UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA
ESCUELA DE MATEMÁTICA

TRABAJO DE GRADUACIÓN:

PROCESOS ESTOCÁSTICOS CON APLICACIONES A GENÉTICA DE POBLACIONES

PRESENTADO POR:

WILBER ALEXANDER ORTIZ CORTEZ

PARA OPTAR AL GRADO DE:

LICENCIATURA EN MATEMÁTICA

Ciudad Universitaria, 8 de Noviembre de 2019

UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA
ESCUELA DE MATEMÁTICA

TRABAJO DE GRADUACIÓN:

PROCESOS ESTOCÁSTICOS CON APLICACIONES A GENÉTICA DE POBLACIONES

PRESENTADO POR:

WILBER ALEXANDER ORTIZ CORTEZ

ASESORES:

DR. ADRIÁN GONZÁLEZ CASANOVA.

MSc.CARLOS ERNESTO GÁMEZ RODRÍGUEZ.

Ciudad Universitaria, 8 de Noviembre de 2019

AUTORIDADES

UNIVERSIDAD DE EL SALVADOR

RECTOR UNIVERSITARIO:

MSc. ROGER ARMANDO ARIAS ALVARADO

SECRETARIA GENERAL:

LIC. CRISTOBAL HERNÁN RÍOS BENITEZ

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

DECANO:

LIC. MAURICIO ERNESTO HERNÁN LOVO

SECRETARIA:

LIC. MELANY TURCIOS

ESCUELA DE MATEMÁTICA

DIRECTOR:

DR.NERYS FUNES

SECRETARIA:

LIC. ALBA IDALIA CUELLAR

DEDICATORIA

A mi madre por todo su amor, cariño, consejos, paciencia y apoyo incondicional en todos estos años de estudios.

AGRADECIMIENTOS

Agradezco a Dios por la fuerza, sabiduría y salud que me ha dado en todos esto años de estudio su bondad y misericordia fueron, son y serán nuevas cada día.

A mi amada madre por su apoyo incondicional, sus consejos y su paciencia, gracias por absolutamente todo. También a mi novia por ser el motor que impulsa mi vida, por acompañarme siempre con sus consejos en los momentos difíciles.

Mi mayor agradecimiento para el Dr. Adrián Gonzalez Casanova por su apoyo académico y poder compartir sus conocimientos, sin usted no hubiera sido posible terminar, ni empezar esta tesis. Además debo de agradecer mucho al MSc. Carlos Ernesto Gamez por su apoyo y paciencia que me demostró cada semana en las reuniones de asesorías.

A todos ustedes,
INFINITAS GRACIAS

Índice general

Capitulos Página

RESUMEN

Muchos relacionan un proceso estocástico a una serie de acontecimientos aleatoria y en algún grado están en lo cierto. Existen procesos llamados determinísticos, son procesos en los que conociendo las condiciones iniciales siempre siguen el mismo curso y producen el mismo resultado final, es decir elementos no aleatorios están presentes, podemos predecir en el tiempo todos los posibles estados y el estado final siempre será el mismo dado unas mismas condiciones iniciales. En procesos estocásticos no estamos presente a un simple curso de acontecimientos en el tiempo, es decir que hay un grado de indeterminación en los posibles cursos o secuencias de pasos que tome el proceso y este escenario múltiple puede ser descrito por distribuciones y densidades probabilísticas. En un proceso estocástico aún cuando tenemos las mismas condiciones iniciales en el comienzo hay diferente secuencias de acontecimientos que el proceso puede tomar y por ende se puede arribar a diferentes estados finales partiendo de unas mismas condiciones iniciales. Es importante notar que todos estas secuencias no poseen la misma probabilidad de ocurrir en el tiempo, algunos pueden acontecer mas frecuente que otros. Por eso en sistemas donde ocurren procesos estocásticos es importante identificar y caracterizar cada posible secuencia de acontecimiento por su probabilidad de ocurrir. En general, todo proceso estocástico es sometido a análisis de probabilidades.

En estos capítulos introductorios de procesos estocásticos en genética de poblaciones el objetivo del trabajo es dar una visión general de los antecedentes teóricos de conceptos

de dualidad, coalescente de Kigman, difusión de Wright-Fisher, generadores, dualidad de generadores de los procesos de Markov y presentar conexiones fundamentales de manera unificada, para concluir con el modelo de Wright-Fisher generalizado, es decir con población variable, para ello será de mucha utilidad la convergencia en probabilidad de los procesos de frecuencia y ancestría con límite de escala cuando N tiende al infinito. También tratamos de ayudar a comprender resultados como la frecuencia de una población, el tiempo hasta el ancestro común (MRCA) o el tiempo de fijación de una población. Esperamos que sirva de referencia para los estudiantes de matemática con énfasis en la Biomatemática. Por último, pero no menos importante, espero que este trabajo desencadene nuevas investigaciones en este área multifacética y ampliamente aplicable de la teoría de la probabilidad.

INTRODUCCIÓN

El modelo de Wright-Fisher es un modelo clásico dentro de la genética de poblaciones. Éste modela la evolución de la frecuencia de un alelo en una población de tamaño N como una cadena de Markov discreta y con espacio de estados finitos. La dinámica genética de poblaciones es un proceso propio de aleatoriedad, por lo que el uso de modelos estocásticos para su estudio resulta natural y eficiente. Desde hace más de 100 años la genética de poblaciones es un área en la frontera entre matemáticas y biología. En esta investigación revisaremos y estudiaremos un modelo clásico que se han construido en el área. El modelo estudia como cambia el perfil genético de una población bajo la hipótesis de que cada individuo es del mismo tipo que su padre y de que en cada generación hay la misma cantidad de individuos.

La genética de poblaciones depende de modelos matemáticos lo cual permite hacer predicciones. Los genetistas construyen modelos matemáticos abstractos que incorporan los efectos de la selección, deriva genética, flujo génico y mutación sobre la dinámica de la frecuencia génica. A partir de estos modelos se extraen conclusiones sobre los posibles patrones de la variabilidad genética en las poblaciones reales.

Se puede predecir la frecuencia alélica de una generación en función de la frecuencia de la generación anterior. En cambio, la deriva genética es un proceso microevolutivo de tipo aleatorio o estocástico, admite un tratamiento matemático descrito en término de probabilidades. Conociendo la frecuencia alélica de una generación no se puede saber la

frecuencia alélica de la siguiente generación. Se puede predecir, eso sí, la probabilidad de tener cierta frecuencia alélica

La versión más fácil del modelo de Wright-Fisher data de la década de 1940, es decir, antes del descubrimiento de la estructura de doble hélice de ADN por Crick y Watson en 1950 (que proporciona la base molecular de la evolución). Sin embargo, se sabía mucho antes de 1950 que el ADN es el portador de información hereditaria. El modelo de Wright-Fisher se utiliza para describir la evolución de una población de individuos, de dos tipos diferentes, denotados por $+$ y $-$. Estos tipos son neutrales, es decir, su éxito reproductivo no depende del tipo, y su reproducción es aleatorio.

Además, estudiaremos con una visión general de los antecedentes teóricos y así conectarlo con los fundamentos de manera unificada del concepto de dualidad en los procesos de Markov. Cabe mencionar que la dualidad de los procesos de Markov con respecto a una función apareció por primera vez en la literatura a finales de los años 40 y principios de los 50, y se ha formalizado y generalizado durante las siguientes décadas. Desde entonces se ha aplicado en una variedad de campos que van desde sistemas de partículas interactivas, teoría de colas, procesos a genética de poblaciones entre otros.

Capítulo 1

Definición del Problema

1.1. Planteamiento del Problema

En el Modelo Wright-Fisher, se ha supuesto que el tamaño de la población es finito y constante. En cada unidad de tiempo, cada individuo elige aleatoriamente a uno de los individuos de la generación anterior, y adopta su tipo. Esta es una forma de reproducción aleatoria. Hay varias preguntas interesantes sobre el modelo Wright-Fisher. Para investigar, tanto en dirección hacia adelante como hacia atrás en el tiempo. Algunas preguntas son:

1. El Modelo de Wrigth-Fisher con tamaño de población variable, ¿es aplicable para cualquier tipo de población?
2. Dada la frecuencia de un individuo en cualquier generación g, ¿cuál es la probabilidad de que el tipo se extinga ó domine?
3. ¿Cuál es la esperanza del tiempo que tarda uno de los individuos con su frecuencia en fijarse en la población?

4. ¿Cómo debe identificarse los procesos de frecuencia y ancestría usando límite en el tiempo y comprobarse la convergencia?

5. Conociendo una muestra de la población en la genración g, ¿hace cuánto tiempo vivió el ancestro común más reciente (MRCA: the most recent common ancestor)?

6. ¿Cuál es la distribución de la variable aleatoria MRCA?

Interesa ver los fundamento de los Procesos Estocástico , la teoría de la estructura Matemática que se encuentra en la Genética de Poblaciones, y así poder dar respuesta a las interrogantes antes citadas.

1.2. Antecedentes

- Valdez Álvarez, Ana Beatriz (2006) Modelos epidemiológicos determinísticos. Tesis de Licenciatura, Universidad de El Salvador.

- Herrera Polanco, Diana Marcela (2013) Ajuste de un modelo SEIR para la tuberculosis en El Salvador. Tesis de Licenciatura, Universidad de El Salvador.

- Chávez Mancia, Javier Eliseo y Santos Nolasco, Mario Israel (2015) Aplicación del análisis discriminante para la detección de factores de riesgo en pacientes con diabetes mellitus en la región del bajo Lempa de El Salvador. Tesis de Licenciatura, Universidad de El Salvador.

- Moreno Ramos, María Elizabeth y Campos Martínez, Karen Brizeida (2015) Métodos de optimización para estimación de parámetros en modelos epidemiológicos. Tesis de Licenciatura, Universidad de El Salvador

1.3. Justificación

La relación entre la Matemática y la Biología no ha sido tan estrecha como debería. Con trabajos de muchos investigadores se ha ido observando la necesidad de relacionar estas dos importantes ramas del conocimiento, y en especial, durante el último siglo se han hecho grandes avances en cuanto a este propósito. Para mantener esta tendencia es necesario conocer cada vez más la utilidad y las posibilidades que se generan al construir estos vínculos. En biología se ha tratado de explicar tanto las bases evolutivas como la genética y formas de que los animales se especializan y logren sacar el máximo provecho a su condición. No es azar, ni suerte si no que hay leyes matemáticas detrás de los saltos evolutivos. Las abejas no construyen colmenas hexagonales (más bien prismáticas) por azar, sino que esa forma les permite almacenar la mayor cantidad de miel con la menor cantidad posible de cera. Las cornamentas de muchos animales presentan formas que obedecen a las matemáticas y a las leyes físicas, lo mismo que las rayas de una cebra o las manchas de un leopardo.

Darwin, D´Arcy Thompson o el mismo Mendel trabajaban con las matemáticas para explicar la evolución y los principios básicos del ADN. Hoy en día, por ejemplo, las leyes de Mendel son un hito en el desarrollo de la teoría de la probabilidad y la estadística. En otras palabras en la biología hay muchas matemáticas para ser explicada. Las matemáticas están presentes en muchos más aspectos de la biología, y han dado lugar a lo que ahora se llama biología matemática. Robert Malthus, por ejemplo, encontró la ley matemática que analiza el crecimiento de una población, ya sean seres humanos o bacterias en un cultivo. Las matemáticas y la biología van de la mano desde hace ya muchos años y ese encuentro promete ser cada vez más fructífero. Con esta trabajo de investigación se enriquecerá el área de la matemática aplicada específicamente la de matemática biológica que muy poco o casi nada se a investigado en nuestra escuela de matemática.

1.4. Objetivos

1.4.1. Generales:

- Ilustrar la estructura matemática que une los Procesos Estocásticos y Genética de Poblaciones.
- Comprender el efecto de fuerza evolutiva, como lo es deriva genética, sobre el cambio en las frecuencias alélicas dentro de una población.

1.4.2. Específicos:

- Proporcionar respuestas a preguntas del modelo Wright-Fisher, para describir los procesos estocásticos que surgen a lo largo del tiempo.
- Analizar como cambia el perfil genético de una población con tamaño constante y con población de tamaño variable.
- Mostrar como el Modelo de Wright-Fisher estudia en dirección al futuro o en dirección al pasado.

1.5. Metodología de Investigación

- *TIPO DE INVESTIGACIÓN.*

  La investigación es bibliográfica y esta basada en las notas del curso Estructuras Probabilísticas en Evolución impartido en la UNAM por el Dr. Adrián González Casanova, y además se ha hecho una recopilación de libros, artículos, y otras fuentes que han permitido obtener información relevante del tema. Todo esto para recopilar de manera ordenada la información más útil y destacada del tema.

- *MANERA DE TRABAJO.*

  Tendré reuniones semanales con mi Asesor interno para discutir las dudas y revisar los avances cada semana y con mi Asesor externo poder hacer videollamadas cada dos semanas y cada semana enviar avances para las respectivas correcciones.

- *PRESENTACIONES.*

  Realizaré exposiciones a los asesores de la presentación del perfil y la defensa final.

1.6. Marco Teórico

1.6.1. Bosquejo Histórico

La historia de la genética de poblaciones está ligada al desarrollo y formalización de la teoría evolutiva, resultando en la Teoría Sintética de la Evolución en las décadas de 1930 y 1940. Comienza a desarrollarse a partir de comienzos del siglo XX, cuando independientemente Godfrey Harold Hardy y Wilhelm Weinberg utilizan principios mendelianos de la segregación y de la probabilidad para explicar las relaciones entre las frecuencias alélicas y genotípicas en una población.

Entre 1918 y 1932 la larga polémica entre biométricos y mendelianos se zanja finalmente cuando Ronald Aymer Fisher (1890-1962), John Burdon Sanderson Haldane (1892-1964) en Inglaterra y Sewall Green Wright (1889-1988) en Estados Unidos llevaron a cabo la síntesis del darwinismo, el mendelismo y la biometría y fundaron la teoría de la genética de poblaciones. Estos autores lograron la armonización de las teorías desarrolladas por las dos escuelas que existían a comienzos del siglo XX. Por un lado, la escuela mendeliana, liderada por William Bateson y sus discípulos, consideraba que los caracteres de herencia simple y naturaleza cualitativa eran los únicos con relevancia evolutiva. Para esta escuela, las pequeñas diferencias entre individuos no eran de naturaleza genética sino ambiental. La otra escuela, la escuela biométrica, liderada por Francis Galton, Walter Weldon y Karl Pearson consideraba que las diferencias hereditarias eran cuantitativas y continuas, negando la necesidad de la existencia de los genes como entidades discretas.

Las herramientas y técnicas matemáticas de la genética de poblaciones se refinaron

progresivamente a lo largo del siglo pasado hasta el presente, y continúan siendo parte importante de todo estudio genético o genómico. En las poblaciones naturales, los métodos moleculares han facilitado enormemente el análisis de los patrones de variación genética y proporcionan una oportunidad sin precedentes para el estudio empírico de la evolución y los procesos demográficos que la forman. Además, el uso de estas herramientas moleculares se ha convertido en un enfoque fundamental para la conservación de las especies.

