

DOCTORADO EN CIENCIA Y TECNOLOGIA

Evaluado y acreditado por la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU).
Resolución N° 1178/11. Calificación "B".

*Deriva genética: nuevos modelos con superposición de
generaciones y aplicaciones*

Trabajo de tesis para optar por el título de Doctor en Ciencia y Tecnología de la Universidad
Nacional de General Sarmiento

Autor: **Ana Laura Maffei**

Director: Dr. Tomás Tetzlaff

Fecha: 10 de Marzo de 2022



DOCTORADO EN CIENCIA Y TECNOLOGIA

Evaluado y acreditado por la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU).
Resolución N° 1178/11. Calificación "B".

FORMULARIO "E" TESIS DE POSGRADO

Niveles de acceso al documento autorizados por el autor

El autor de la tesis puede elegir entre las siguientes posibilidades para autorizar a la UNGS a difundir el contenido de la tesis: __a)__

- a) *Liberar el contenido de la tesis para acceso público.*
- b) *Liberar el contenido de la tesis solamente a la comunidad universitaria de la UNGS:*
- c) *Retener el contenido de la tesis por motivos de patentes, publicación y/o derechos de autor por un lapso de cinco años.*

- a. Título completo del trabajo de Tesis: Deriva genética: nuevos modelos con superposición de generaciones y aplicaciones
- b. Presentado por (Apellido/s y Nombres completos del autor): Ana Laura Maffei
- c. E-mail del autor: analauramaffei@gmail.com
- d. Estudiante del Posgrado (consignar el nombre completo del Posgrado): Doctorado en Ciencia y Tecnología.
- e. Institución o Instituciones que dictaron el Posgrado (consignar los nombres desarrollados y completos): Universidad Nacional de General Sarmiento.
- f. Para recibir el título de (consignar completo):
a) Grado académico que se obtiene: Doctor
b) Nombre del grado académico: Doctor en Ciencia y Tecnología
- g. Fecha de la defensa: 13 / 05 /2022
día mes año
- h. Director de la Tesis (Apellidos y Nombres): Dr. Tomás Tetzlaff
- i. Tutor de la Tesis (Apellidos y Nombres):
- j. Colaboradores con el trabajo de Tesis: Dr. Carlos De Brasi
- k. Descripción física del trabajo de Tesis (cantidad total de páginas, imágenes, planos, videos, archivos digitales, etc.): 142 páginas y 14 imágenes.
- l. Alcance geográfico y/o temporal de la Tesis:
- m. Temas tratados en la Tesis (palabras claves): Deriva genética, Deriva cromosómica, Heterocigosidad, Tamaño poblacional efectivo, Número efectivo de generaciones.

DOCTORADO EN CIENCIA Y TECNOLOGIA

Evaluado y acreditado por la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU).
Resolución N° 1178/11. Calificación "B".

n. Resumen en español (hasta 1000 caracteres):

Definimos dos procesos estocásticos de deriva con renovación parcial en una población de tamaño constante. El primero generaliza el modelo de Wright-Fisher con el objetivo de describir los cambios de una frecuencia alélica en una plaga con posible mutación. El segundo aborda la deriva en una población celular de dos tipos describiendo el fenómeno de inactivación de cromosomas X en mujeres. Para ambos modelos cuantificamos la pérdida de diversidad mediante cálculo de la heterocigosidad, el tamaño poblacional efectivo y el número efectivo de generaciones. Finalmente proponemos una variante del modelo de urna de Polya-Eggenberger aplicado a la deriva en una población celular pequeña en crecimiento. Encontramos que la probabilidad de sesgo extremo en el patrón de inactivación inicial, suponiendo que la división celular no es sincrónica, tiende a aumentar por deriva durante la expansión clonal y calculamos las probabilidades límite para el caso de inactivación inicial binomial.

o. Resumen en portugués (hasta 1000 caracteres):

Definimos dois processos de deriva estocástica com renovação parcial em uma população de tamanho constante. O primeiro generaliza o modelo de Wright-Fisher para descrever as mudanças de uma frequência alélica em uma praga com uma possível mutação. O segundo aborda a deriva em uma população de células de dois tipos, descrevendo o fenômeno da inativação do cromossomo X em mulheres. Para ambos os modelos, quantificamos a perda de diversidade calculando a heterozigossidade, o tamanho efetivo da população e o número efetivo de gerações. Finalmente, propomos uma variante do modelo de urna Polya-Eggenberger aplicado à deriva em uma pequena população celular em crescimento. Descobrimos que a probabilidade de viés extremo no padrão de inativação inicial, assumindo que a divisão celular não é síncrona, tende a aumentar por deriva durante a expansão clonal e calculamos as probabilidades no limite para o caso de inativação inicial binomial.

