright-Fisher en tiempo y espacio. A la izquierda, un proceso de Wright-Fisher $Z = \{Z_n\}$ de parámetros $N = 10$ y $z_0 = 10$. A la derecha, el proceso continuo $X = \{X_t\}_{t \geq 0}$ tal que $X_t = \frac{1}{2N}Z_{\lfloor 2Nt \rfloor}$.

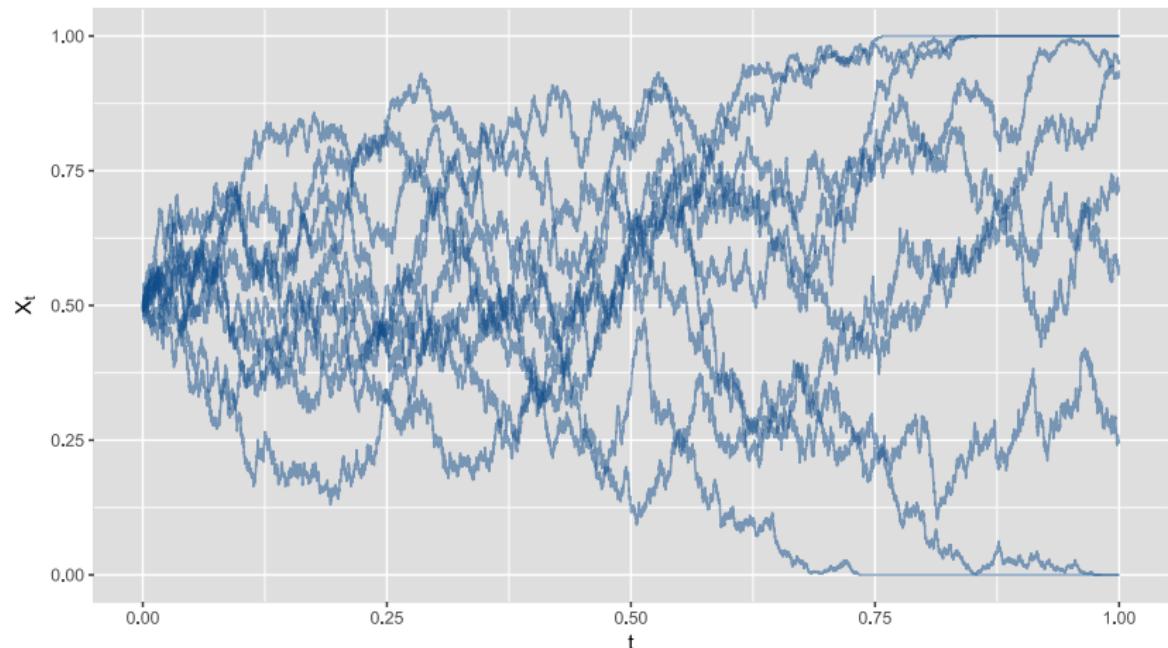
Sea $X = \{X_t\}$ la solución de la ecuación diferencial estocástica

$$\begin{cases} dX_t = \sqrt{X_t(1-X_t)} dW_t \\ X_0 = p_0 \in (0, 1) \end{cases}$$

Entonces X_N converge débilmente a X en $D_{[0,1]}[0, +\infty)$ con

$$D_{[0,1]}[0, +\infty) = \{f : [0, +\infty) \rightarrow [0, 1] : f \text{ es càdlàg}\}$$

Aproximación por difusión del proceso de Wright-Fisher



Trayectorias de la difusión límite de un proceso de Wright-Fisher en $[0, 1]$.

- ¿Cuál es la probabilidad de que se alcancen los estados absorbentes?
- ¿Cuál es el tiempo esperado para que esto ocurra?
- ¿Existe una distribución estacionaria de esta cadena?

Tiempo esperado de absorción

Sea $\tau = \inf_{t \geq 0} \{X_t \in \{0, 1\}\}$ el tiempo esperado para llegar a un estado absorbente de la difusión límite del proceso de Wright-Fisher.