1.6.1.1. Principio de Hardy-Weinberg

El Principio de Hardy-Weinberg es la base de la genética matemática de poblaciones, fue descrito en 1908 en forma independiente por el matemático británico Godfrey Harold Hardy y el médico alemán Wilhelm Weinberg. El principio de Hardy-Weinberg, utilizando una aplicación de la distribución binomial, describe y predice las frecuencias genotípicas y alélicas de una población que no evoluciona. Relaciona la frecuencia alélica con la frecuencia de genotipos y explica matemáticamente porqué en la población las mutaciones dominantes no reemplazan a las recesivas. El modelo asume los siguientes supuestos:

1. No hay mutación
2. El apareamiento es al azar
3. la población es de tamaño infinito, no hay deriva genética
4. No hay selección: todos sobreviven y dejan igual número de descendientes
5. No hay flujo génico : la población está aislada de otras poblaciones de la misma especie

Hardy y Weinberg demostraron que, si se cumplen los supuestos, las frecuencias génicas de la población no dependen del carácter dominante o recesivo de los genes mantenién-

dose constantes generación tras generación. A pesar que la reproducción sexual supone mezcla de genes, la constante reorganización de éstos no cambia la frecuencia génica en las sucesivas generaciones manteniéndose la variabilidad presente en la población. Es decir, la herencia mendeliana por si misma, no engendra cambio evolutivo, no es un mecanismo que altera las frecuencias génicas en las poblaciones.

La relación general entre la frecuencia alélica y genotípica puede describirse en términos algebraicos: si p es la frecuencia del alelo A_1 y q es la frecuencia del alelo A_2, se cumple que $p + q = 1$ si existen solo 2 alelos. Las frecuencias genotípicas de equilibrio vienen dadas por $p^2A_1A_1$, $2pqA_1A_2$, $q^2A_2A_2$, estas frecuencias resultan del desarrollo del binomio $(p + q)^2$. En el caso de alelos múltiples la condición de equilibrio esta dada por una multinomial.

Al presentar las características de una población no influenciada por fuerzas evolutivas, el principio de Hardy-Weinberg permite comparar la estructura genética de una población real con la estructura genética que se espera de una población en equilibrio genético. Si las frecuencias genotípicas observadas se desvían en forma significativa de las calculadas de acuerdo a la asunción de equilibrio del modelo, alguna fuerza evolutiva está actuando en la población de estudio. El modelo de Hardy-Weinberg constituye, entonces, la hipótesis nula de la genética de poblaciones.

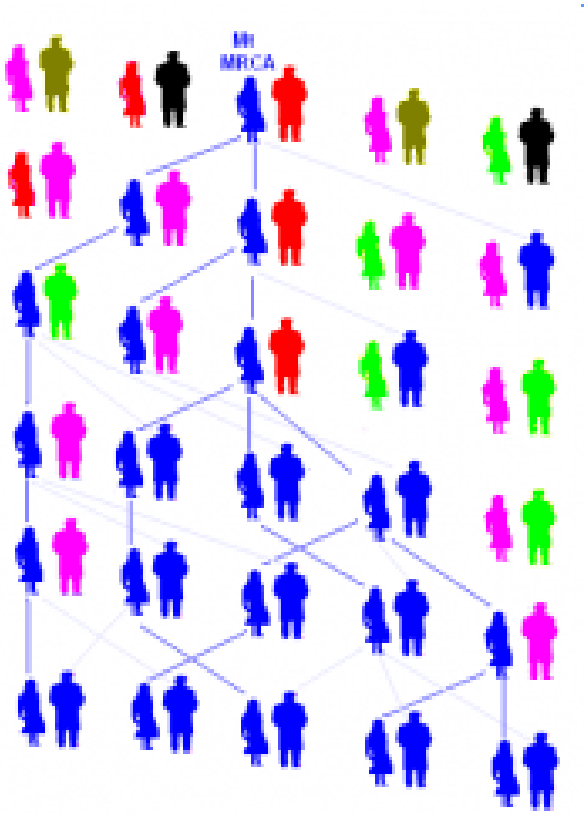
1.6.1.2. Teoría de la coalescencia

La teoría de la coalescencia o simplemente coalescente, definido por John Kingman, un matemático, en 1982 es considerado el avance más importante de la genética de poblaciones de las últimas tres décadas y sustenta gran parte de la investigación actual en genética de poblaciones. Se basa en el hecho que, asumiendo que hubo un origen único

para todos los seres vivos, todos los genes iguales de una población derivan por copia de un antepasado común..

La teoría de la coalescencia se generó como una alternativa al análisis prospectivo genético-poblacional tradicional que permite hacer predicciones sobre la evolución futura de una población. El coalescente es un modelo matemático retrospectivo basado en la genealogía de genes. A partir de modelos matemáticos describe como se van uniendo los linajes hacia atrás en el tiempo hasta alcanzar el antepasado común (coalescencia) y establece la relación entre tiempo de coalescencia y tamaño poblacional, edad del ancestro en común más reciente y otros parámetros como mutación, migración, etc.

La teoría de la coalescencia aplicada al estudio de la variedad nucleotídica ha permitido reconstruir la historia evolutiva y la estructura poblacional de muchas especies. Entre ellas, la historia reciente de la especie humana. El grupo de Wilson estudió el polimorfismo de restricción del ADN mitocondrial (es transmitido solo por las hembras) de 147 individuos humanos de origen geográfico y étnico muy diverso. Sus resultados muestran que todos los tipos de ADN mitocondrial actuales coalescen en una copia perteneciente a una mujer que vivió hace unos 200,000 de años en África, conocida como la Eva mitocondrial. La población en tiempos de Eva se estima en 7,000 a 14,000 individuos. La interpretación correcta de estos resultados es que de todos los linajes de mujeres sólo uno ha conseguido transmitir su ADN mitocondrial a nuestra generación, habiéndose extinguido todos los demás. En conclusión, la Eva mitocondrial no estaba sola y nuestra herencia nuclear proviene de muchas otras mujeres y hombres que coexistieron o no con ella.


MRCA

Década	Desarrollo
1850	Darwin y Wallace formulan la teoría de evolución por selección natural.
1860	Mendel establece los fundamentos de la genética.
1900	Las leyes de Mendel son redescubiertas y finalmente valoradas. Hardy y Weinberg establecen los fundamentos de la genética de poblaciones teórica.
1910	Fisher y Wright desarrollan la genética de poblaciones, expresando la teoría darwinista de la selección en modelos poblacionales. Se inicia la formulación de la moderna teoría sintética de la evolución.
1920	Se consolida la noción de que los genes están localizados en los cromosomas, pese a que se desconoce su naturaleza.
1920-1930	Müller establece que la radiación ultravioleta produce mutaciones. Se establece que el ADN es el material hereditario.
1940	Modelo de doble hélice del ADN, por Crick y Watson.
1950	Determinación del código genético vinculando la información del ADN con la secuencia de las proteínas. Se consolida la idea del reloj molecular. La electroforesis de proteínas provee la primera vía sistemática de examinar la variación genética en las poblaciones naturales. Kimura formula su teoría neutralista de la evolución molecular.
1960	Primeras secuencias de ADN.
1970	Estudios de ADN mitocondrial sugieren el origen africano de la especie humana moderna. Mullis inventa la reacción de PCR (polymerase chain reaction).
1980	Kingman define el coalescente como modelo de análisis genético poblacional.

1.6.2. Preliminares

En esta sección se presenta el marco teórico bajo el que se desarrolla este trabajo. En su mayoría, los resultados aquí presentados se dan sin demostración pues corresponden a teoremas clásicos en el área de matemática y en el contexto en que aparecen es más general de lo necesario. Aún así, emergen como resultado de la revisión bibliográfica y como complemento a los conocimientos adquiridos durante la carrera.

- Procesos Estocásticos

Un proceso estocástico es una colección de variables aleatorias $\{X_t : t \in T\}$ parametrizada por un conjunto T, llamado espacio parametral, en donde las variables toman valores en un conjunto S llamado espacio de estados.

Ejemplo 1.6.3. *A un deudor de una tarjeta de crédito le ofrecen distintos métodos para finiquitar su deuda. Sea*

A: Un abono mensual sin altos interese

F: Finiquitar la deuda completa

I: Pagar un mínimo con altos intereses

B: No pagar nada y quedarse en el buró de crédito

X_t: Decisión a tomar en el mes t.

S:{ A,F,I,B} Decisión a tomar Discreto

T: { 0,1,2,...,12} Mes Discreto

Por lo tanto es una serie estocástica con espacio de estado discreto.

Clasificación de Procesos Estocásticos

Por ultimo los procesos estocástico se pueden clasificar de la siguiente forma:

Procesos Estocásticos Estacionarios: Un proceso estocástico estacionario es aquel

cuya distribución de probabilidad varía de forma más o menos constante a lo largo de cierto periodo de tiempo. Con otras palabras, una serie de números puede parecer (y ser) caótica pero tomar valores dentro de un rango limitado.

Procesos Estocásticos no Estacionarios: Un proceso estocástico no estacionario es aquel cuya distribución de probabilidad varía de forma no constante. Si una serie de números se comporta de forma totalmente caótica, podríamos decir que es aleatorio no estacionario.

S/T	t Discreto	t Continuo
X Discreto	Proceso de estado discreto y tiempo discreto (Cadena) (Unidades producidas mensualmente de un producto)	Proceso de estado discreto y tiempo continuo (Proceso de Saltos Puros) (Unidades producidas hasta el instante t)
X Continuo	Proceso de estado continuo y tiempo discreto (Toneladas de producción diaria de un producto)	Proceso de estado continuo y tiempo continuo (Proceso Continuo) (Velocidad de un vehículo en el instante t)

- Procesos de Markov

Las cadenas de Markov y los procesos de Markov son un tipo especial de procesos estocásticos que poseen la siguiente propiedad:

Propiedad de Markov: Conocido el estado del proceso en un momento dado, su comportamiento futuro no depende del pasado. Dicho de otro modo, "dado el presente, el futuro es independiente del pasado"

Cadena de Markov: es un proceso estocástico a tiempo discreto $X_n : n = 0, 1, ...$, con espacio de estados discreto, y que satisface la propiedad de Markov, esto es, para cualquier entero $n \geq 0$, y para cualesquiera estados $x_0, ..., x_{n+1}$, se cumple

$$P(x_{n+1}|x_0, ..., x_n) = P(x_{n+1}|x_n)$$

- Martingala

Una martingala a tiempo discreto es, en términos generales, un proceso $\{x_n : n = 0, 1, ...\}$ que cumple la condición

$$E(X_{n+1}|X_0 = x_0, ..., X_n = x_n) = x_n$$

En palabras, esta igualdad significa que el valor promedio del proceso al tiempo futuro $n + 1$ es el valor del proceso en su último momento observado, es decir, x_n. Esto es, se trata de una ley de movimiento aleatorio que es equilibrada o simétrica, pues en promedio el sistema no cambia del último momento observado.

- Ecuaciones Diferenciales Estocásticas

Una ecuación diferencial estocástica es una ecuación de la forma

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t$$

definida para valores de T en un intervalo $[0, T]$, y con condición inicial la variable aleatoria X_0 que se asume F_0-medible e independiente del movimiento Browniano. La incógnita de esta ecuación es el proceso X_t. Los coeficientes $b(t, x)$ y $\sigma(t, x)$ son funciones reales definidas sobre $[0, T] \times \mathbb{R}$, y se conocen como los coeficientes de tendencia (o también deriva) y de difusión respectivamente. La ecuación diferencial Estocástica se interpreta como la ecuación integral

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dB_s$$

en donde la primera es una integral de Riemann mientras que la segunda es una integral estocástica de Itô. El proceso solución puede interpretarse como el estado de un sistema que evoluciona de manera determinista gobernado por la parte no aleatoria de

la ecuación pero perturbado por un ruido aditivo dado por la integral estocástica. A un proceso de esta forma última se le llama proceso de Itô y para que esta ecuación tenga alguna solución se deben imponer condiciones en los coeficientes. De manera análoga al caso determinista, existen teoremas básicos de existencia y unicidad para ecuaciones diferenciales estocásticas que establecen condiciones de regularidad para los coeficientes b y σ, bajo las cuales la ecuación Estocástica tiene solución única.

Teorema (Formula de Itô).

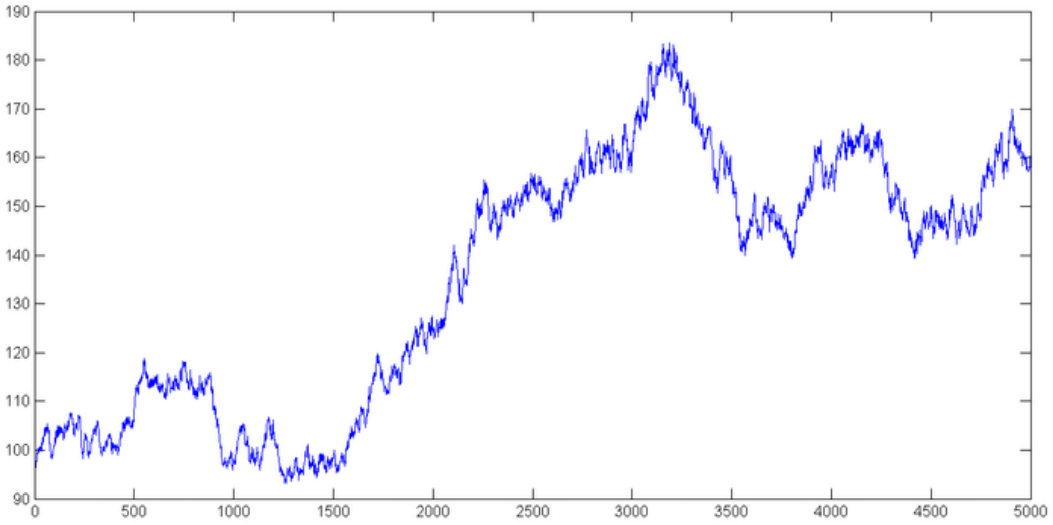
Si X_t es un proceso de Itô dado por la Ecuación Estocástica y $f(t,x)$ es una función de clase C^1 en t y de clase C^2 en x, entonces el proceso $Y_t = f(t, X_t)$ es también un proceso de Itô y satisface la ecuación

$$dY_t = f_t(t, X_t)dt + f_x(t, X_t)dX_t + \frac{1}{2}f_{xx}(t, X_t)d\langle X_t, X_t\rangle$$

- Movimiento Browniano

Un movimiento Browniano unidimensional de parámetro σ^2 es un proceso estocástico $\{B_t : t \geq 0\}$ con valores en $\mathbb{R}$ que cumplen las siguientes propiedades:

1. $B_0 = 0$
2. Las trayectorias son continuas
3. El proceso tiene incrementos independientes
4. Para cualesquiera tiempos $0 \leq s < t$, la variable incremento $B_t - B_s$ tiene distribución $N(0, \sigma^2(t-s))$



- Ley Fuerte de los Grandes Números

La idea de probabilidad está íntimamente relacionada a la frecuencia relativa. Algunos resultados sobre esta relación son los llamados "Leyes de los Grandes Números", que establecen el tipo de convergencia, “debil” o “fuerte”, en que las frecuencias relativas de un evento se aproximan a la probabilidad de este.