p. Resumen en inglés (hasta 1000 caracteres):


We define two stochastic drift processes with partial renewal in a population of constant size. The first generalizes the Wright-Fisher model in order to describe the changes of an allelic frequency in a pest with a possible mutation. The second addresses drift in a cell population of two types by describing the phenomenon of X chromosome inactivation in women. For both models we quantify the loss of diversity by calculating the heterozygosity, the effective population size and the effective number of generations. Finally, we propose a variant of the Polya-Eggenberger urn model applied to drift in a small growing cell population. We find that the probability of extreme bias in the initial inactivation pattern, assuming that cell division is not synchronous, tends to increase by drift during clonal expansion and we calculate the limit probability for the binomial initial inactivation case.

q. Aprobado por (Apellidos y Nombres del Jurado): Pablo Augusto Ferrari, Alberto Leonardo Maltz y Juan Ignacio Túnez.

DOCTORADO EN CIENCIA Y TECNOLOGIA

Evaluado y acreditado por la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU).
Resolución N° 1178/11. Calificación "B".

Firma y aclaración de la firma del Presidente del Jurado:

 Pablo Augusto
Ferrari

Firma del autor de la tesis:



Ana Laura Maffei



DOCTORADO EN CIENCIA Y TECNOLOGIA

Evaluado y acreditado por la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU).
Resolución N° 1178/11. Calificación "B".

Deriva genética: nuevos modelos con superposición de generaciones y aplicaciones

Publicaciones:

Abstracts extendidos

- De Brasi C., Maffei A. L. y Tetzlaff T., Dispersion of drift processes with application to X inactivation in different tissues. Congreso latinoamericano de biomatemática, SOLABIMA 2015.
- De Brasi C., Maffei A. L. y Tetzlaff T., Números efectivos para un modelo de deriva con renovación parcial de la población. Congreso latinoamericano de biomatemática, SOLABIMA 2017.

Aportes Originales:

Se proponen dos modelos de deriva para poblaciones de tamaño constante e individuos de dos tipos, en los cuales en cada generación, una fracción aleatoria de la población muere y otra se reproduce para restablecer el tamaño poblacional (Capítulo II). El primero es una generalización del modelo de Wright-Fisher por la naturaleza estocástica del esquema de reproducción. En el segundo, el grupo de individuos que se reproduce restaura el número poblacional con la misma proporción de tipos presente en el grupo. Para ambos modelos obtuvimos expresiones para las funciones de heterocigosidad y varianza a lo largo de las generaciones, así como para el número efectivo poblacional. Para completar el estudio basado en la comparación con el proceso clásico de Wright-Fisher, definimos y calculamos una magnitud que llamamos número efectivo de generaciones.

Aplicamos el esquema de expansión aleatoria al contexto de control de plagas con eventual mutación y mostramos que ante episodios repetidos de fumigación, la deriva de una frecuencia alélica puede ser significativa aún si una parte considerable de la población sobrevive en cada episodio y en ausencia de selección. El proceso con expansión proporcional determinística se aplica a la dinámica del sesgo de inactivación de cromosomas X permitiendo cuantificar la deriva durante la vida de una mujer a partir del momento en que los tejidos alcanzan un número estable de células (Capítulo IV).

Definimos un tercer proceso que consiste en una variante del modelo de urna de Polya-Eggenberger en el que la población inicial tiene una composición estocástica (Capítulo III), con el fin de representar la distribución de células con inactivación del cromosoma X materno

DOCTORADO EN CIENCIA Y TECNOLOGIA

Evaluable y acreditado por la Comisión Nacional de Evaluación y Acreditación Universitaria (CONEAU).
Resolución N° 1178/11. Calificación "B".

que resulta del fenómeno de Lyonización y posterior crecimiento poblacional con reproducción asincrónica (Capítulo IV). El cálculo de las probabilidades de sesgo en la inactivación permite inferir un aumento del mismo en la etapa inmediatamente posterior a la Lyonización, de interés en hemofilia y otras enfermedades ligadas al cromosoma X.