Por un resultado de la teoría de difusiones se tiene que

$$\mathbb{E}_x(\tau) = g(x)$$

es la solución de la ecuación diferencial

$$\begin{cases} \frac{1}{2}x(1-x)g''(x) = -1 \\ g(1) = g(0) = 0 \end{cases}$$

Consecuentemente,

$$\mathbb{E}_x(\tau) = -2[x \log(x) + (1-x) \log(1-x)]$$

Tiempo esperado de absorción

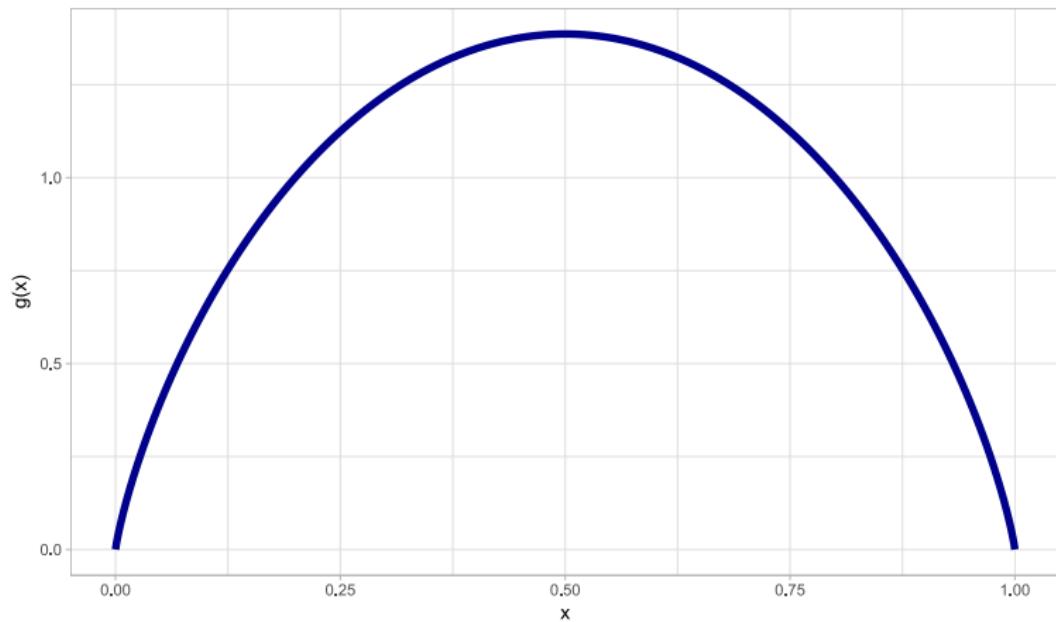
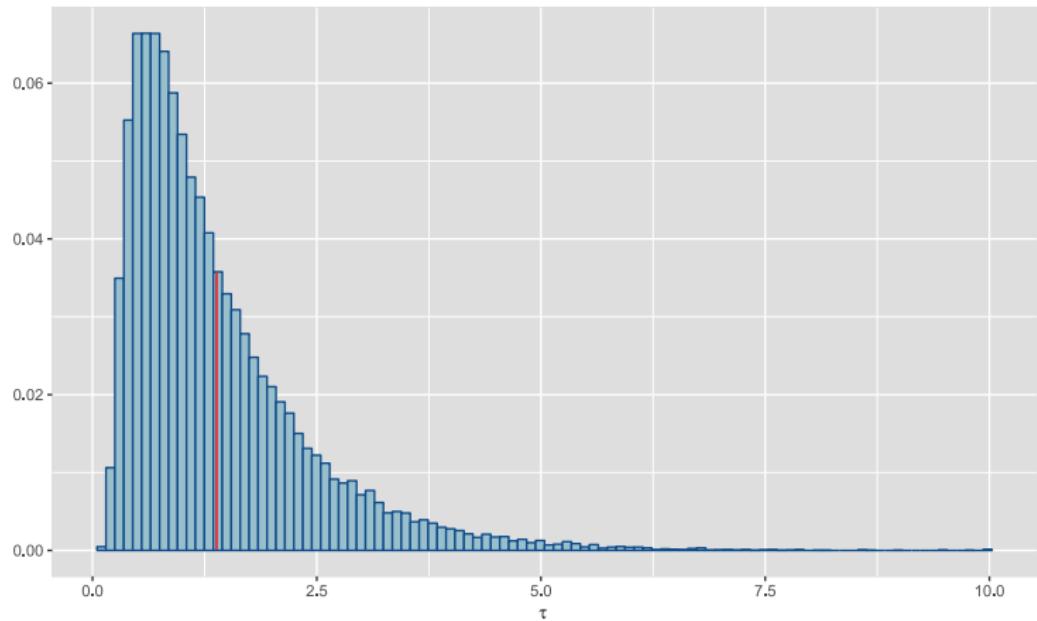


Gráfico de $g(x) = -2[x \log(x) + (1-x) \log(1-x)]$.