La Ley Fuerte de los Grandes Números, llamada también Teorema de Borel, se puede enunciar así:

La frecuencia relativa con que ocurre un hecho en pruebas independientes y en las mismas condiciones converge a la probabilidad del hecho observado con probabilidad 1.

Si la convergencia que se da es en forma “casi segura” la ley a la que de lugar se conocerá como ley fuerte de los grandes números.

Una sucesión de variables aleatorias, X_n , converge con probabilidad 1, o de forma casi segura, a una variable aleatoria X (que puede degenerar en una constante K) cuando se cumple que:

$$P(\lim_{x\to\infty} X_n = X) = 1$$

De esta forma interpretamos que $X_n \xrightarrow{c.s} X$, cuando la probabilidad de que en el límite la sucesión de variables aleatorias y aquella a la que converge sean iguales es uno.

- Proceso de Feller

La definición de los procesos de Feller implica poner restricciones de continuidad en la función de transición , para lo cual es necesario restringir la atención a los procesos que se encuentran en un espacio topológico $(E, \mathcal{T}_E)$. Se supondrá que E es localmente compacto , Hausdor , y tiene una base contable (lccb, para abreviar). Dichos espacios siempre poseen una colección contable de funciones continuas no desaparecidas $f\colon E \to \mathbb{R}$ que separan los puntos de E. Los espacios de Lccb incluyen muchos de los espacios topológicos que podemos considerar, como $\mathbb{R}^n$, múltiples topológicos y, de hecho, cualquier subconjunto abierto o cerrado de otro espacio lccb.
Dado un espacio topológico E , $C_0(E)$ denota las funciones continuas de valor real que desaparecen en el infinito . Es decir, $f\colon E \to \mathbb{R}$ está en $C_0(E)$ si es continuo y, para cualquiera $\epsilon > 0$, el conjunto $\{x\colon |f(x)| \geq \epsilon\}$ es compacto. De manera equivalente, su extensión a la compactación $E^* = E \cup \{\infty\}$ de un punto de E dada por $f(\infty) = 0$ es continua. El conjunto $C_0(E)$ es un espacio de Banach bajo la norma uniforme,

$$\|f\| \equiv \sup_{x \in E} |f(x)|.$$

Ahora podemos establecer la definición general de las funciones y procesos de transición de Feller. Un espacio topológico $(E, \mathcal{T}_E)$ es también considerado como un espacio medible equipándolo con su Borel sigma álgebra $\mathcal{B}(E) = \sigma(\mathcal{T})$, así que tiene sentido hablar de las probabilidades de transición y funciones de E.

Definición 1.6.4. *Sea E ser un espacio lccb. Entonces, una función de transición $\{P_t\}_{t \geq 0}$ es Feller si, para todos $f \in C_0(E)$,*

1. $P_t f \in C_0(E)$.

2. $t \mapsto P_t f$ *es continuo con respecto a la topología de la norma* $C_0(E)$.

3. $P_0 f = f$.

Un proceso X de Markov cuya función de transición es Feller es un proceso de Feller .

Nota: Los procesos de Feller, como se definen aquí, a veces se denominan procesos de Feller-Dynkin (y de manera similar para las funciones de transición de Feller). El término proceso de Feller a veces se usa para referirse a la clase más general de procesos obtenidos al reemplazar $C_0(E)$ por el espacio $C_b(E)$ de funciones limitadas continuas en la definición anterior.

Ahora mostramos que, en la definición de los procesos de Feller, no es necesario imponer la condición $t \mapsto P_t f$ continua bajo la topología de la norma, para $f \in C_0(E)$. De hecho, la condición mucho más débil que, $P_t f(x) \to f(x)$ como $t \to 0$ es suficiente.

Teorema 1.6.5. *Una función de transición* $\{P_t\}_{t\geq 0}$ *en un espacio lccb* E *es Feller si y solo si, para todos* $f \in C_0(E)$*,*

1. $P_t f \in C_0(E)$ *para todos* $t \geq 0$.

2. $P_t f(x) \to f(x)$ *como* $t \to 0$*, para cada uno* $x \in E$.

Un proceso de Markov se define por una colección de probabilidades de transición $P_{s,t}$, una para cada $s \leq t$, que describe cómo pasa de su estado en el tiempo s a una distribución en el tiempo t . Aquí solo considero el caso homogéneo, lo que significa que $P_{s,t}$ depende solo del tamaño $t - s$ del incremento de tiempo y no explícitamente de los tiempos de inicio o finalización s, t , por lo que la notación $P_{s,t}$ puede ser reemplazada por P_{ts}. Esto no es una gran restricción porque, dado un proceso X de Markov no homogéneo , siempre es posible observar su proceso de espacio-tiempo (t, X_t) tomando valores $\mathbb{R}_+ \times E$, que serán Markov homogéneos.

Definición 1.6.6. *Una función de transición homogénea en* $(E, \mathcal{E})$ *es una colección* $\{P_t\}_{t\geq 0}$ *de probabilidades de transición en* $(E, \mathcal{E})$ *tal que* $P_s P_t = P_{s+t}$ *para todos* $s, t \geq 0$.

Un proceso X es Markov con función de transición $P = \{P_t\}$, y con respecto a un espacio de probabilidad filtrado $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ si está adaptado y

$$\mathbb{E}\left[f(X_t) \mid \mathcal{F}_s\right] = P_{ts} f(X_s)$$

(casi seguro), para todos los tiempos $s < t$.

La identidad $P_{s+t} = P_s P_t$ se conoce como la ecuación de Chapman-Kolmogorov. Alternativamente $\{P_t\}$ forma un semi-grupo .

- Generador infinitesimal (procesos estocásticos)

En matemáticas, específicamente en el análisis estocástico, el generador infinitesimal de un proceso estocástico es un operador diferencial parcial que codifica una gran cantidad de información sobre el proceso.

Definición 1.6.7. *Sea $X : [0, \infty) \times \Omega \to \mathbb{R}^n$ definido en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ si una difusión Itô que satisfaga una ecuación diferencial estocástica de la forma:*

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t \mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t$$

dónde B es un movimiento browniano m-dimensional y $b : \mathbb{R}^n \to \mathbb{R}^n$ y $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ son los campos de deriva y difusión respectivamente. Por un punto $x \in \mathbb{R}^n$, sea $\mathbb{P}^x$ denotar la ley de X con valor inicial dado $X_0 = x$, y sea $\mathbb{E}^x$ la expectativa con respecto a $\mathbb{P}^x$

El generador infinitesimal de X es el operador $\mathcal{A}$, que se define para actuar sobre funciones adecuadas $f : \mathbb{R}^n \to \mathbb{R}$ por:

$$\mathcal{A}f(x) = \lim_{t\to 0} \frac{\mathbb{E}^x[f(X_t)] - f(x)}{t}$$

El conjunto de todas las funciones f para el cual existe este límite en un punto x se denota $D_A(x)$, mientras D_A denota el conjunto de todos f para lo cual existe el límite para todos $x \in \mathbb{R}^n$.

- Convergencia

Consideraremos $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad fijo. Las sucesiones de variables aleatorias estarán definidas en este espacio.

Definición 1.6.8. *Convergencia Casi Segura. Una sucesión de variables aleatorias* $(X_n)_{n\geq 1}$ *se dice que converge casi seguramente si su conjunto de convergencia tiene probabilidad* 1.

La convergencia casi segura la denotaremos por $X_n \xrightarrow{c.s} X$.
Un tipo de convergencia más débil que la convergencia casi segura es la llamada convergencia en probabilidad.

Definición 1.6.9. *Convergencia en Probabilidad. Una sucesión* $(X_n)_{n\geq 1}$ *de variables aleatorias se dice que converge en probabilidad a la variable aleatoria* X *si para cada* $\epsilon > 0$ *se satisface:*

$$\lim_{n\to\infty} P[|X_n - X| > \epsilon] = 0$$

La convergencia en probabilidad será denotada por $X_n \xrightarrow{P} X$.
En las definiciones de convergencia casi segura y en probabilidad, se consideró un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ fijo en donde estaban definidas todas las variables aleatorias. La convergencia en distribución que se definirá a continuación es un concepto que se refiere no a una propiedad de convergencia de las variables aleatorias sino de las funciones de distribución. Así, las variables aleatorias en consideración pueden estar definidas en distintos espacios de probabilidad.

Definición 1.6.10. *Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias y $(F_n)_{n\geq 1}$ la sucesión correspondiente de funciones de distribución. Diremos que X_n converge en distribución a (la variable aleatoria) X con función de distribución F, si*

$$\lim_{n\to\infty} F_n(x) = F(x),$$

para todo $x \in \mathbb{R}$, punto de continuidad de F. La convergencia en distribución la denotaremos $X_n \xrightarrow{D} X$.

Capítulo 2

Dualidad en la Gráfica Wright-Fisher

2.1. La Gráfica Aleatoria de Wright-Fisher

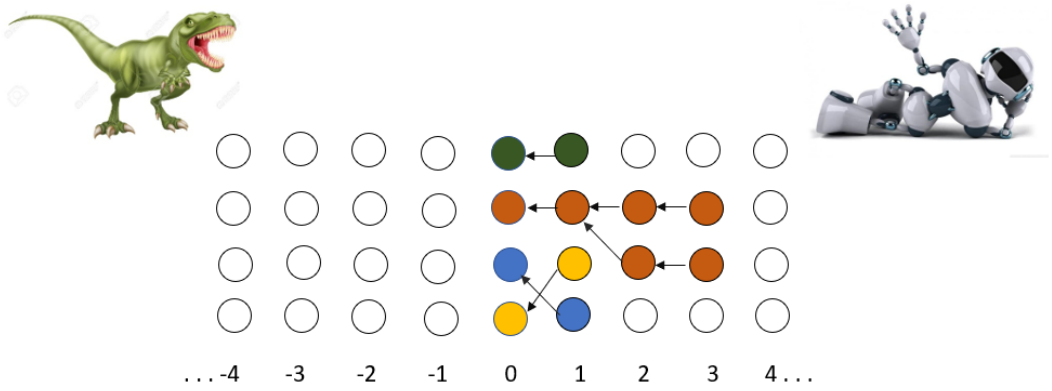
Definición 2.1.1. *(El modelo de Wright-Fisher).*

Sea $N \in \mathbb{N}$, N es el tamaño de la población. Consideremos el conjunto de vértices $V = \mathbb{Z} \times [N]$, donde $[N] = \{1, 2, 3, ..., N\}$. El vértice $(g, i) \in V$ representa al i-ésimo individuo de la generación g. A cada vértice $v \in V$ le asociamos una variable aleatoria U_v distribuida uniforme en [N]. La familia de variables aleatorias $\{U_v\}_{v \in \mathbb{N}}$ es una familia independiente. La madre del individuo (g, i) es el individuo $(g-1, U_{(g,i)})$. Hay una arista entre cada individuo y su madre, es decir, el conjunto de aristas es

$$A := \{(v, v') : v = (g, i) \in V, v' = (g - 1, U_v)\}$$

La gráfica de Wright-Fisher es

$$G := (V, A)$$



La gráfica de Wright-Fisher modela las relaciones de ancestria de una población, bajo las hipótesis de que cada individuo elige a su madre uniformemente al azar y numero de individuos por generación iguales (el tamaño de la población es constante).

2.2. Proceso de Frecuencia y Ancestría

La gráfica se puede estudiar en dirección del futuro, analizando como cambia el perfil genético de una población, o en dirección del pasado, estudiando las relaciones de ancestría de una muestra de la población en cierta generación. El proceso de frecuencia de dos tipos es la manera más simple de estudiar la gráfica en dirección del futuro.

Definición 2.2.1. *(El proceso de frecuencia de dos tipos).*
Asumamos que existen dos tipos de individuos $+$ *y* $-$. $t(g,i) \in \{+,-\}$*es el tipo del individuo (g, i). En el espacio de probabilidad de la gráfica de Wright-Fisher (V, A), fijamos una generación inicial* $a \in \mathbb{Z}$ *y asignamos tipos a todos los individuos en las generaciones posteriores a "*a*" de la siguiente forma:*

1. *Asignamos de manera arbitraria tipos a cada individuo en la generación* $a \in \mathbb{Z}$. *Denotamos al número de individuos tipo* $-$ *en la generación* a *por* xN. *Esto*

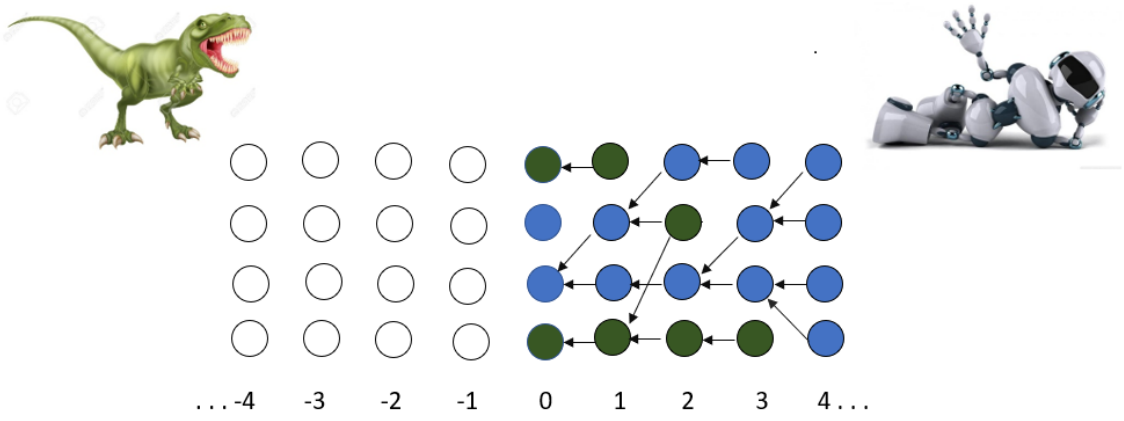
implica que $x \in [0,1]$.

2. *Cada individuo copia el tipo de su madre.*

El Proceso de Frecuencia $(X(a,g))_{g\geq 0}$ y valor inicial $X(a,0) = x$ se define como

$$X(a,g) = \frac{\sum_{i=1}^{N} 1_{\{t(i,g)=-\}}}{N}$$

Resulta que el proceso de frecuencia es una cadena de Markov a tiempo discreto.