Deriva genética: nuevos modelos con
superposición de generaciones y aplicaciones

Autor: Ana Laura Maffei
Director: Dr. Tomás Tetzlaff

Universidad Nacional de General Sarmiento
Instituto de Ciencias

10 de Marzo de 2022

Agradecimientos

Gracias:

A mi director de tesis, Tomás, por todo lo que me enseñó en estos años, por la paciencia que supo tenerme y por acompañarme a lo largo de mi carrera.

A Carlos De Brasi, por aceptar ser mi director de beca Conicet y siempre estar predispuesto a ayudarme.

Al jurado por tomarse el tiempo de leer la tesis y por sus devoluciones.

A mis amigas Cristina y Virginia porque siempre me escucharon, aconsejaron y estuvieron acompañándome todos estos años.

A mis compañeros de la UNGS, Lucia, Martín, Meli, Mariana, Jheisson, Mónica y Raquel que siempre estuvieron atentos a mis avances. Siempre fue muy lindo compartir un mate con ellos con largas charlas de por medio.

A mis compañeros de la UNLu, Adri, Euge, Lucia, Nico y Seba que siempre estuvieron dispuestos a ayudarme y a escucharme. Siempre es lindo ir a la UNLu y estar con ellos.

A Alejandro Varela porque es un gran docente y fue con quien descubrí que existía la posibilidad de hacer investigación.

A mi familia, a mi hermana, a mi papá, a mis sobrinos y a mi ahijado.

A mi mamá que siempre es la primera en confiar en mí, es quien siempre se alegra con cada uno de mis logros y quien desde que inicié mis estudios me acompañó con sus mates de madrugada sin importar el día. Todo lo que soy es gracias a ella.

A mi hermano Fer, que sin entender mis largas horas de estudio siempre me acompañó y a mi sobrina Donatella, ser su tía es una de las cosas más lindas que me pasaron.

A Ri, que siempre me acompañó en este largo camino.

A Juan que siempre estuvo al lado mío y me acompaña, sin duda es el mejor compañero. Este trabajo en gran parte lo pude lograr por su apoyo incondicional.

A Juan Martín, quién puso mi vida patas para arriba y cambió todas mis prioridades, pero sin duda es lo más hermoso que tengo, es quien hace que mis días sean perfectos.

Resumen

En este trabajo revisamos resultados existentes sobre el proceso estocástico de deriva, proponemos nuevos modelos y explicamos sus aplicaciones. En primer lugar definimos dos procesos que se caracterizan por la renovación parcial de la población y superposición de generaciones. El primero generaliza al modelo de Wright-Fisher con el objetivo de describir los cambios de una frecuencia alélica en una plaga en condiciones similares a las de un cultivo intensivo con posible mutación. Con el segundo abordamos el problema de deriva en una población celular de dos tipos, motivado por el fenómeno de inactivación de cromosomas X en mujeres, de interés en el estudio de enfermedades ligadas al cromosoma X como la hemofilia. El recambio celular es descrito como un proceso estocástico de muestreo y expansión clonal con reemplazo parcial de la población en cada generación. Para ambos modelos cuantificamos la pérdida de diversidad en la población a partir del cálculo de las expresiones de la heterocigosidad, el tamaño poblacional efectivo y el número efectivo de generaciones, en comparación con el proceso de Wright-Fisher clásico. Finalmente proponemos una variante del modelo de urna de Polya-Eggenberger aplicado a la división celular mitótica. Se trata de una alternativa al supuesto usualmente aceptado de distribución binomial en el patrón de inactivación inicial del cromosoma X con mantenimiento de dicho patrón por expansión clonal sincrónica. Encontramos que la probabilidad de sesgo extremo en el patrón de inactivación inicial, suponiendo que la división celular mitótica no es sincrónica, tiende a aumentar por deriva durante la expansión clonal y calculamos las probabilidades límite para el caso de inactivación inicial binomial.