Tiempo esperado de absorción

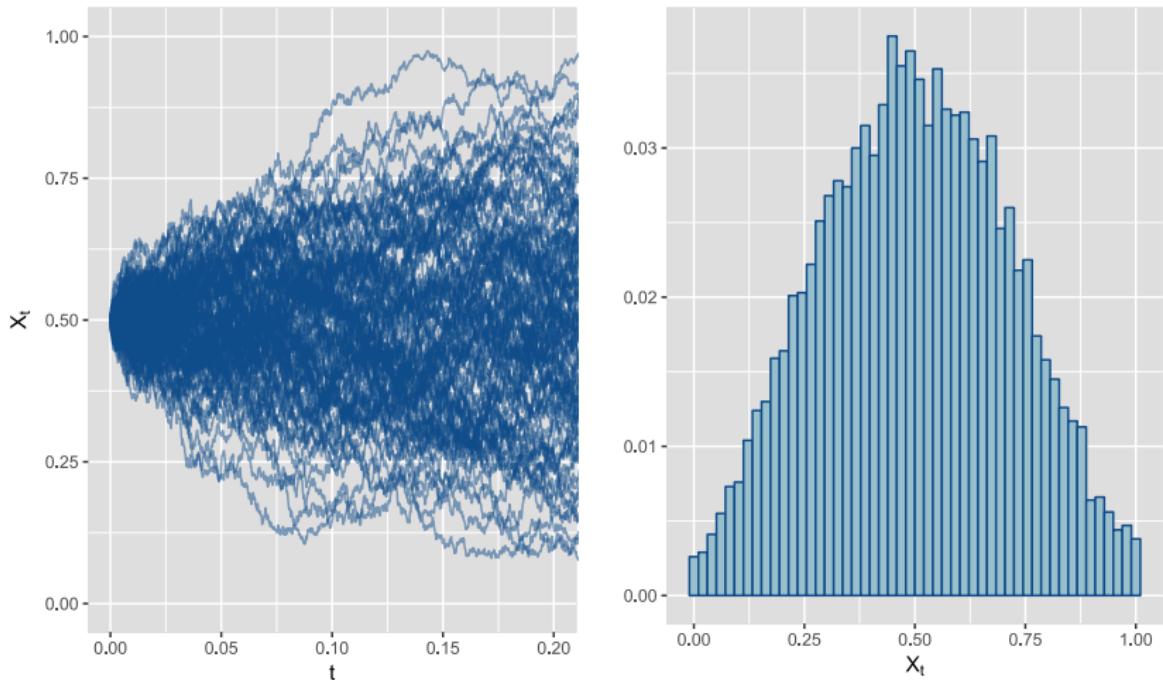


Densidad estimada $\tau = \inf\{t \geq 0 : X_t \in \{0, 1\}\}$ para el proceso de difusión $dX_t = \sqrt{X_t(1-X_t)}dW_t$ y $X_0 = 0, 5$. En rojo, $\mathbb{E}(\tau) = -2 \log(1/2) \approx 1,38$. Para la estimación de la densidad se simularon 40,000 observaciones mediante el esquema de Milstein.

Distribución de X_t

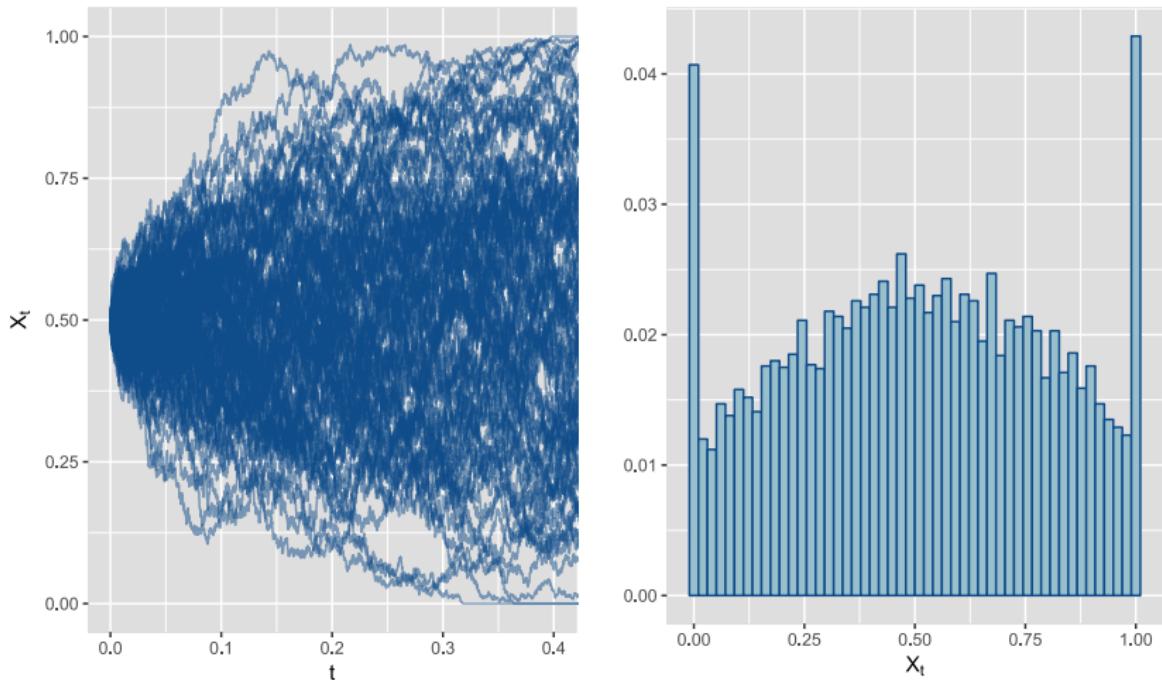
Estamos interesados ahora en encontrar la distribución del proceso de difusión $X = \{X_t\}$ que aproxima un proceso de Wright-Fisher para cualquier $t \geq 0$ y, en caso de que ésta exista, encontrar la distribución límite del proceso.