Proposición 2.2.2. *Para toda* $a \in \mathbb{Z}$, $g \in \mathbb{N}$ *y* $x \in [N]/N$ *fijas, la distribución de* $X(a,g+1)$*, dado* $X(a,g) = x$*, esta caracterizada por:*

$$NX(a,g+1)|(X(a,g)) = x \stackrel{d}{=} B(N,x) \tag{2.1}$$

donde $B(N,x)$ *es una variable aleatoria Binomial de parámetros* N *y* x*.*

Demostración. El individuo $(i,g+1)$ para cualquier $i \in [N]$, es de tipo $-$ si y sólo si elige a un individuo tipo $-$ como su padre. Recordemos que cada individuo elige a su padre uniformemente al azar, entonces si la frecuencia de individuos tipo $-$ en la

generación anterior es x,

$$\mathbb{P}(t(i,g+1) = -|X(a,g) = x) = x$$

Como sólo hay dos tipos, concluimos que

$$1_{\{t(i,g)|(X(a,g)=x)=-\}} \stackrel{d}{=} b_i$$

donde b_i es una variable aleatoria Bernoulli con parámetro x. Además $\{b_i\}_{i\in[N]}$ es una familia independiente de variables aleatorias, pues se construye utilizando una familia aleatoria de variables aleatorias (las variables aleatorias uniformes en $[N]$ que usa cada individuo para elegir a su padre en la generación anterior). La prueba se concluye observando que $N(X(a,g+1))|(X(a,g)) = x$ es el número de individuos tipo $-$ en la generación $g+1$ y por lo tanto

$$N(X(a,g+1))|(X(a,g)) = x = \sum_{i=1}^{N} 1_{\{t(i,g)=-\}|(X(a,g))=x} \stackrel{d}{=} \sum_{i=1}^{N} b_i \stackrel{d}{=} B(N,x) \tag{2.2}$$

■

Una consecuencia inmediata de la proposición anterior es que

$$\mathbb{E}[X(a,g+1)|(X(a,g)) = x] = x \tag{2.3}$$

$$Var[X(a,g+1)|(X(a,g)) = x] = \frac{1}{N}x(1-x) \tag{2.4}$$

El modelo de Wright-Fisher es un modelo neutral. Esto significa que ser de un tipo o de otro $(+,-)$ no hace mas probable que un individuo se reproduzca. En el lenguaje matemático esto significa simplemente que $X(a,g) = x$ es una Martingala, es decir para todo $s, g \in \mathbb{N}$

$$\mathbb{E}[X(a, g+s)|X(a,g)] \overset{a.s}{=} X(a,g) \tag{2.5}$$

Para demostrar que $X(a,g)$ en efecto es una Martingala, es posible utilizar la construcción de la gráfica aleatoria (no es el método mas rápido, pero nos permite entender mejor a la gráfica de Wright-Fisher y a los procesos que en ella habitan). La observación es que, condicionalmente es $X(a,g)$,

$$1_{\{t(i,g+s)=-\}|(X(a,g))=x} \overset{d}{=} c_i$$

donde de nuevo c_i es una variable aleatoria Bernoulli con parámetro $X(a,g)$. Para concluir esto noten que el tipo del individuo $(i, g+s)$ es el mismo que el de su padre y este a su vez es el mismo que el de su padre. Entonces, el tipo de el individuo $(i, g+s)$ es el mismo que el del padre del padre del padre... de su padre que vivió en la generación g, y quien, por construcción, es un individuo elegido uniformemente al azar en la generación g, al cual podemos llamar el ancestro del individuo $(i, g+s)$ en la generación g. Entonces, la probabilidad de que el ancestro del individuo $(i, g+s)$ en la generación g sea tipo $-$ es simplemente $X(a,g)$.

$$\mathbb{E}[X(a,g+s)|X(a,g)] = \frac{1}{N}\mathbb{E}[\sum_{i=1}^{N} 1_{\{t(i,g+s)=-\}}|X(a,g)] = \frac{1}{N}\mathbb{E}[\sum_{i=1}^{N} c_i|X(a,g)] =$$
$$= \mathbb{E}[c_i|X(a,g)] = X(a,g)$$

Cabe mencionar, $\{c_i\}_{i\in[N]}$ no es una familia independiente de variables aleatorias, hay bastante correlación en el tipo de distintos individuos. Esto se debe a que el ancestro del individuo $(i, g+s)$ en la generación g y el ancestro del individuo $(j, g+s)$ en la generación g pueden ser el mismo. Este argumento nos indica que conocer la genealogía de una población nos permite decir cosas sobre cómo evoluciona el proceso de frecuen-

cia $X(a, g)$. Antes de enfocarnos en el proceso de ancestros, hablemos del equilibrio de Hardy-Weinberg (1908), el cual es un principio clásico en genética de poblaciones que establece que la composición genética de una población permanece en equilibrio mientras no actué la selección natural.

Este principio se puede formalizar en la siguiente proposición, trivial desde el punto de vista matemático, pero muy importante desde el punto de vista biológico.

Proposición 2.2.3. *Para todo* $a \in \mathbb{Z}$, $g \in \mathbb{N}$ *y* $x \in [N]/N$ *fijas,*

$$\lim_{N\to\infty} X(a, g+1)|(X(a, g) = x) \overset{a.s}{=} x \tag{2.6}$$

Demostración. La prueba es una aplicación inmediata de la Ley fuerte de los grandes números. También podemos probarla usando la desigualdad de Tschevichev, pues para todo N, la esperanza de $\mathbb{E}[X(a, g+1)|(X(a, g)) = x] = x$ y $\lim_{N\to\infty} Var[X(a, g+1)|(X(a, g)) = x] = 0$

■

El principio de Hardy-Weinberg implica que cualquier desviación de la esperanza es una señal de que alguno de los dos tipos tiene ventaja selectiva, alguno de los dos tipos esta transformándose en el otro (por ejemplo por que esta sufriendo mutaciones) o hay alguna otra fuerza de la evolución que hace que el proceso de frecuencia no sea neutral. Resulta que, en la escala correcta, la suerte puede determinar que un tipo se fije, incluso si este no tiene ventaja selectiva. A la suerte, como una fuerza de la evolución, se le conoce como deriva génica. Profundizaremos en este tema más adelante.

Hemos definido un proceso para estudiar como evoluciona el perfil genético de una población, o para ser más precisos, como cambia la frecuencia de individuos tipo $-$ en la gráfica de Wright-Fisher. Ahora estudiaremos las relaciones de ansestría de individuos en la gráfica.

Definición 2.2.4. *La relación $\sim_g$ entre individuos $v, v' \in V_g := \{(g', i) \in V : g' > g\}$ se define como: $v \sim_g v'$ si y solo si el ancestro del individuo v en la generación g y el ancestro del individuo v' en la generación g son el mismo individuo.*

Noten que la relación $\sim_g$ es una relación de equivalencia. Esto se debe a que cada individuo $v \in V_g$ tiene un único ancestro en la generación g, luego entonces cada individuo esta relacionado con el mismo, si $v \sim_g v'$, entonces el ancestro de v' en la generación g es el mismo que el de v, es decir $v' \sim_g v$. Ahora, si $v \sim_g v'$ y $v' \sim_g v''$, entonces los tres individuos tienen al mismo ancestro y por lo tanto $v \sim_g v''$.

Una relación de equivalencia para los números en $[N]$ es equivalente a una partición de los números en $[N]$. Por ejemplo, la partición $\{\{1, 2, 4\}\{3\}\{5\}\}$ induce la relación de equivalencia $1 \sim 2 \sim 4$ y los demás miembros de $[N]$ no son equivalentes a nada. Además, si decimos que dos números están en el mismo bloque, si están son equivalente según la equivalencia $1 \sim 2 \sim 4$ y los demás miembros de $[N]$ no son equivalentes, entonces obtendremos la partición $\{\{1, 2, 4\}\{3\}\{5\}\}$. Por esta razón diremos algunas veces que i y j están en el mismo bloque de una partición, y otras diremos que son equivalentes, es lo mismo.

Los procesos de ancestría se consideran como procesos con valores en las particiones de $[N]$, principalmente por la razón de que dicho espacio admite una métrica que se puede extender al caso cuando N tiende a infinito. Dicha métrica se construye definiendo para toda partición π de $[N]$, la proyección de π' en $[M]$, para $M < N$, como $\pi'|_M$ tal que para todo $i, j \in [M]$, i y j están en el mismo bloque de $\pi'|_M$ si y solo si están en el mismo bloque de π. La distancia entre dos particiones de $[N]$, π y π' se define como:

$$d(\pi, \pi') = -\frac{1}{N} + \underset{\{M \leq N : \pi|_M = \pi'|_M\}}{min} \{\frac{1}{M}\}$$

Definición 2.2.5. *Sea π una partición aleatoria $[N]$, decimos que π es intercambiable si su distribución es invariante a permutaciones de $[N]$.*

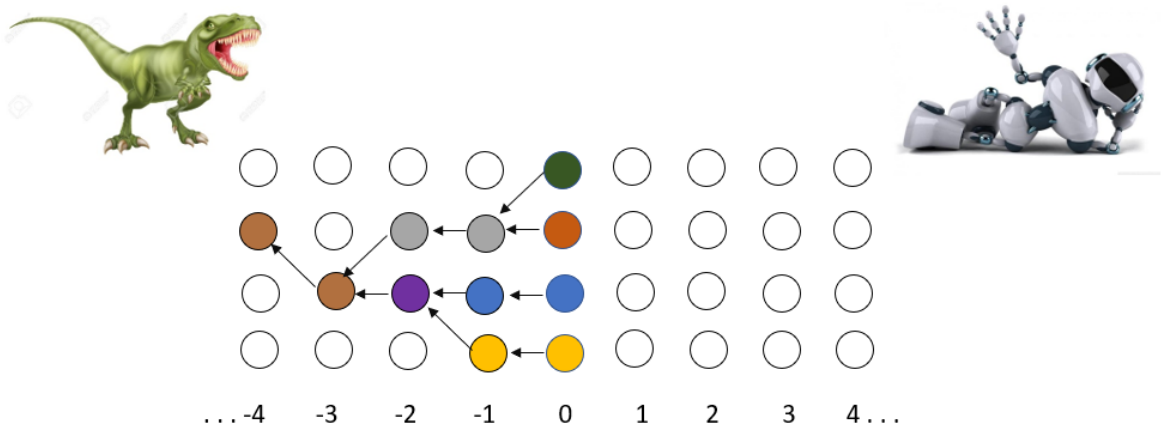
En esta investigación nos vamos a concentrar en el proceso de conteo de bloques, el cual se construye proyectando el espacio de las particiones al espacio de los números naturales, asignando a cada partición el numero de bloques que la conforma. Por ejemplo $|\{\{1,2,4\}\{3\}\{5\}\}| = 3$ y $|\{\{1\}\{2,4\}\{3\}\{5\}\}| = 4$.

Definición 2.2.6. *En el espacio de probabilidad de la gráfica de Wright-Fisher* (V, A)*, fijamos una generación inicial* $b \in \mathbb{Z}$ *y muestreamos* $n \leq N$ *individuos en dicha generación. Definimos* $\pi(b, g)$ *a la partición inducida por* $\sim_g$ *en* $[n]$*, donde* n *es el número de individuos que muestreamos en la generación* g*, y definimos al proceso del número de ancestros* $A(b, g)$ *como*

$$A(b, g) = |\pi(b, g)|$$

Ilustración.

Para $g = 1$ se esta agrupando bloques de hermanos y para $g = 2$ se estaría agrupando bloques de primos.



Claramente el estado $\{1\}$ es absorbente para el proceso $A(b, g)$. El resto de sus probabilidades de transición también son fáciles de escribir, siempre y cuando nos interesen solo los términos de orden superior.

Proposición 2.2.7. *Para todo* $b \in \mathbb{Z}$, $g \in \mathbb{N}$ *y* $n \in [N]$ *fijas, tales que* $n << N$,

$$\mathbb{P}(A(b, g+1) = n-1 | A(b,g) = n) = \binom{n}{2}\frac{1}{N} + O(1/N) \tag{2.7}$$

$$y$$

$$\mathbb{P}(A(b, g+1) = n | A(b,g) = n) = 1 - \binom{n}{2}\frac{1}{N} + O(1/N) \tag{2.8}$$

Demostración. La primera observación es que la probabilidad de que el i-ésimo bloque y el j-ésimo bloque coalescan y que ningún otro bloque coalesca, cuando hay n bloques es:

$$\Rightarrow \frac{1}{N}\prod_{i=1}^{n-2}\frac{N-i}{N} = \frac{1}{N} + O(\frac{1}{N})$$

Resolviendo el lado izquierdo

$$\Rightarrow \frac{1}{N}(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-n+2}{N})$$

$$\Rightarrow \frac{1}{N}(1-\frac{1}{N})(1-\frac{2}{N})...(1-\frac{n-2}{N})$$

$$\Rightarrow \frac{1}{N}(1-\frac{\binom{n-1}{2}}{N}) + O(\frac{1}{N})$$

como $N \to \infty$ y n es constante

$$\Rightarrow \frac{1}{N} + O(\frac{1}{N})$$

Como se pueden formar $\binom{n}{2}$ pares de los bloques de n bloques, concluimos que

$$\mathbb{P}(A(b, g+1) = n-1 | A(b,g) = n) = \binom{n}{2}\frac{1}{N} + O(\frac{1}{N}) \tag{2.9}$$

Ahora notemos que para pasar de n bloques a menos de $n-1$ bloques en una generación es necesario que al menos una de las siguientes dos cosas sucedan: dos parejas de bloques coalescan o tres bloques coalescan. Noten que cada uno de estos eventos tiene

probabilidad $1/N^2$. Por lo tanto

$$\mathbb{P}(A(b,g+1) < n-1 | A(b,g) = n) < 2/N^2 = O(\frac{1}{N}) \tag{2.10}$$

Juntando las ecuaciones (2.9) y (2.10) obtenemos que

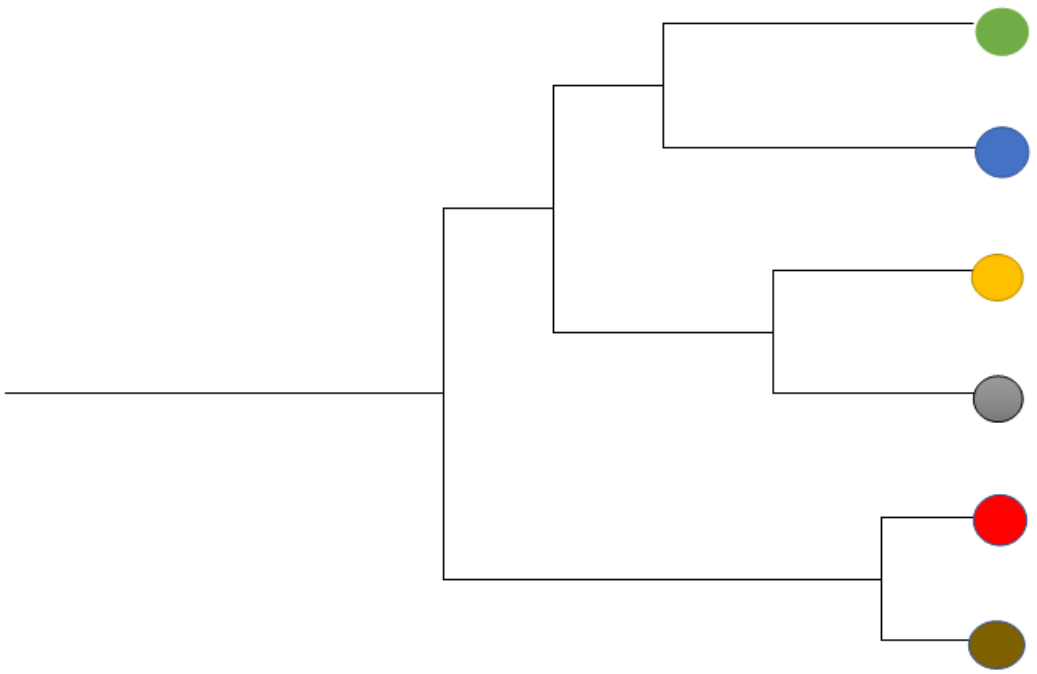
$$\mathbb{P}(A(b,g+1) = n | A(b,g) = n) = 1 - \binom{n}{2}\frac{1}{N} + O(\frac{1}{N})$$

■

Una variable aleatoria, funcional de la gráfica de Wright-Fisher así como del proceso de ancestría, es el tiempo hasta el ancestro común.