Índice general

Resumen	I
1. Deriva Genética	1
1.1. Modelo de Wright-Fisher	3
1.1.1. Cadena de Markov absorbente y martingala	3
1.1.2. Heterocigosidad	10
1.1.3. Coalescencia	14
1.1.4. Tamaño efectivo	15
1.2. Procesos de Cannings	16
1.2.1. La varianza en los procesos de Cannings	18
1.2.2. Autovalores de la matriz de transición	20
1.2.3. Heterocigosidad	27
1.2.4. Tamaño efectivo	28
1.2.5. El modelo de Wright-Fisher como caso particular del modelo de Cannings	30
2. Generalización y variante del modelo de Wright-Fisher	33
2.1. Expansión determinística y expansión aleatoria	33
2.1.1. El número de descendientes por individuo	37
2.1.2. Autovalores	44
2.1.3. Deriva genética y heterocigosidad	45
2.1.4. Tamaño efectivo y número efectivo de generaciones	47
2.2. Distribuciones extremas de las variables y_i en nuestros modelos	60
3. Variante del modelo de urna de Polya-Eggenberger	61

ÍNDICE GENERAL

3.1. El modelo de Polya-Eggenberger	61
3.1.1. Convergencia con $c = 1$	62
3.2. Variante del modelo de Polya-Eggenberger	62
4. Aplicaciones	73
4.1. Evolución por deriva en plagas	73
4.1.1. Deriva	74
4.1.2. Mutación	81
4.2. Inactivación del cromosoma X: Lyonización	92
4.2.1. Patrón de inactivación del cromosoma X	93
4.2.2. Causas del sesgo en el patrón de inactivación del cromosoma X	94
4.3. Procesos de deriva aplicados al X	96
4.3.1. Deriva cromosómica en la renovación celular	96
4.3.2. Deriva en el crecimiento por expansión clonal	107
4.3.3. Modelo de deriva de Polya-Eggenberger aplicado a la expansión clonal	107
4.3.4. Evolución del sesgo a partir de la inactivación en el blastocisto	110
4.4. Enfermedades ligadas al cromosoma X	114
4.4.1. Hemofilia	114
4.4.2. Modelo de deriva y correlación entre tejidos	116
Conclusión	119
Apéndice	121
Bibliografía	124

Capítulo 1

Deriva Genética

La genética de poblaciones comienza a desarrollarse en la primera mitad del siglo XX. Estudia el destino de la variación genética que surge por mutación en las poblaciones biológicas, como también las fuerzas evolutivas que afectan esa variación. En este capítulo estudiaremos la deriva genética, un proceso evolutivo independiente de los agentes de selección, que se debe al tamaño finito de la población y actúa cambiando aleatoriamente las frecuencias alélicas de las especies en el tiempo ([21] y [15]). Sewall Wright (1930) [38], genetista, y Ronald Fisher (1930) [13], estadístico y biólogo, principales fundadores de la genética de poblaciones, desarrollaron un modelo probabilístico de deriva genética, que luego fue generalizado por Chris Cannings (1973) [3], matemático y estadístico.

Para describir los modelos de genética poblacional repasemos primero los conceptos de gen y alelo. La información genética de cada célula (los genes) está contenida en segmentos de ADN llamados cromosomas, y el lugar en el cromosoma que ocupa un gen se llama *locus*. Un alelo es una variante de cómo se expresa un gen. Hay tantos alelos como posibles expresiones haya de un gen. Por ejemplo, para un *locus* determinado se tiene el gen que establece el color de pelo de un individuo. Existen variaciones respecto al color de pelo (rubio, negro, castaño). En los organismos haploides, que se multiplican por reproducción asexual, cada célula tiene un juego de cromosomas y por lo tanto un solo alelo está presente para determinar cada característica del

CAPÍTULO 1. DERIVA GENÉTICA

individuo. En los organismos diploides, que tienen reproducción sexual, hay dos juegos de cromosomas en cada célula, uno de herencia materna y otro de herencia paterna, excepto los gametos (óvulos y espermatozoides), que sólo tienen un juego (células haploides).