Distribución de X_t



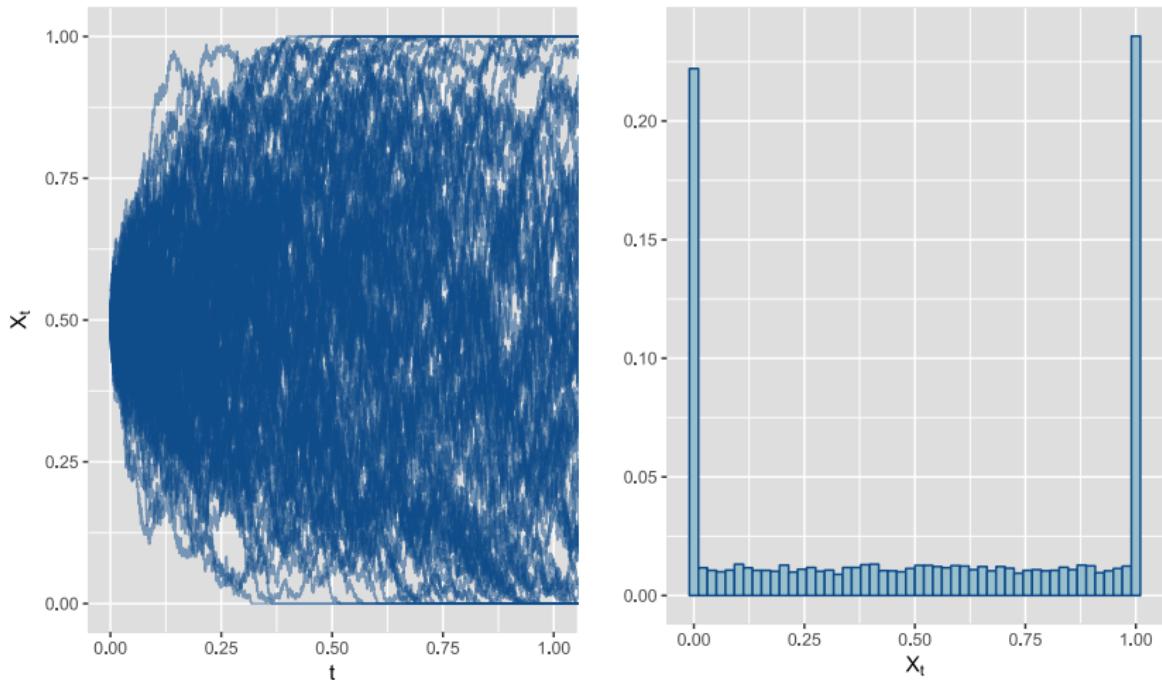
A la izquierda, trayectorias simuladas de $X = \{X_t\}_{t \geq 0}$ en $[0, 0,2]$. A la derecha, densidad estimada de X_t con $t = 0,2$.

Distribución de X_t



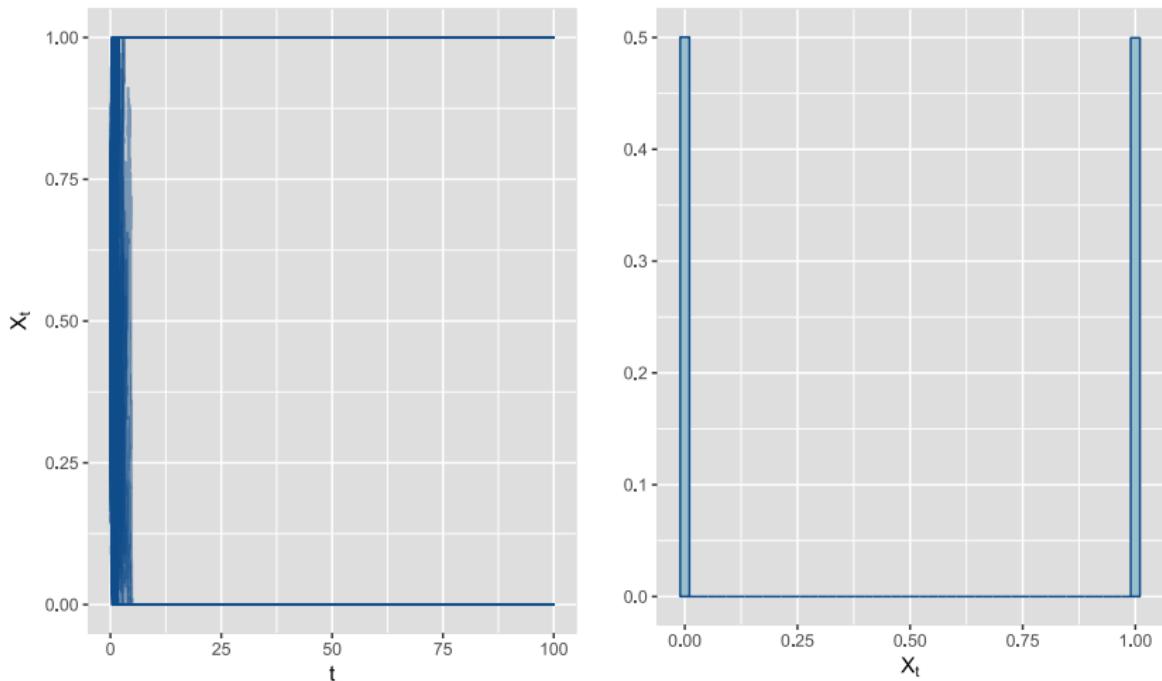
A la izquierda, trayectorias simuladas de $X = \{X_t\}_{t \geq 0}$ en $[0, 0,4]$. A la derecha, densidad estimada de X_t con $t = 0,4$.