Definición 2.2.8. *El tiempo hasta el ancestro común de una muestra de n individuos en la generación b es*

$$T_{MRCA}[n,b] = inf\{g \in \mathbb{N}_0 : A(b,g) = 1 | A(b,0) = n\}$$



Las siguientes dos propiedades son bastante obvias, pero nos serán útiles para calcular límites de escala.

Corolario 2.2.9. *Para todo* $n \in [N], b \in \mathbb{Z}$

1. $T_{MRCA}[2,b] \stackrel{d}{=} G(1/N)$ *es una variable aleatoria geométrica con parámetro 1/N*
2. $\mathbb{E}[T_{MRCA}[n,b]] < \infty$

Demostración. Notemos que $T_{MRCA}[n,b]$ tiene la propiedad de perdida de memoria, es decir $\mathbb{P}(T_{MRCA}[2,b] = i) = \mathbb{P}(T_{MRCA}[2,b] = i+k | T_{MRCA}[2,b] > k)$, además $\mathbb{P}(_{TMRCA}[2,b] = 1) = 1/N$ entonces la distribución de $T_{MRCA}[2,b]$ tiene que ser Geométrica de parámetro $1/N$ (recuerden que la única distribución con valores en $\mathbb{N}$ y la propiedad de perdida de memoria es la Geométrica). Ahora notemos que $\mathbb{E}[T_{MRCA}[n,b]] < n\mathbb{E}[T^i_{MRCA}[2,b]] \leq \mathbb{E}[G^i(1/N)] = nN \leq \infty$, donde T^i_{MRCA} son copias independientes de T_{MRCA}

■

En el proceso de frecuencia, usualmente llamaremos a su tiempo de absorción, tiempo de fijación (el tiempo en que algún tipo de individuo se fija en la población).

Definición 2.2.10. *El tiempo de fijación, cuando la frecuencia inicial de individuos tipo* $-$ *es* $x \in [N]/N$ *en la generación* a*, es*

$$T_{fix}[x,a] = inf\{g \in \mathbb{N}_0 : X(a,g)(1-X(a,g)) = 0 | X(a,0) = x\}$$

2.3. Dualidad

Definición 2.3.1. *Sea* $(X_i)_{i\in I}$ *y* $(A_i)_{i\in I}$ *dos procesos estocásticos con valores en* E_1 *y* E_2 *respectivamente y espacio de índices* I *(Normalmente consideramos* $I = \mathbb{N}_0$ *o* $I = \mathbb{R}^+$*). Sea* $H : E_1 \times E_2 \to \mathbb{R}$*.* $(X_i)_{i\in I}$ *y* $(A_i)_{i\in I}$ *son H-duales para todo* $x \in E_1$*,* $n \in E_2$ *e* $i \in I$

$$\mathbb{E}_x[H(X_i, n)] = \mathbb{E}_n[H(x, A_i)]$$

Ejemplo 2.3.2. *Si* $E_1 = [0, 1]$, $E_2 = \mathbb{N}$ *con la función* $H(x, n) = x^n$

$$H : [0, 1] \times \mathbb{N} \to \mathbb{R}$$

$$\Rightarrow H(x, n) = x^n$$

$$\Rightarrow \mathbb{E}_x[H(X_i, n)] = \mathbb{E}_n[H(x, A_i)]$$

$$\Rightarrow \mathbb{E}_x[X_i^n] = \mathbb{E}_n[x^{A_i}]$$

entonces $(X_i)_{i\in I}$ *y* $(A_i)_{i\in I}$ *cumplen la dualidad de los momentos, para todo* $x \in E_1$, $n \in E_2$ *e* $i \in I$.

Ejemplo 2.3.3. *Sea H(x,n) una función simétrica(con respecto a las entradas) y sea* X_i, A_i *el mismo proceso estocástico, claramente son H-duales.*

Ejemplo 2.3.4. *Siguiendo la notación de la definición de la gráfica de Wright Fisher, el proceso de frecuencia y el proceso de número de ancestros, sea a* $E_1 = [N]/N$, $E_2 = [N]$ *definimos la función de dualidad de muestreo como*

$$S(x, n) = \mathbb{P}(\text{ } n \text{ individuos en la generación } g + 1 \text{ sean tipo } -$$
$$\text{dado que la frecuencia de tipo } - \text{ en la generación } g \text{ es } x)$$

Los individuos eligen a su padre de manera independiente, la probabilidad de que un individuo sea tipo $-$ *en la generación* $g + 1$, *si la frecuencia de individuos tipo* $-$ *en la generación anterior es* x, *es simplemente* x^n. *Luego entonces,*

$\Rightarrow S(x, n) = x^n$

ahora notemos que $\mathbb{E}_x[S(X_g^N, n)] = \mathbb{E}_x[(X_g^N)^n]$ *y* $\mathbb{E}_n[S(x, A_g^N)] = \mathbb{E}_n[x^{A_g^N}]$

$\Rightarrow \mathbb{E}_x[(X_g^N)^n] = \mathbb{E}_n[x^{A_g^N}]$

$\Rightarrow (X_i)_{i\in I}$ y $(A_i)_{i\in I}$ *son S-duales, entonces cumplen la dualidad de muestreo. Es importante notar que la función* $S(x,n)$ *depende de la manera en la que los individuos eligen a sus padres y por lo tanto* S *sera diferente cuando consideremos generalizaciones de la Gráfica de Wright-Fisher.*

También definimos la función de muestreo generalizada.

$$S(x,n,g) = \mathbb{P}(\ n \text{ individuos en la generación } g+1 \text{ sean tipo } -$$
$$\text{dado que la frecuencia de tipo } - \text{ en la generación } 0 \text{ es } x)$$

la cual sera útil para probar dualidad en varios casos.

Es interesante notar que en el caso de la Gráfica de Wright-Fisher, la dualidad de muestreo coincide con la dualidad de los momentos restringida a los primeros N momentos. Es decir, para $x = [N]/N$, y $n = [N]$, la probabilidad de que un individuo sea tipo $-$ en la generación $g+1$, si la frecuencia de individuos tipo $-$ en la generación anterior es x, es x (pues cada individuo elige a su padre uniformemente al azar).

Teorema 2.3.5. *(Dualidad en la gráfica de Wright Fisher (Moehle 1999)).*
Para todo $a \in \mathbb{Z}$ *y* $b \in \mathbb{Z}$*, el proceso de frecuencia* $X(g,a)$ *y el de ancestria* $A(g,b)$ *asociados a la Gráfica de Wright Fisher son momento duales, es decir, para todo* $x \in [N]/N$*,* $n \in [N]$ *y* $g \in \mathbb{N}$

$$\mathbb{E}_x[X(g,a)^n] = \mathbb{E}_n[x^{A(g,b)}] \tag{2.11}$$

Demostración. Para cualquier $x \in [N]/N$, $n \in [N]$ y $g \in \mathbb{N}$ fijos, fijamos una muestra de n individuos en la generación $g+1$ y asignamos tipos a los individuos en la generación 0 de tal forma que la frecuencia de individuos tipo $-$ en la generación 0 sea x. Usaremos la función de muestreo generalizada, (x, n, g) la cual introducimos en la ecuación del primer ejemplo. Noten que, condicionando en el proceso de ancestría obtenemos

$$\mathbb{P}(S(x,n,g)) = \mathbb{E}_n[S(x, A(g, g+1))] = \mathbb{E}_n[S(x, A(g,b))] = \mathbb{E}_n[x^{A(g,b)}]$$

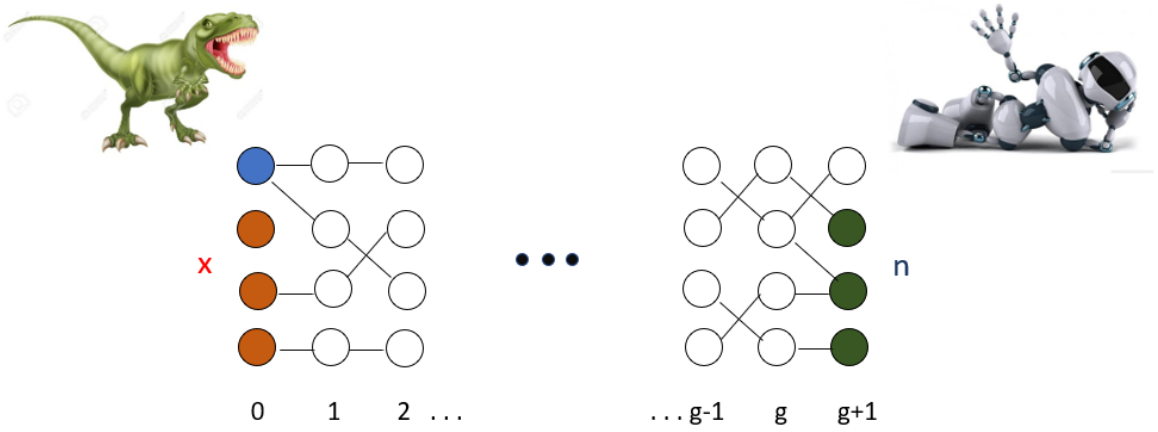
y condicionando en el de frecuencia

$$\mathbb{P}(S(x,n,g)) = \mathbb{E}_x[S(X(g,0),n)] = \mathbb{E}_x[S(g,a),n] = \mathbb{E}_x[X(g,a)^n]$$

Por lo tanto,

$$\mathbb{E}_x[X(g,a)^n] = \mathbb{E}_n[x^{A(g,b)}]$$

■



Capítulo 3

El Coalescente y la Difusión

3.1. Generadores

Definición 3.1.1. *Sea $(X_t)_{t\in\mathbb{R}^+}$ un proceso de Markov con valores en $\mathbb{R}$. El semi-grupo de operadores asociado a (X_t), aplicado a cualquier función $f : \mathbb{R} \to \mathbb{R}$, es:*

$$P_t(f(x)) = \mathbb{E}_x[f(X_t)]$$

Proposición 3.1.2. *$\{P_t\}_{t\geq 0}$ es un semi-grupo, donde la operación binaria es la composición.*

Demostración. Tenemos que probar que $P_tP_s = P_{t+s}$. Esto es consecuencia de la propiedad de Markov

$$P_{t+s}(f(x)) = \mathbb{E}_x[f(X_{t+s})] = \mathbb{E}_x[\mathbb{E}_x[f(X_{t+s})|X_t]] = \mathbb{E}_x[\mathbb{E}_{X_t}[f(X_s)]] = P_tP_s(f(x))$$

En la ultima igualdad aplicamos el operador P_t a la función $P_s(f(x)) = \mathbb{E}_x[f(X_t)]$

■

El semi-grupo asociado a (X_t) es una manera de ordenar todas las distribuciones finito dimensionales de (X_t). El generador de este semi-grupo es un objeto mucho mas pequeña que todo el semi-grupo, que de alguna forma resume toda esta información.

Definición 3.1.3. *El generador del semi-grupo $\{P_t\}_{t\geq 0}$ es un operador $\mathcal{A}$ que se define aplicado a cada $f : \mathbb{R} \to \mathbb{R}$ como el siguiente limite en la norma del supremo:*

$$\mathcal{A}f(x) := \lim_{t\to 0} \frac{P_t f(x) - P_0 f(x)}{t}$$

El conjunto de funciones para el cual dicho límite (en la norma del supremo) existe, se define como el dominio del generador y se denota $D(\mathcal{A})$.