El planteo básico de la deriva genética consiste en considerar una población de individuos con dos variantes alélicas, A y a , de un único gen. Un alelo puede aumentar o reducir su frecuencia en la población sin que para ello sea necesario la selección natural. Estas variaciones se deben a razones puramente aleatorias, como consecuencia del muestreo en un conjunto finito que tiene lugar en la transmisión de genes de una población biológica a su descendencia. Este fenómeno se denomina deriva genética. A largo plazo la deriva produce la eliminación de la variación, por fijación de alguno de los alelos presentes en la población original. Los efectos de la deriva suelen ser mayores en poblaciones pequeñas como consecuencia de una reducción de su tamaño a causa de, por ejemplo, epidemias, cambios climáticos o aislamiento de poblaciones, afectando la variabilidad genética de la población original. A la restricción de tamaño por la que puede atravesar una población se la denomina cuello de botella, mientras que si un grupo pequeño se separa de la población principal para formar una nueva colonia, el fenómeno se denomina efecto fundador.

En los modelos de deriva que desarrollaremos en este capítulo consideraremos poblaciones de tamaño constante y con individuos abstractos de dos tipos, A y a . En organismos haploides, un individuo biológico corresponde a un sólo alelo. Los organismos diploides en cambio, pueden ser AA , Aa , aA o aa pero suponiendo cruzamientos al azar, los mismos modelos (como el de Wright-Fisher y sus generalizaciones y variantes) se pueden aplicar para analizar las frecuencias alélicas de A y a tanto a poblaciones de organismos haploides como diploides. También se pueden utilizar los modelos, no para alelos sino para poblaciones de células de dos tipos, como lo haremos en el Capítulo 4.

1.1. Modelo de Wright-Fisher

El modelo de Wright-Fisher considera originalmente una población de organismos diploides de tamaño finito $N/2$ con reproducción sexual. Estudia el cambio por puro azar a lo largo de las generaciones de las frecuencias alélicas de un gen con dos variantes, A y a , dentro de los N alelos presentes. La población es panmíctica, lo cual significa que los apareamientos son al azar. Los organismos diploides producen gran cantidad de gametos, que se combinan al azar para formar cigotos. El número de gametos es mucho mayor que N pero tiene las mismas frecuencias alélicas que la generación de la que provienen. Entonces muestrear sin reposición $N/2$ cigotos de la población de cigotos resultante es equivalente a muestrear al azar, con reposición, N de los alelos presentes en la generación precedente. En el caso haploide, el mismo modelo de Wright-Fisher se aplica a una población de N individuos de tipo A o a , que en cada generación deja una descendencia que es una muestra con reposición de la población precedente. Al ser equivalente el escenario diploide de $N/2$ individuos al haploide de N individuos, consideramos sólo este último en la siguiente subsección.

Presentaremos el modelo de Wright-Fisher y desarrollaremos los principales resultados que se obtienen del estudio de dicho modelo de deriva genética ([33] y [9]). Estos resultados serán retomados en el Capítulo 2 para ser aplicados a uno de nuestros modelos de deriva, que como veremos es una generalización del modelo de Wright-Fisher.

1.1.1. Cadena de Markov absorbente y martingala

En el modelo de Wright-Fisher se considera una población de tamaño constante, N , a lo largo de las generaciones, con dos variantes alélicas, A y a . La generación $n + 1$ es el resultado de un muestreo con reposición de tamaño N de la generación n . Por lo tanto, la probabilidad p_{ij} de que haya j alelos de tipo A en la generación $n + 1$ dados i alelos de tipo A en la generación n

sigue la ley binomial:

$$p_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}. \quad (1.1.1)$$

Si se define X_n como el número de alelos de tipo A en la generación n , entonces $\{X_n, n = 0, 1, \dots\}$ es una cadena de Markov homogénea con matriz de transición $P = \{p_{ij}\}$ y espacio de estados $S = \{0, 1, \dots, N\}$.

A partir de (1.1.1) se tiene que

$$X_{n+1} \mid X_n = i \sim Bi(N, i/N) \quad (1.1.2)$$

lo que implica que

$$E(X_{n+1} \mid X_n = i) = N \frac{i}{N} = i, \quad (1.1.3)$$

y de donde se obtiene

$$E(X_{n+1} \mid X_n) = X_n. \quad (1.1.4)$$

El número de alelos A podría pasar a ser 0 indicando la pérdida de dicho alelo de la población, o N , indicando la pérdida del alelo a de la población. Cuando uno de los dos tipos de alelos se pierde de la población, nunca retorna, es por eso que 0 y N son estados absorbentes. Además es fácil ver que 0 y N son accesibles (en un paso) desde cualquier otro estado. Luego, la probabilidad de que alguno de los dos alelos se fije en la población es 1 (ver apéndice). Sin embargo, el valor esperado de X_n es constante a lo largo de las generaciones, es decir,

$$E(X_{n+1}) = E(X_0) \quad (1.1.5)$$

para cada $n \geq 0$. En efecto, teniendo en cuenta propiedades de la esperanza, si se toma esperanza en (1.1.4) se obtiene que $E(X_{n+1}) = E(X_n)$, de donde por inducción resulta (1.1.5).