Distribución de X_t



A la izquierda, trayectorias simuladas de $X = \{X_t\}_{t \geq 0}$ en $[0, 1]$. A la derecha, densidad estimada de X_t con $t = 1$.

Distribución de X_t



A la izquierda, trayectorias simuladas de $X = \{X_t\}_{t \geq 0}$ en $[0, 100]$. A la derecha, densidad estimada de X_t con $t = 100$.

¿Qué pasa cuando se agrega selección natural y tasas de mutación positivas al modelo?

- Población diploide de tamaño N .
- Locus en el genoma con dos alelos A y a .
- $2N$ copias de los alelos.
- $\omega = (\omega_{11}, \omega_{12}, \omega_{22})$: vector de probabilidades de supervivencia.



- $\mu = (\mu_1, \mu_2)$: probabilidades de mutación.

$$\begin{array}{c} \mu_1 \quad \mu_2 \\ A \rightarrow a \quad a \rightarrow A \end{array}$$

Modelo de Wright-Fisher con selección y mutación

- $Z = \{Z_n\}_{n \in \mathbb{N}}$ donde Z_n es el número de alelos A en la generación n .
- Z es una cadena de Markov con espacio de estados $\{0, 1, \dots, 2N\}$.
- La matriz de transición $\mathbb{P} = ((p_{ij}))_{i,j=0,1,\dots,2N}$ es tal que

$$p_{ij} = \mathbb{P}(Z_{n+1} = j | Z_n = i) = \binom{2N}{j} (p_{\text{sel}})^j (1 - p_{\text{sel}})^{2N-j}$$

con

$$p_{\text{sel}} = \frac{p_{\text{mut}}^2 w_{11} + p_{\text{mut}}(1 - p_{\text{mut}})w_{12}}{p_{\text{mut}}^2 w_{11} + 2p_{\text{mut}}(1 - p_{\text{mut}})w_{12} + (1 - p_{\text{mut}})^2 w_{22}}$$

y

$$p_{\text{mut}} = \frac{i}{2N} (1 - \mu_1) + \left(1 - \frac{i}{2N}\right) \mu_2.$$

Aproximación por difusión

$\omega = (\omega_{11}, \omega_{12}, \omega_{22})$ un vector de viabilidades y $\mu = (\mu_1, \mu_2)$ un vector de probabilidades de mutación. Para cada $N \in \mathbb{N}$ consideraremos el proceso de Wright-Fisher con selección de parámetros N, ω, μ y p_0 , $Z^{(N)} = \{Z_n^{(N)}\}_{n \in \mathbb{N}}$. Asumiremos que las viabilidades ω_{ij} son $O(1/2N)$ realizando la reparametrización

$$\omega_{ij} = 1 + \frac{\sigma_{ij}}{2N}.$$

Consideraremos también que las mutaciones μ_i son también $O(1/2N)$ realizando la reparametrización

$$\mu_i = \frac{u_i}{2N}$$

Aproximación por difusión del proceso de Wright-Fisher

Para cada $N \in \mathbb{N}$ se considera $Z^{(N)} = \{Z_n^{(N)}\}_{n \in \mathbb{N}}$ el proceso de Wright-Fisher de parámetros N y p_0 que toma valores en $E_N = \{0, 1, \dots, 2N\}$. Sea

$$X_t^{(N)} = \frac{1}{2N} Z_{\lfloor 2Nt \rfloor}^{(N)}$$

un reescalado de tiempo y espacio del proceso original.

Sea $X = \{X_t\}$ la solución de la ecuación diferencial estocástica

$$\begin{cases} dX_t = (-u_1 X_t + u_2(1 - X_t) + \\ \quad X_t^2(1 - X_t)\sigma_{11} + X_t(1 - X_t)(1 - 2X_t)\sigma_{12} - X_t(1 - X_t)^2\sigma_{22})dt \\ \quad + \sqrt{X_t(1 - X_t)} dW_t \\ X_0 = p_0 \in (0, 1) \end{cases}$$

Entonces X_N converge débilmente a X en $D_{[0,1]}[0, +\infty)$.