El Generador y el semi-grupo se relacionan a través de la Ecuación del Generador, la cual es valida para toda $f \in D(\mathcal{A})$.

$$P_t(f(x)) - P_0(f(x)) = \int_0^t \mathbb{E}_x[\mathcal{A}f(X_s)]ds \tag{3.1}$$

Ejemplo 3.1.4. *(Cadena de Markov a tiempo continuo). Un proceso (X_t) con valores en un espacio discreto $E \subset \mathbb{R}$, tiene generador*

$$\mathcal{A}f(x) = \sum_{y\in E} r_{xy}[f(y) - f(x)] \tag{3.2}$$

para todo $x \in E$, donde $\{r_{xy}\}_{x,y\in E}$ es una colección de números no negativos tales que para todo x, $\sum_{y\in E} r_{xy} < \infty$, es ni mas ni menos que una cadena de Markov a tiempo continuo que va del estado x al estado y a tasa r_{xy}. Si el espacio de estados E no es discreto, la formula 3.2 se convierte en

$$\mathcal{A}f(x) = \int_{y\in E} [f(y) - f(x)]\mu_x(dy) \tag{3.3}$$

Ejemplo 3.1.5. *(Cadena de Markov a tiempo discreto y contínuo en el espacio de estados(generador discreto)).*

El λ-Generador o Generador discreto de una cadena de Markov a tiempo discreto X_g se define como el generador de X_{N_t}, donde N_t es un proceso de Poisson con intensidad λ. Si $X_1|_{X_0=x}$ tiene densidad μ_x, el generador discreto de X_g es

$$\mathcal{A}f(x) = \lambda \int_{y\in E} [f(y) - f(x)]\mu_x(dy) \tag{3.4}$$

Ejemplo 3.1.6. *(Ecuación diferencial).*

Una ecuación diferencial, es un ejemplo algo degenerado de un proceso estocástico. ¿Cual es su generador? Supongamos que $\{X_t^x\}_{x\in E}$ existe para todo x y es la única solución de

$$dX_t = g(X_t)dt$$

con valor inicial x, para una función continua $g : \mathbb{R} \to \mathbb{R}$. Para usar la notación usual de ecuaciones diferenciales y recordar que X_t es determinista, definimos la función $v^x(t) := X_t^x$, entonces el generador de X_t es

$$\mathcal{A}f(x) := \lim_{t\to 0} \frac{P_t f(x) - P_0 f(x)}{t} = \lim_{t\to 0} \frac{f(v^x(t)) - f(v^x(0))]}{t} = v'(t)f'(v^x(t)) = g(x)f'(x)$$

Si una familia de procesos estocásticos X_t^N converge débilmente a un proceso estocástico X_t, entonces sus distribuciones finito dimensionales convergen y por lo tanto sus semi-grupos y también sus generadores deben converger. Todo esto sucede bajo las siguientes hipótesis:

Teorema 17.25[1] Sea $X, X^1, X^2, ...$ procesos de Feller en S con semi-grupos $(T_t), (T_{1,t}), (T_{2,t})...$ y generadores $\mathcal{A}, \mathcal{A}_1, \mathcal{A}_2, ...,$ y fijar un núcleo D para $\mathcal{A}$. Entonces estas condiciones son equivalentes:

i) Si $f \in D$, existe algún $f_n \in dom(A_n)$ con $f_n \to f$ y $\mathcal{A}_n f_n \to \mathcal{A}f$;

ii) $T_{n,t} \to T_t$ fuertemente por cada $t > 0$;

[1]Foundations of Modern Probability, Pág. 331 (5)

iii) $T_{n,t}f \to T_t f$ para cada $f \in C_0$, uniformemente para funciones acotadas $t > 0$;

iv) Si $X_0^n \xrightarrow{d} X_0$ en S, entonces $X^n \xrightarrow{d} X$ en $D(\mathbb{R}_+, \hat{S})$

Teorema 17.28[2] (Aproximación de Cadenas de Markov) Sea $Y^1, Y^2, ...$ cadenas de Markov a tiempo discreto en S con operadores de transición $U_1, U_2, ...,$ y considerar el proceso Feller X en S con semi-grupo (T_t) y generador $\mathcal{A}$, fijos el nucleo D para $\mathcal{A}$ y suponga que $0 < h_n \to 0$. Entonces las condiciones $I)$ a $IV)$ del Teorema 17.25 siguen siendo equivalentes para los operadores y procesos

$$A_n = h_n^{-1}(U_n - I), \qquad T_{n,t} = U_n^{t/h_n}, \qquad X_t^n = Y_{[t/h_n]}^n$$

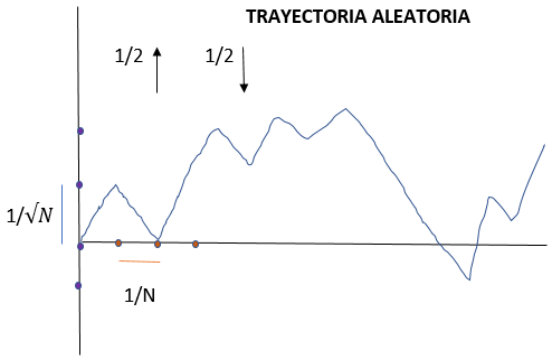
Usemos esta idea del ejemplo 3.1.6. para encontrar el generador del movimiento Browniano.

Ejemplo 3.1.7. *(Movimiento Browniano).*

Sea (S_t) *una caminata aleatoria simple, es decir su generador es*

$$\mathcal{A}f(n) = \frac{1}{2}[f(n - 1/2) - f(n)] + \frac{1}{2}[f(n + 1/2) - f(n)], \tag{3.5}$$

Definamos el proceso $(X_t^N) = \frac{1}{\sqrt{N}} S_{N_t}$, entonces su generador es



[2] Foundations of Modern Probability, Pág. 334 (5)

$$\mathcal{A}^N f(x) = \frac{1}{1/N}[\mathbb{P}_x(S_1 = x+\frac{1}{\sqrt{N}})[f(x+\frac{1}{\sqrt{N}})-f(x)]+\mathbb{P}_x(S_1 = x-\frac{1}{\sqrt{N}})[f(x-\frac{1}{\sqrt{N}})-f(x)]]$$

$$\mathcal{A}^N f(x) = N[\frac{1}{2}[f(x-\frac{1}{\sqrt{N}})-f(x)]+\frac{1}{2}[f(x+\frac{1}{\sqrt{N}})-f(x)]] \tag{3.6}$$

Supongamos que f es dos veces diferenciable $(f \in C_2)$ y apliquemos el teorema de Taylor

$$\begin{aligned}
\mathcal{A}^N f(x) &= N[\frac{1}{2}f(x-\frac{1}{\sqrt{N}})-\frac{1}{2}f(x)+\frac{1}{2}f(x+\frac{1}{\sqrt{N}})-\frac{1}{2}f(x)] \\
&= \frac{N}{2}[f(x-\frac{1}{\sqrt{N}})-2f(x)+f(x+\frac{1}{\sqrt{N}})] \\
&= \frac{N}{2}[f(x)-\frac{1}{\sqrt{N}}f'(x)+\frac{1}{2N}f''(x)-\frac{1}{3!N^{3/2}}f'''(x)+... \\
&\quad f(x)+\frac{1}{\sqrt{N}}f'(x)+\frac{1}{2N}f''(x)+\frac{1}{3!N^{3/2}}f'''(x)+... \\
&\quad -2f(x)] \\
&= \frac{N}{2}[\frac{1}{N}f''(x)+O(N^{-2})] \\
\mathcal{A}^N f(x) &= \frac{1}{2}f''(x)+O(1/N)
\end{aligned}$$

En efecto, el generador del movimiento browniano es

$$\mathcal{A}^N f(x) = \frac{1}{2}f''(x)+O(1/N)$$

y su dominio contiene a todas las funciones dos veces diferenciables.

Una forma de ver esto directamente es utilizando la formula de Ito. Si $f \in C_2$ y B_t es el movimiento Browniano, tenemos que

$$f(B_t) = f(x)+\int_0^t f'(B_s)dB_s+\int_0^t \frac{1}{2}f''(B_s)ds$$

Notar que $\mathbb{E}[f(B_t)] = f(x)+\mathbb{E}[\int_0^t \frac{1}{2}f''(B_s)ds]$, ya que $\mathbb{E}[\int_0^t f'(B_s)dB_s]$ es una martingala su esperanza es cero. Entonces,

$$\mathcal{A}^N f(x) := \lim_{t\to 0} \frac{P_t f(x) - P_0 f(x)}{t} = \lim_{t\to 0} \frac{\mathbb{E}[\int_o^t \frac{1}{2} f'' B_s ds]}{t} = \frac{1}{2} f''(x)$$

Ejemplo 3.1.8. *(Difusiones).*

Ahora imaginemos que tenemos una secuencia de cadenas de Markov a tiempo continuo con espacio de estados $\mathbb{R}$, muy parecidas a caminatas aleatorias con una pequeña deriva y que tarda en saltar en algunos sitios mas que en otros. Sea X_t^N (Un proceso que da saltos mas grandes en una dirección que en otra) el procesos estocástico con generador

$$\mathcal{A}^N f(x) = \sqrt{N}[\frac{1}{2}[f(x - \frac{1}{\sqrt{N}}) - f(x)] + \frac{1}{2}[f(x + \frac{1+\mu(x)}{\sqrt{N}}) - f(x)]], \tag{3.7}$$

para una función "Bonita"(Lipschitz-continuas, elipticas...) $\mu(x) : \mathbb{R} \to \mathbb{R}$. Aplicando Taylor obtenemos

$$\begin{aligned}
\mathcal{A}^N f(x) &= \sqrt{N}[\frac{1}{2} f(x - \frac{1}{\sqrt{N}}) - \frac{f(x)}{2} + \frac{1}{2} f(x + \frac{1+\mu(x)}{\sqrt{N}}) - \frac{f(x)}{2}] \\
&= \frac{\sqrt{N}}{2}[f(x - \frac{1}{\sqrt{N}}) - 2f(x) + f(x + \frac{1+\mu(x)}{\sqrt{N}})] \\
&= \frac{\sqrt{N}}{2}[f(x) - \frac{1}{\sqrt{N}} f'(x) + \frac{1}{2!N} f'' - \frac{1}{3!N^{3/2}} f'''(x) + \ldots \quad f(x) + \frac{1+\mu(x)}{\sqrt{N}} f'(x) + \\
&\quad (\frac{1+\mu(x)}{\sqrt{N}})^2 f''(x) + (\frac{1+\mu(x)}{\sqrt{N}})^3 f'''(x) + \ldots \quad - 2f(x)] \\
&= \frac{\sqrt{N}}{2}[\frac{1}{2N} f''(x) - \frac{1}{3!N^{3/2}} f'''(x) + \ldots \quad \frac{\mu(x)}{\sqrt{N}} f'(x) + \frac{(1+\mu(x))^2}{2N} f''(x) + \\
&\quad (\frac{1+\mu(x)}{\sqrt{N}})^3 f'''(x) + \ldots] \\
&= \frac{\sqrt{N}}{2}[\frac{\mu(x)}{\sqrt{N}} f'(x) + O(N^{-1/2})]
\end{aligned}$$

$$\Rightarrow \mathcal{A}^N f(x) = \frac{1}{2}\mu(x) f'(x) + O(N^{-1/2}), \tag{3.8}$$

Si ahora consideramos a X_t^N (un proceso que se tarda mas en saltar desde algunos lugares) como el procesos estocástico con generador

$$\mathcal{A}^N f(x) = N\sigma^2(x)[\frac{1}{2}[f(x - \frac{1}{\sqrt{N}}) - f(x)] + \frac{1}{2}[f(x + \frac{1}{\sqrt{N}}) - f(x)]] \tag{3.9}$$

para funciones "Bonitas" $\sigma(x) : \mathbb{R} \to \mathbb{R}$, obtenemos

$$\mathcal{A}^N f(x) = \frac{\sigma^2(x)}{2} f''(x) + O(N^{-1}) \tag{3.10}$$

Se puede considerar procesos que den saltos mas grandes hacia alguna dirección y que tarden mas en saltar desde algunos lugares, como X_t^N con generador

$$\mathcal{A}^N f(x) = N\sigma^2(x)[\frac{1}{2}[f(x - \frac{1}{\sqrt{N}}) - f(x)] + \frac{1}{2}[f(x + \frac{1 + (\mu(x)/\sqrt{N}\sigma^2(x))}{\sqrt{N}}) - f(x)]], \tag{3.11}$$

Para funciones "Bonitas"(Lipschitz-continuas, elipticas...) $\sigma(x) : \mathbb{R} \to \mathbb{R}$ y $\mu(x) : \mathbb{R} \to \mathbb{R}$. Aplicando Taylor obtenemos

$$\mathcal{A}^N f(x) = \mu(x) f'(x) + \frac{\sigma^2(x)}{2} f''(x) + O(N^{-1}) \tag{3.12}$$

Esta serie de ejemplos nos permite interpretar la deriva $\mu(x) : \mathbb{R} \to \mathbb{R}$ y la difusividad $\sigma(x) : \mathbb{R} \to \mathbb{R}$ de una difusión X_t, la cual se define como la solución (si es que existe) de la ecuación diferencial estocástica

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t,$$

pues es fácil usar la formula de Ito para verificar que el generador de dicha difusión es

$$\mathcal{A} f(x) = \mu(x) f'(x) + \frac{\sigma^2(x)}{2} f''(x) \tag{3.13}$$

Definición 3.1.9. *(Dualidad para generadores).*

Decimos que dos generadores $\mathcal{A}_1$, $\mathcal{A}_2$, de procesos con valores en E_1, $E_2 \subset \mathbb{R}$, son duales con respecto a una función $H : E_1 \times E_2 \to \mathbb{R}$, si para toda $x \in E_1$ y $n \in E_2$, $H(\cdot, n) \in D(\mathcal{A}_1)$ y $H(x, \cdot) \in D(\mathcal{A}_2)$ y se cumple la igualdad

$$\mathcal{A}_1 H(x, n) = \mathcal{A}_2 H(x, n)$$

donde $\mathcal{A}_1$ actúa en la primera entrada de H y $\mathcal{A}_2$ en la segunda.

Observación 3.1.10. *La dualidad de los generadores, más una serie de condiciones técnicas, implica la dualidad de los procesos. Este resultado puede encontrarse en la Proposición 1.2.*[3] *Sea $(X_t), (Y_t)$ procesos de Markov con generadores L^X, L^Y, sea $H : E \times F \to \mathbb{R}$ acotadas y continuas. Si $H(x, \cdot), P_t H(x, \cdot) \in D(L^Y)$ para toda $x \in E$, $t \geq 0$ y $H(\cdot, y), Q_t H(\cdot, y) \in D(L^X)$ para toda $y \in F$, $t \geq 0$, y si*

$$L^X H(\cdot, y)(x) = L^Y H(x, \cdot)(y) \qquad \forall x \in E, y \in F,$$

entonces (X_t) y (Y_t) son duales con respecto a H.

Ejemplo 3.1.11. *(La difusión de Wright Fisher y el coalescente de Kingman). Considere X_t a la solución de la ecuación diferencial estocástica*

$$dX_t = \sqrt{X_t(1 - X_t)} dB_t$$

X_t es la difusión de Wright Fisher y tiene generador

$$\mathcal{A}(f(x)) = x(1 - x)\frac{1}{2} f''(x)$$

[3] On the notion(s) of duality for Markov processes, Pág. 62 (4)

y su dominio contiene a las funciones dos veces derivables. Consideremos la función $H(x,n) = x^n$. Notemos que la segunda derivada de $H(x,n)$ con respecto a x es $H_{xx}(x,n) = n(n-1)x^{n-2}$. Entonces,

$$\mathcal{A}(H(x,n)) = x(1-x)\frac{n(n-1)}{2}x^{n-2} = \binom{n}{2}[x^{n-1} - x^n] = \binom{n}{2}[H(x,n-1) - H(x,n)].$$

El lado izquierdo es el generador de una cadena de Markov a tiempo continuo que va del estado n al $n-1$ a tasa $\binom{n}{2}$, aplicado a H. Es el generador del proceso de conteo de bloques del coalescente de Kingman.

Los generadores también nos ayudan a demostrar convergencia débil.

Definición 3.1.12. *Decimos que la secuencia de procesos de Markov (X_t^N) converge en el sentido de generadores al proceso X_t, si para toda $f \in C_2^b$(dos veces diferenciable y tal que $\lim_{x\to\infty} f(x) = 0$),*

$$\mathcal{A}^N(f(x)) \to \mathcal{A}(f(x))$$

uniformemente en x.