En este capítulo suponemos que el número de alelos A en la generación cero, X_0 , no es aleatorio, y por lo tanto se tiene que $E(X_0) = X_0$ como también que la varianza de X_0 es cero, $Var(X_0) = 0$.

Por otro lado veremos que

$$E_i X_n = i, \quad (1.1.6)$$

donde $E_i X_n = E(X_n | X_0 = i)$. Es decir, el número esperado de alelos A en la generación n , es igual al número inicial de alelos A en la población. Esto puede interpretarse como que en muchas realizaciones del experimento, algunas poblaciones tenderán al estado 0 y otras al estado N , pero en proporciones que mantendrán un promedio igual a i , lo cual es un fenómeno característico de los procesos de deriva.

Para demostrar (1.1.6) basta ver que

$$E_i E(X_{n+1} | X_n) = E_i X_{n+1}. \quad (1.1.7)$$

Primero probemos (1.1.6) a partir de (1.1.7) y luego probemos (1.1.7).

Como

$$E_i E(X_{n+1} | X_n) = E_i X_n \quad (1.1.8)$$

por (1.1.4), si igualamos las expresiones (1.1.7) y (1.1.8) se tiene $E_i X_{n+1} = E_i X_n$, de donde por inducción se deduce $E_i X_n = E_i X_0 = i$, que es el resultado (1.1.6).

Para completar la prueba nos falta mostrar que $E_i E(X_{n+1} | X_n) = E_i X_{n+1}$. Desarrollando las esperanzas y utilizando P_i para denotar probabilidades condicionadas al evento $\{X_0 = i\}$ se tiene:

$$\begin{aligned} E_i E(X_{n+1} | X_n) &= \sum_{x_n} E(X_{n+1} | X_n = x_n) P_i(X_n = x_n) \\ &= \sum_{x_n} \sum_{x_{n+1}} x_{n+1} P(X_{n+1} = x_{n+1} | X_n = x_n) P_i(X_n = x_n) \quad (1.1.9) \\ &= \sum_{x_n} \sum_{x_{n+1}} x_{n+1} P(X_{n+1} = x_{n+1} | X_n = x_n, X_0 = i) P_i(X_n = x_n) \end{aligned}$$

$$\begin{aligned} &= \sum_{x_{n+1}} x_{n+1} \sum_{x_n} P_i(X_{n+1} = x_{n+1} | X_n = x_n) P_i(X_n = x_n) \\ &= \sum_{x_{n+1}} x_{n+1} \sum_{x_n} P_i(X_{n+1} = x_{n+1}, X_n = x_n) \end{aligned} \quad (1.1.10)$$

$$\begin{aligned}
 &= \sum_{x_{n+1}} x_{n+1} P_i(X_{n+1} = x_{n+1}) \\
 &= E_i X_{n+1}
 \end{aligned}$$

y obtenemos (1.1.7) como se quería probar. Las expresiones (1.1.9) y (1.1.10) son iguales porque el proceso en $n+1$ condicional a n es independiente de $n-1, \dots, 0$, es decir, es markoviano, por lo tanto, agregar $X_0 = i$ condicionando en $P(X_{n+1} = x_{n+1} | X_n = x_n)$ no cambia la probabilidad de que X_{n+1} valga x_{n+1} .

El proceso $\{X_n\}_{n \geq 0}$ es martingala, es decir, cumple:

$$E(|X_n|) < \infty \quad (1.1.11)$$

$$E(X_{n+1} | X_n, \dots, X_0) = X_n. \quad (1.1.12)$$

En efecto,

$$0 \leq X_n \leq N$$

y

$$E(X_{n+1} | X_n = x_n, \dots, X_0 = x_0) = \sum_{x_{n+1}} x_{n+1} P(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) \quad (1.1.13)$$

$$= \sum_{x_{n+1}} x_{n+1} P(X_{n+1} = x_{n+1} | X_n = x_n) \quad (1.1.14)$$

$$= E(X_{n+1} | X_n = x_n) \quad (1.1.15)$$

$$= x_n. \quad (1.1.16)$$

Las expresiones (1.1.13) y (1.1.14) son iguales por la propiedad de Markov, y (1.1.15) y (1.1.16) lo son por (1.1.3).