Se distinguen dos casos de interés: la difusión que cumple que $u_1 = u_2 = 0$ y la que cumple que existe $u_j > 0$ para $j \in \{0, 1\}$.

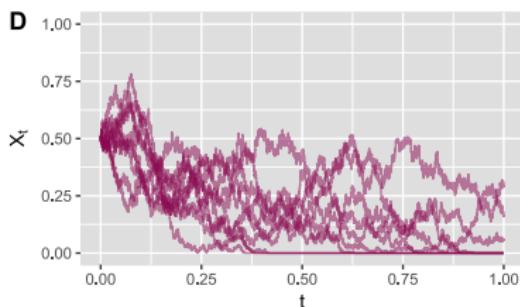
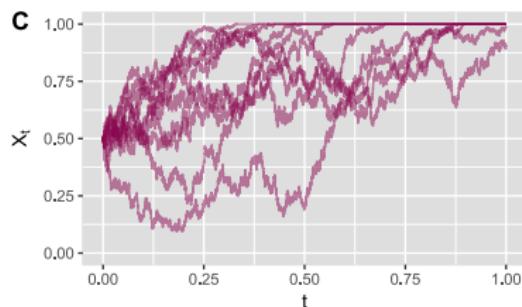
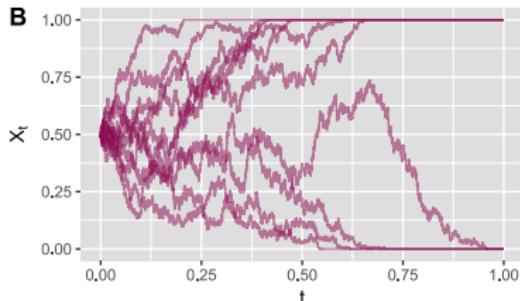
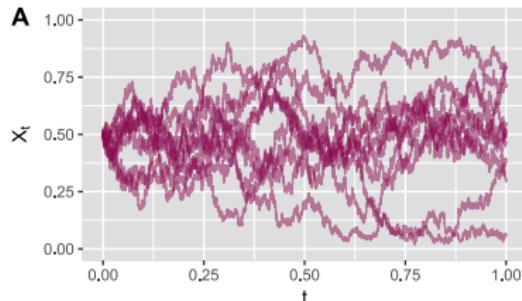
Caso

u_1

=

u_2

0



Trayectorias de la difusión $X = \{X_t\}_{t \geq 0}$ correspondiente al proceso de Wright-Fisher con selección. (A) Selección balanceadora. (B) Selección disruptiva. (C) A dominante. (D) a dominante.