Observación 3.1.13. *Si (X_t^N) converge en el sentido de generadores al proceso (X_t), y se cumplen algunas condiciones técnicas, entonces $(X_t^N) \Rightarrow (X_t)$. Los detalles pueden encontrarse en el Teorema 17.25*

Para tener una interpretación concisa de la difusión de Wright Fisher (Como un movimiento Browniano cambiado de tiempo), definiremos al Modelo de Moran. En este modelo hay una población de N individuos pero el tiempo es continuo y a tasa uno sucede un evento. En cada evento un individuo uniformemente al azar es elegido para reproducirse y un individuo es elegido de la misma forma para morir (es posible que el mismo individuo se reproduzca y muera en un solo evento). El individuo recién producido toma el lugar del recién fallecido. La población no cambia hasta el siguiente

evento. Si asignamos tipos $\{+,-\}$ a cada individuo al tiempo cero y decimos que cada individuo copia el tipo de su padre, obtenemos el proceso de frecuencia del modelo de Moran.

Definición 3.1.14. *(El proceso de frecuencia del modelo de Moran).*
Para $N \in \mathbb{N}$ fijo, el proceso de frecuencia X_t^N del modelo de Moran es la cadena de Markov a tiempo continuo con espacio de estados $[N]/N$ y generador

$$\mathcal{A}^N f(x) = x(1-x)[f(x+1/N)-f(x)]+(1-x)x[f(x-1/N)-f(x)] = \frac{1}{N^2}x(1-x)f''(x)+O(\frac{1}{N})$$

Claramente $X^N_{N^2t}$ converge a la difusión de Wright Fisher en el sentido de los generadores y en este caso no es difícil extender este resultado a convergencia débil.
Decimos que dos secuencias de procesos son asintóticamente equivalentes si convergen al mismo límite.
En el siguiente ejemplo veremos que el modelo de Moran es asintóticamente equivalente al de Wright Fisher. Además, veremos porque en una difusión la función $\mu(x)$ se conoce como esperanza infinitesimal y la función $\sigma^2(x)$ como varianza infinitesimal.

Ejemplo 3.1.15. *El límite (en sentido de generadores) del proceso de frecuencia del modelo de Wright Fisher es fácil de calcular usando la Proposición 2.2.2 del segundo capítulo. En particular usaremos las Ecuaciones (2.2), (2.3) y (2.4). Usando la ecuación (2.2), notamos que el generador discreto de X_g^N, el proceso de frecuencia asociado a la gráfica de Wright Fisher con tamaño de población $N \in \mathbb{N}$, aplicado a $f \in C_2$, para todo $x \in [N]/N$, es (recordemos que $B(N,x)$ es una variable aleatoria binomial de parámetros N y x)*

$$\mathcal{A}f(x) = \mathbb{E}[f(B(N,x))-f(x)] = \mathbb{E}[B(N,x)-x]f'(x)+\mathbb{E}[(B(N,x)-x)^2]\frac{1}{2}f''(x)+O(1/N)$$

donde la segunda igualdad se sigue del Teorema de Taylor. Ahora usamos la ecuación (2.4) para concluir que

$$\mathcal{A}f(x) = \frac{1}{N}\frac{x(1-x)}{2}f''(x) + O(1/N)$$

Por lo tanto $X^N_{\lfloor Nt \rfloor}$ converge en el sentido de los generadores a la difusión de Wright Fisher. Por otro lado, el proceso de número de ancestros también converge en el sentido de los generadores. Se sigue de la Proposición 2.2.7 que el generador del proceso de ancestria asociado a la gráfica de Wright Fisher con tamaño de población $N \in \mathbb{N}$, aplicado a $f \in C_2$, para todo $n \in [N]$ es

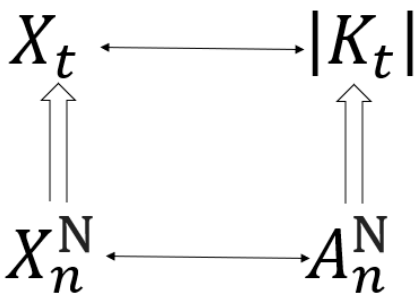
$$\mathcal{A}f(n) = \frac{1}{N}\binom{n}{2}[f(n-1) - f(n)] + O(1/N)$$

Por lo tanto el proceso de ancestria $A^N_{\lfloor Nt \rfloor}$ converge al proceso de conteo de bloques del coalescente de Kingman $|K_t|$. En resumen se muestra la siguiente diagrama

$$A^{\mathrm{N}}_{\lfloor Nt \rfloor} \Longrightarrow |K_t|$$

$$X^{\mathrm{N}}_{\lfloor Nt \rfloor} \Longrightarrow X_t$$

Una forma de probar que el coalescente de Kingman es momento dual de la difusión de Wright Fisher es utilizando el siguiente diagrama



En efecto, noten que $H(x,n) = x^n \in C_2^b$ cumple (en este caso la convergencia de los

generadores implica convergencia débil)

$$\mathbb{E}_n[x^{|K_t|}] = \lim_{N\to\infty} \mathbb{E}_n[x^{A^N_{\lfloor Nt\rfloor}}] = \lim_{N\to\infty} \mathbb{E}_x[(X^N_{\lfloor Nt\rfloor})^n] = \mathbb{E}_x[X^n_t]$$

3.2. El Coalescente de Kingman y la Difusión de Wright Fisher

Podemos usar la dualidad de los momentos para calcular los momentos de la difusión de Wright Fisher, recuerden que como (X_t) toma valores en $[0,1]$, los momento de X_t caracterizan su distribución.

Proposición 3.2.1. *Definimos* $M_n(x,t) = E_x[X^n_t]$. *Entonces para toda* $x \in [0,1]$, $M_1(x,t) = x$, $M_2(x,t) = x + e^{-t}x(1-x)$ *y* $\lim_{t\to\infty} M_n(x,t) = x$ *para toda* n.

Demostración. La prueba consiste en aplicar dualidad de los momentos

$$M_1(x,t) = \mathbb{E}_x[X_t] = \mathbb{E}_1[x^{|K_t|}] = x$$

Ahora, si el proceso de conteo de bloques inicia en el estado 2, hay dos opciones que hay se quede hasta el tiempo t o que se mueva al estado 1. Recordemos que $|K_t|$ salta del estado 2 al 1 a tasa 1. Entonces,

$$\begin{aligned} M_2(x,t) = \mathbb{E}_x[X^2_t] = \mathbb{E}_2[x^{|K_t|}] &= \mathbb{P}(|K_t| = 1)x + \mathbb{P}(|K_t| = 2)x^2 \\ &= (1-e^{-t})x + e^{-t}x^2 = x + e^{-t}x(1-x) \end{aligned}$$

Finalmente, notemos que $\lim_{t\to\infty} |K_t| \overset{a.s}{=} 1$, por lo tanto

$$\lim_{t\to\infty} M_n(x,t) = \lim_{t\to\infty} \mathbb{E}_n[x^{|K_t|}] = x$$

El Coalescente de Kingman nos gusta por varias razones, las dos principales son que es muy fácil de estudiar y que es un límite universal. Calculemos el tiempo esperado hasta el ancestro común en el Coalescente de Kingman.

Definición 3.2.2. *El tiempo hasta el ancestro común de un coalescente con* (C_t) *tal que* $C_0 = \{\{1\}, \{2\}, ..., \{n\}\}$ *es*

$$T_{MRCA}[n] = inf\{t > 0 : C_t = \{1, 2, ..., n\}\}$$

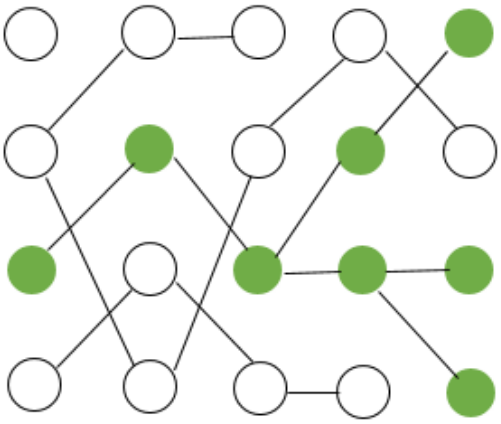
El tiempo de fijación de una difusión con saltos, Y_t*, con valores en* $[0, 1]$ *y estados absorbentes contenidos en* $\{0, 1\}$*, tal que* $Y_0 = x$*, es*

$$T_{fix}(x) = inf\{t > 0 : Y_t(1 - Y_t) = 0\}$$

.

Lema 3.2.3. *El tiempo hasta el ancestro común del coalescente de Kingman cumple*

$$\mathbb{E}[T_{MRCA}[n]] = 2(1 - \frac{1}{n})$$



Demostración.

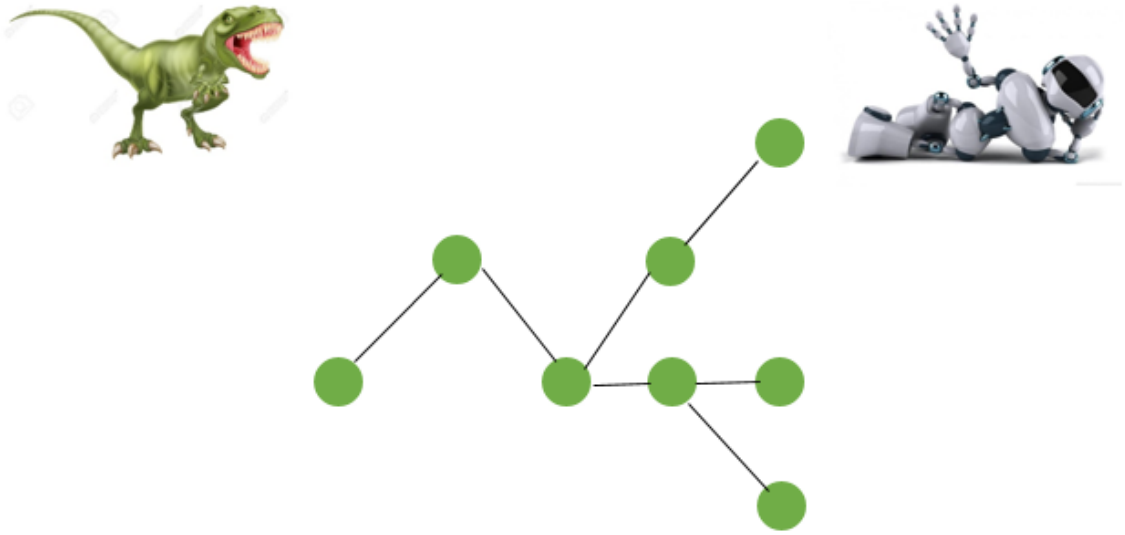
Definimos $\tau_i = inf_{t>0}\{|K_t| = n - i\}$, para todo $i \in [n-1]$. Notemos que $\tau_i - \tau_{i-1} = inf_{i,j\in[n-i+1]} e^i_{i,j} \sim exp(\binom{n}{2})$, donde $e^i_{i,j}$ es una variable aleatoria exponencial de parámetro 1. Entonces

$$T_{MRCA}[n] \overset{d}{=} \tau_n + \tau_{n-1} + ... + \tau_2 \qquad \mathbb{E}[\tau_1] = \frac{1}{\binom{n}{2}}, \mathbb{E}[\tau_2 - \tau_1] = \frac{1}{\binom{n-1}{2}}, ..., \mathbb{E}[\tau_i - \tau_{i-1}] = \frac{1}{\binom{n-i+1}{2}}$$

$$\mathbb{E}[T_{MRCA}[n]] = \mathbb{E}[\tau_{n-1}] = \sum_{i=1}^{n-1} \mathbb{E}[\tau_i - \tau_{i-1}] = \sum_{i=2}^{n} \frac{1}{\binom{i}{2}} = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\sum_{i=2}^{n} \frac{1}{i-1} - \frac{1}{i}$$

El resultado se sigue al observar que $\sum_{i=2}^{n} \frac{1}{i-1} - \frac{1}{i} = 1 - \frac{1}{n}$, pues es una suma telescópica.

■



También podemos calcular la esperanza de la longitud de árbol del coalescente de kingman.

Definición 3.2.4. *La longitud de árbol de un coalescente* (C_t) *tal que* $C_0 = \{\{1\}, \{2\}, ..., \{n\}\}$

$$L(n) = \int_0^{T_{MRCA}[n]} |C_t|$$

Lema 3.2.5. *La longitud de árbol del coalescente de Kingman cumple*

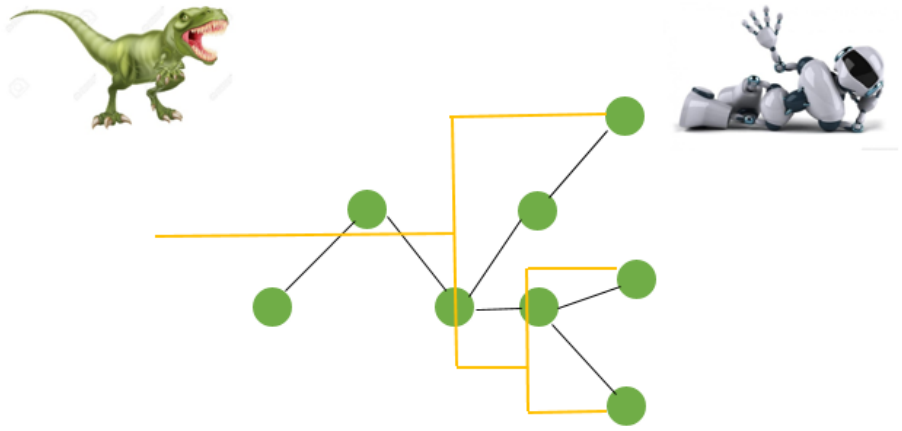
$$\mathbb{E}[L(n)] \sim 2log(n)$$

Demostración. Usando la misma notación que en el lema anterior tenemos que

$$L(n) \stackrel{d}{=} n\tau_n + (n-1)\tau_{n-1} + ... + 2\tau_2$$

$$\mathbb{E}[L(n)] = \sum_{i=2}^{n} i\mathbb{E}[\tau_i - \tau_{i-1}] = \sum_{i=2}^{n} \frac{i}{\binom{i}{2}} = 2\sum_{i=2}^{n-1} \frac{1}{i-1} = 2\sum_{i=1}^{n} \frac{1}{i} \sim 2log(n)$$

La última serie es conocido en matemática como el número Armónico, crece igual de rápido que el logaritmo natural de n. La razón es que la suma está aproximada por la integral, cuyo valor es $log(n)$. ■



El Lema 3.2.5. implica que el tiempo de fijación de la difusión de Wright Fisher tiene que ser finito y más aun, su esperanza tiene que ser menor a 2 para todo estado inicial. Podemos hacer un cálculo mucho mas preciso del tiempo de fijación.

Lema 3.2.6. *El tiempo de fijación de la difusión de Wright Fisher cumple para todo* $x \in (0,1)$.

$$\mathbb{E}[T_{fix}(x)] = g(x) = -2(xlogx + (1-x)log(1-x))$$

Demostración. El truco es condicionar el tiempo de fijación en la posición de la difusión después de un instante.

$$\mathbb{E}[T_{fix}(x)] = \lim_{\epsilon \to 0} \mathbb{E}[T_{fix}(x) | T_{f_{ix}}(x) > \epsilon] = \lim_{\epsilon \to 0} \mathbb{E}[T_{fix}(X_\epsilon)] + \epsilon = \lim_{\epsilon \to 0} P_\epsilon g(x) + \epsilon$$

Donde P_t es el semi-grupo de X_t, ahora dividiendo entre ϵ obtenemos:

$$0 = \lim_{\epsilon \to 0} \frac{P_\epsilon g(x) - g(x)}{\epsilon} + 1 = \mathcal{A} g(x) + 1 \tag{3.14}$$

donde $\mathcal{A}$ es el generador de X_t. Cabe resaltar que hasta ahora no hemos utilizado que X_t es la difusión de Wright Fisher y por lo tanto la ecuación (3.14) es válida para todo proceso de Markov. Lo único que falta es verificar que g cumple la ecuación (3.14). Para eso notemos que

$$g'(x) = -2(log(x) + 1 - log(1-x) - 1) = -2(log(x) - log(1-x))$$

y

$$g''(x) = -2(\frac{1}{x} + \frac{1}{1-x}) = \frac{-2}{x(1-x)}$$

La prueba se termina al recordar que $Ag(x) = \frac{1}{2} x(1-x) f''(x)$ y utilizar las condiciones de frontera $g(0) = g(1) = 0$

■

Capítulo 4

Modelo de Wright-Fisher Generalizado

Para finalizar este trabajo de investigación, en este capítulo estudiaremos el modelo de Wright-Fisher generalizado, es decir, con tamaño de población variable, tal vez una de las hipótesis más cuestionables (en términos de aplicabilidad biológica) del modelo de Wright-Fisher, es que la población es constante; las poblaciones raras veces són constantes. Los siguiente resultados nos permite estudiar la universalidad del coalescente de Kingman en términos de las variaciones del tamaño de la población.

4.1. Modelo Cannings

El modelo de Cannings es una generalización del modelo de Wright-Fisher. Son fáciles de estudiar y sirven para darse una idea de que tan universal es el Coalescente de Kingman como límite.

Definición 4.1.1. *(Modelo de Cannings).*

Sea $(\nu_{1,g}^N, \nu_{2,g}^N, ..., \nu_{N,g}^N)_{g\in\mathbb{Z}}$ *una familia de variables aleatorias independientes e idénticamente distribuidas, con distribución* μ_N*, y con valores en* $\mathbb{N}_0 \times N$ *tales que* $(\nu_{1,g}^N, \nu_{2,g}^N, ..., \nu_{N,g}^N)$ *es un vector intercambiable y* $\sum_{i=1}^N \nu_{i,g} = N$*. Consideremos el orden* $v = (i, g) < v' = (i', g')$ *si* $g > g'$ *ó* $g = g'$*, es decir,* $i < i'$*. Una vez que todos los individuos menores*

a $v = (i, g)$ han elegido a sus hijos, el individuo v elige uniformemente al azar ν_v individuos de los individuos en la generación $g + 1$ que aún no tienen padre. El modelo de Cannings con parámetros N y μ_N es la gráfica aleatoria con vértices $V = [N] \times \mathbb{Z}$ y aristas E, tal que para todo $v < v'$, $(v, v') \in E$ si y solo si v es padre de v'.

Definimos $C_N = \mathbb{E}[\frac{\nu_{i,g}^N(\nu_{i,g}^N-1)}{N-1}]$, observemos que C_N es la posibilidad de que dos individuos de la misma generación sean hermanos. De hecho, esta probabilidad puede calcularse sumando

$$\mathbb{E}[\sum_{i=1}^{N} \frac{\nu_{i,g}^N(\nu_{i,g}^N - 1)}{N(N-1)}] = \sum_{i=1}^{N} \mathbb{E}[\frac{\nu_{i,g}^N(\nu_{i,g}^N - 1)}{N(N-1)}]$$

Además, definimos $B_N = \mathbb{E}[\frac{\nu_{i,g}^N(\nu_{i,g}^N-1)(\nu_{i,g}^N-2)}{(N-1)(N-2)}]$, que es la probabilidad de que tres individuos en la misma generación sean hermanos

$$\mathbb{E}[\sum_{i=1}^{N} \frac{\nu_{i,g}^N(\nu_{i,g}^N - 1)(\nu_{i,g}^N - 2)}{N(N-1)(N-2)}] = \sum_{i=1}^{N} \mathbb{E}[\frac{\nu_{i,g}^N}{N} \frac{(\nu_{i,g}^N - 1)}{(N-1)} \frac{(\nu_{i,g}^N - 2)}{(N-2)}]$$

Teorema 4.1.2. *(Möhle).*

Sea (A_g^N) el proceso de número de ancestro asociado al modelo de Cannings con parámetros N y μ_N . Supongamos que

1. $C_N \to 0$
2. $B_N/C_N \to 0$ *(el propósito es exigir que la tasa a la que se unen tres o más linajes es insignificante en comparación con la tasa de fusiones por pares)*

Entonces, $(A_{\lfloor t/C_N \rfloor}^N) \to |K_t|$, en el sentido de los generadores(definición 3.1.12)

Demostración.

La prueba consiste en calcular el generador discreto, $\mathcal{A}^N$ de $(\mathcal{A}_{\lfloor t/C_N \rfloor}^N)$.

Sea $f : \mathbb{N} \to \mathbb{R} \in C_b$

$$\begin{aligned}
\mathcal{A}^N f(n) &= C_N^{-1}\mathbb{E}_n[f(A_1^N) - f(n)] \\
&= C_N^{-1}[[f(n-1) - f(n)]\mathbb{P}(A_1^N = n-1) + \\
&\quad \mathbb{E}_n[f(A_1^N) - f(n) | A_1^N < n-1]\mathbb{P}(A_1^N < n-1)] \\
&= C_N^{-1}[[f(n-1) - f(n)]\binom{n}{2}C_N + O(C_N^2 + B_N) + \\
&\quad \mathbb{E}_n[f(A_1^N) - f(n) | A_1^N < n-1]O(C_N^2 + B_N)] \\
&= [f(n-1) - f(n)]\binom{n}{2} + O(C_N + C_N^{-1}B_N)
\end{aligned}$$

■

Ejemplo 4.1.3. *El modelo de Wright-Fisher es un modelo de Cannings.*

Sean $(\nu_{1,g}^N, \nu_{2,g}^N, ..., \nu_{N,g}^N) \sim mult(N, 1/N)$ *que tienen distribución multinomial con parámetros* N *y* $1/N$ *(lo que implica que es intercambiable) y* $\nu_{i,g}^N \sim Bin(N, 1/N)$. *Recordemos que el segundo momento de una binomial* $Bin(N, x)$ *es* $M_2 = Nx(1 - x + Nx)$. *Entonces*

$$C_N = \mathbb{E}[\frac{\nu_{i,g}^N(\nu_{i,g}^N - 1)}{N-1}] = \frac{1}{N-1}\mathbb{E}[(\nu_{1,g}^N)^2 - \nu_{1,g}^N] = \frac{1}{N-1}(2 - \frac{1}{N} - 1) = \frac{1}{N}$$

Por otro lado, el tercer momento de una binomial es $M_3 = Nx(1 - 3x + 3Nx + 2x^2 - 3Nx^2 + N^2x^2)$ *que con los parámetros que nos ocupan es de orden* N. *Por lo tanto* $B_N = O(N^{-2})$ *y* $B_N/C_N \to 0$

Ahora estudiaremos otra familia para la cual no siempre el límite de escala es el coalescente de Kingman, pero muchas veces si.

Definición 4.1.4. *(El modelo de Wright Fisher con tamaño de la población IID).*

Sea W una variable aleatoria con espacio de estados $(0, 1]$. *Sea* W^N *una variable aleatoria con espacios de estados* $[N]$, *tal que para todo* $i \in [N]$, $\mathbb{P}(W^N = i) = \mathbb{P}(W \in$

$((i-1)/N, i/N))$. *Sea* $(W_g^N)_{g\in Z}$ *una familia de variables aleatorias independientes e idénticas en distribución a* W^N*. El modelo de Wright-Fisher con población IID, con parámetros* W *y* N *es la gráfica aleatoria* (V, E)*, donde* $V = \{(i, g) : i \in [W_g^N]\}$*, cada individuo* $(i, g) \in V$ *elige a su padre de manera uniforme entre los* W_{g-1}^N *individuos de la generación* $g-1$ *y el conjunto de aristas* $E := \{((j, g-1)(i, g)) : (j, g-1)\}$ *es el padre de* (i, g)

El caso $\mathbb{P}(W = 1) = 1$ es simplemente el modelo de Wright Fisher. Observamos que si existe $\epsilon > 0$ tal que $\mathbb{P}(W > \epsilon) = 1$, el proceso de ancestría asociado al modelo de Wright Fisher con tamaño de la población IID converge a un cambio de tiempo constante del proceso de conteo de bloques del coalescente de Kingman, es decir:

$$A^N_{\lfloor N_t \rfloor} \to |K_{ct}|$$

La prueba no es difícil. Noten que en este caso

$$C_N = \sum_{i=1}^{N} \frac{1}{i}\mathbb{P}(W^N \in [\frac{1}{i}, \frac{1}{i+1})) = \sum_{i=\lfloor \epsilon N \rfloor}^{N} \frac{1}{i}\mathbb{P}(W^N \in [\frac{1}{i}, \frac{1}{i+1})) \in [\frac{1}{N}, \frac{1}{\epsilon N}]$$

$$y \quad B_N = \sum_{i=1}^{N} \frac{1}{i^2}\mathbb{P}(W^N \in [\frac{1}{i}, \frac{1}{i+1})) = \sum_{i=\lfloor \epsilon N \rfloor}^{N} \frac{1}{i^2}\mathbb{P}(W^N \in [\frac{1}{i}, \frac{1}{i+1})) \in [\frac{1}{N^2}, \frac{1}{(\epsilon N)^2}]$$

Que pasa si W es una variable aleatoria uniforme en $[0, 1]$ En este caso cada N generaciones esperamos ver una generación en donde solo hay un individuo y en donde todos los linajes deben coalescer. ¿Es posible que el límite en este caso también sea Kingman?, en efecto es el coalescente de Kignman.

Noten que, en general,

$$C_N = \sum_{i=1}^{N} \frac{1}{i}\mathbb{P}(W^N = i) = \sum_{i=1}^{N} \frac{1}{i}\mathbb{P}(W \in ((i-1)/N, i/N)) = \sum_{i=1}^{N} \frac{1}{i}\mathbb{P}(NW \in (i-1, i))$$

$$B_N = \sum_{i=1}^{N} \frac{1}{i^2}\mathbb{P}(NW \in (i-1), i))$$

En el caso en que W es uniforme $C_N = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{i} \sim \frac{log(N)}{N}$ mientras que $B_N = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{i^2} = O(1/N)$, entonces $C_N \to 0$ y $B_N/C_N = O(1/log(N))$ en particular $B_N/C_N \to 0$. El teorema de Mohle implica que

$$A_{\lfloor tN/log(n) \rfloor} \Rightarrow |K_t|$$

Existe W tal que $C_N \to 0$ pero B_N/C_N no tiende a 0. La pregunta es equivalente a encontrar W tal que

$$\frac{\sum_{i=1}^{N} \frac{1}{i}\mathbb{P}(NW \in (i-1, i))}{\sum_{i=1}^{N} \frac{1}{i^2}\mathbb{P}(NW \in (i-1), i))} \nrightarrow 0$$

Para encontrar más ejemplos de estos, se deje como referencia un documento realizado por el Dr. Adrian González Casanova en junio de 2019 titulado “The Symmetric Coalescent and Wright-Fisher models with bottlenecks”(6), en donde se podrán encontrar que $C_N \to 0$ y $B_N/C_N \nrightarrow 0$, por lo tanto se sale de la universalidad del coalescente de Kignman.

CONCLUSIÓN

Los procesos estocásticos es un claro ejemplo en genética de poblaciones que la matemática y sus aplicaciones a la vida real existen. Se concluye que el Modelo Wrgiht-Fisher con población constante, modelo neutral de la genética de poblaciones es la introducción perfecta para esta área tan grande y fascinante que existe. Los procesos estocásticos nos dan la pauta y ayuda necesaria para el análisis del cambio en el perfil genético de una población ya sea constante o variable, resultando así el proceso de frecuencia que nos dice si un tipo de individuo pueda extinguirse o sobrevivir. Por otra parte tomando una muestra de individuos de una población en la generación g, para definir el Proceso de Ancestría, que encontrará el ancestro en común.

Se debe de observar que rescalar el tiempo en el proceso de frecuencia y el proceso de ancestria, convergen a la difusión de Wright-Fisher y Coalecente de Kimgman respectivamente, siendo esto un resultado importante para saber cuanto tiempo tarda en fijarse un individuo de cierto tipo o para conocer el tiempo hasta el ancestro común (TMRCA) de una muestra.

Por último pero no menos importante, el modelo de Cannings, modelo generalizado del modelo Wright-Fisher, es el preámbulo para el modelo de Wright-Fisher independiente e idénticamente distribuidas(IID), se obtienen resultados que son más aplicados a la realidad ya que se estudia y trabaja con poblaciones variables que en biología sería la hipótesis más lógica, que la población varié conforme pase el tiempo.

Bibliografía

[1] Rick Durrett, *Probability Models for DNA Sequence Evolution*, segunda edicion, Copyright 2008, https : //services.math.duke.edu/ rtd/Gbook/PM4DNA0317.pdf

[2] Richard Durrett , «Probability and Stochastic Series», *CRC Press; Edición: 2 (1 de junio de 1996)*, http://www.gbv.de/dms/goettingen/213194988.pdf

[3] Beichelt, F y Fatti, P, *Stochastic Processes and Their Applications*, *CRC Press.N.Y*.

[4] Jansen, Sabine y Kurt, Noemi, *On the notion(s) of duality for Markov processes*, volumen 11, 2014.

[5] Kallenberg, Olav, *Foundations of Modern Probability*, Department of Mathematics, Auburn University, USA, 1997.

[6] González Casanova, Adrián; Miró Pina, Verónica y Siri-Jégousse, Arno, *The Symmetric Coalescent and Wright-Fisher models with bottlenecks*, 18 de junio de 2019, https://arxiv.org/pdf/1903.05642.pdf

[7] Linda J. S. Allen, *An Introduction to Stochastic Processes with Applications to Biology*, Department of Mathematics and Statistics Texas Tech University.

[8] Yilmaz, Bertan, Oti-Aboagye, Richard y Liu, Di, *Partial Dierential Equations with Applications to Finance*, 10 páginas.

[9] Engel, Klaus-Jochen y Nagel, R., *A short course on operator semigroups*, Springer, New York, N.Y., 2006.

[10] Möhle, Martin, *The Concept of Duality and Applications to Markov Processes Arising in Neutral Population Genetics Models*, volumen 5, 2000.

[11] Frank den Hollander, *Stochastic Models For Genetic Evolution*, Mathematical Institute, Leiden University, 2013.