Caso

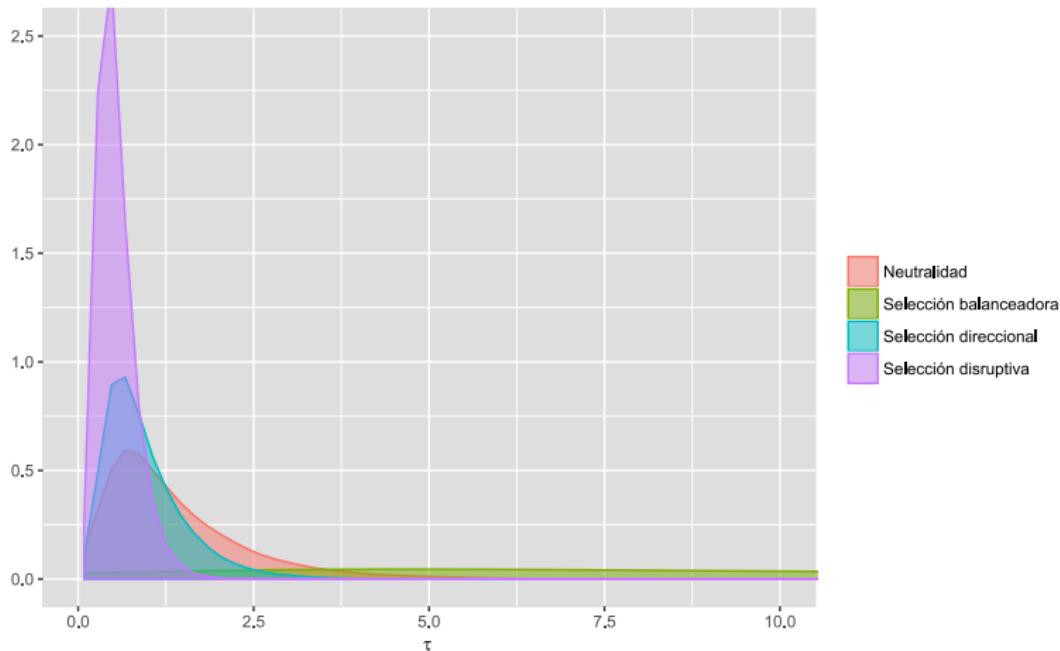
u_1

=

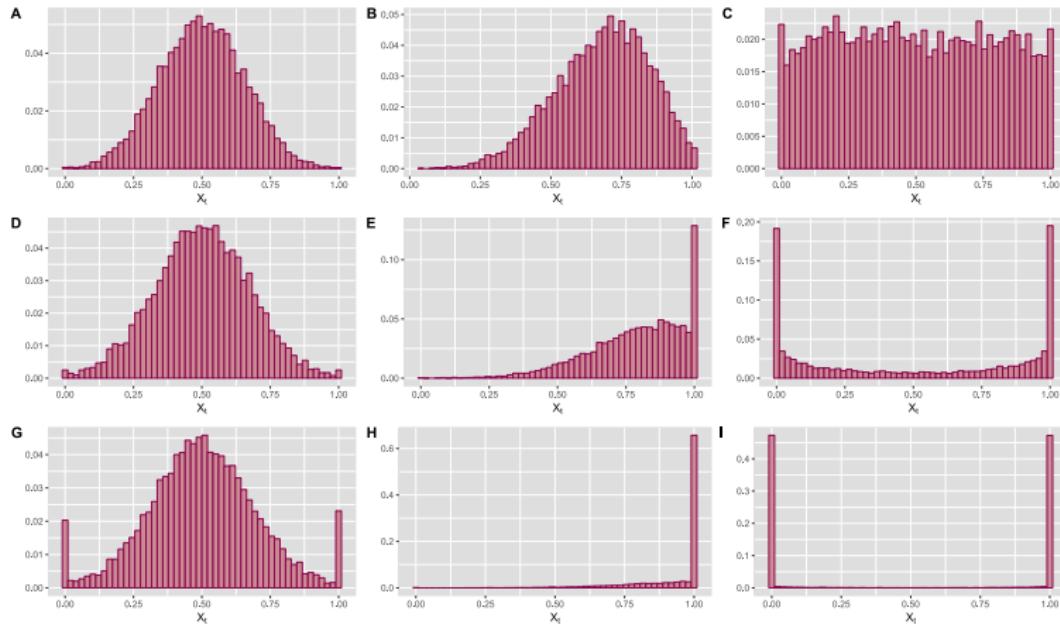
u_2

=

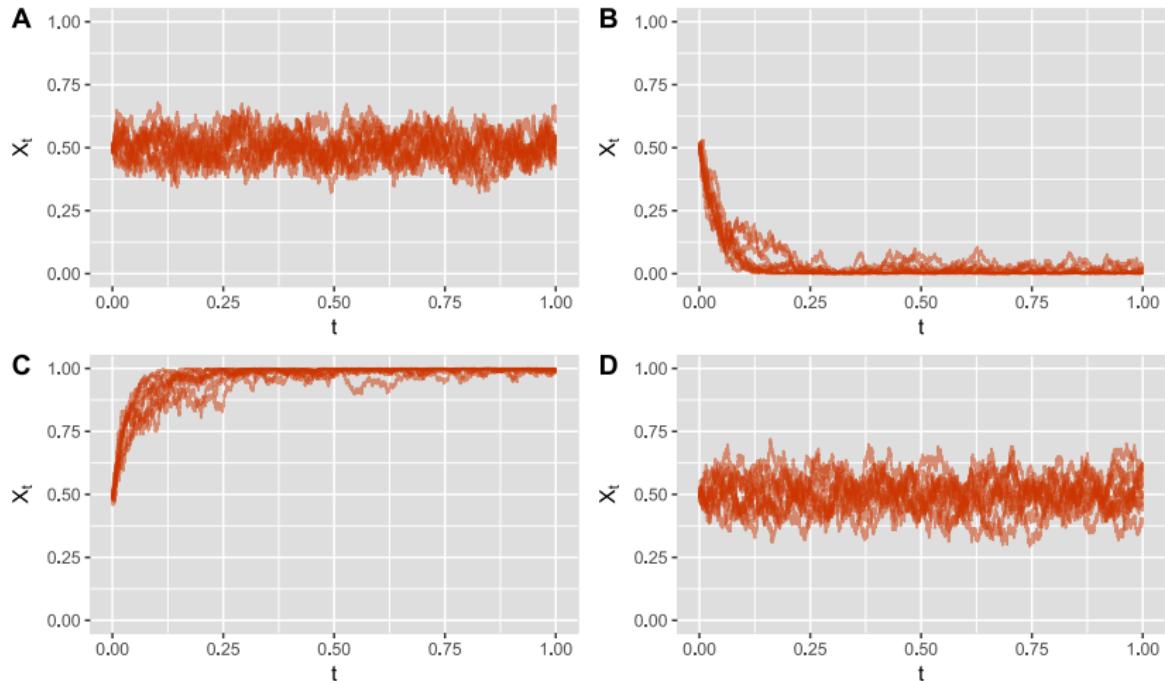
0



Densidad estimada de τ para diversos esquemas de selección. Los valores de τ fueron obtenidos simulando trayectorias de la difusión límite del proceso de Wright-Fisher con selección, $X = \{X_t\}_{t \geq 0}$, mediante el esquema de Milstein con una precisión de $n = 10^6$.



Densidad estimada de X_t para diversos esquemas de selección. Los valores de X_t fueron obtenidos simulando trayectorias de la difusión límite del proceso de Wright-Fisher con selección, $X = \{X_t\}_{t \geq 0}$, mediante el esquema de Milstein con una precisión de $n = 10^6$. La fila superior corresponde a la densidad de X_t para $t = 0, 2$; la fila central corresponde a la densidad de X_t para $t = 0, 4$; la fila inferior corresponde a la densidad de X_t para $t = 1$. Figuras A, D y G: selección balanceadora. Figuras B, E y H: selección direccional. Figuras C, F y I: selección disruptiva.



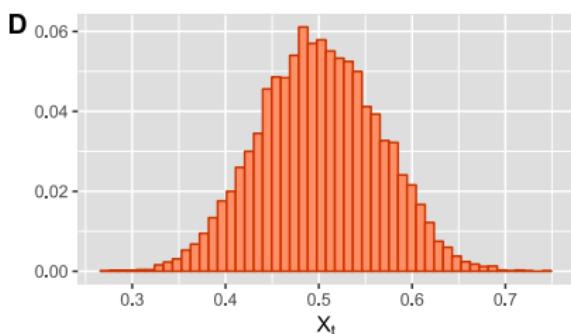
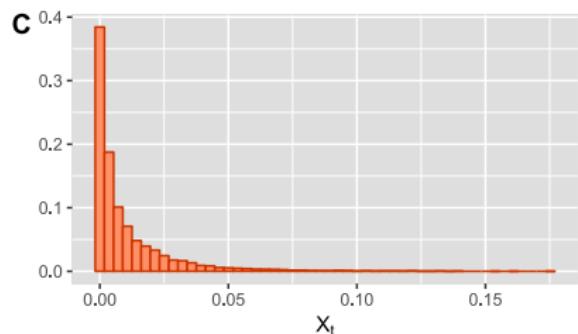
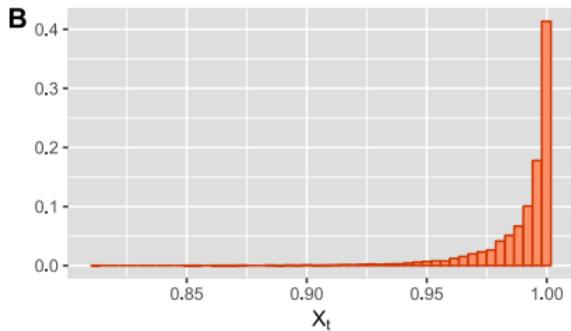
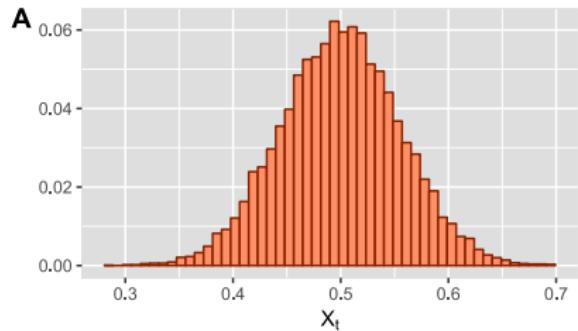
Trayectorias de la difusión $X = \{X_t\}_{t \geq 0}$ correspondiente al proceso de Wright-Fisher con selección y mutación. (A) $u_1 = u_2 > 0$. (B) $u_1 > u_2 > 0$ (C) $0u_1 < u_2$. (D) $u_1 = u_2 = \sigma_{11} = \sigma_{22} > 0$.

Caso

u_j

>

0



Densidad estimada de la distribución estacionaria de la difusión $X = \{X_t\}_{t \geq 0}$. (A) $u_1 = u_2 > 0$. (B) $u_1 > u_2 > 0$ (C) $0u_1 < u_2$. (D) $u_1 = u_2 = \sigma_{11} = \sigma_{22} > 0$.

- Estudiar de forma teórica $P(X_t = 0 \vee X_t = 1 | X_0 = x)$ para todo t .
- ¿Existe una distribución teórica para τ dado $X_0 = x$ con $x \in (0, 1)$?
- Quitar otras hipótesis al modelo: población cerrada, generaciones solapadas, selección natural constante en el tiempo, gen autosómico.
- ¿Qué pasa en el caso cuando un locus es multialélico?

¡Muchas gracias!

El código para reproducir las gráficas de esta presentación puede verse aquí.