Cálculo de la probabilidad de fijación del alelo A

Si bien los resultados presentados hasta ahora indican que en promedio el número de alelos A en la población se mantiene constante a lo largo de las

generaciones, $\{X_n\}_{n \geq 0}$ es una cadena de Markov absorbente, y por lo tanto, es de interés el cálculo de la probabilidad de fijación para cada uno de los dos alelos. Para poder realizar dicho cálculo es necesario definir

$$\tau = \min\{n : X_n = 0 \text{ ó } X_n = N\}$$

como el tiempo de fijación, que corresponde al primer instante en que la población consiste de todos alelos a o de todos alelos A.

En el modelo de Wright-Fisher la probabilidad de fijación para el alelo A es igual a la proporción de alelos A en la población inicial (y análogamente para los a). A continuación vemos el correspondiente resultado y la demostración que da Durrett [9], en la que completamos el desarrollo de algunos aspectos de la deducción.

Teorema 1.1.1.

$$P(X_\tau = N \mid X_0 = i) = i/N. \quad (1.1.17)$$

Demostración. *Para demostrarlo es suficiente ver que*

$$E_i X_\tau = i \quad (1.1.18)$$

donde $E_i X_\tau = E(X_\tau \mid X_0 = i)$. Veamos cómo de este último resultado se deduce lo que queremos probar.

A partir de la definición del tiempo de fijación se tiene que

$$X_\tau \mid X_0 = i \in \{0, N\},$$

y por lo tanto,

$$E_i X_\tau = 0 \cdot P_i(X_\tau = 0) + N \cdot P_i(X_\tau = N).$$

Luego despejando $P_i(X_\tau = N)$ de esta última igualdad y considerando (1.1.18) se obtiene $P_i(X_\tau = N) = i/N$.

Para terminar la prueba demostraremos (1.1.18):

De acuerdo a la definición de τ tenemos que $X_n = X_\tau$ para $n \geq \tau$,

entonces considerando la linealidad de la esperanza se tiene que

$$\begin{aligned} E_i X_n &= E_i(X_n, \tau \leq n) + E_i(X_n, \tau > n) \\ &= \int_{\Omega} X_{\tau} I_{\{\tau \leq n\}} dP_i + \int_{\Omega} X_n I_{\{\tau > n\}} dP_i, \end{aligned}$$

de donde se obtiene $E_i X_{\tau} = i$ haciendo tender n a infinito. En efecto, veamos que $\lim_{n \rightarrow \infty} \int_{\Omega} X_{\tau} I_{\{\tau \leq n\}} dP_i = E_i X_{\tau}$ y $\lim_{n \rightarrow \infty} \int_{\Omega} X_n I_{\{\tau > n\}} dP_i = 0$. Tenemos que, $X_{\tau} I_{\{\tau \leq n\}}$ es una sucesión creciente tal que $\lim_{n \rightarrow \infty} X_{\tau} I_{\{\tau \leq n\}} = X_{\tau}$, luego por el teorema de convergencia monótona:

$$\lim_{n \rightarrow \infty} \int_{\Omega} X_{\tau} I_{\{\tau \leq n\}} dP_i = \int_{\Omega} X_{\tau} dP_i = E_i X_{\tau}. \quad (1.1.19)$$

Por otro lado, $\lim_{n \rightarrow \infty} X_n I_{\{\tau > n\}} = 0$ y $|X_n| \leq N$ para todo n , luego por el teorema de convergencia dominada:

$$\lim_{n \rightarrow \infty} \int_{\Omega} X_n I_{\{\tau > n\}} dP_i = 0. \quad (1.1.20)$$

Entonces

$$\lim_{n \rightarrow \infty} E_i X_n = E_i X_{\tau} \quad (1.1.21)$$

combinando (1.1.21) con el hecho de que $E_i X_n = i$, según (1.1.6), se tiene que

$$E_i X_{\tau} = i \quad (1.1.22)$$

como se quería probar.

Cálculo de la varianza del número de alelos A en la población

Para tener una idea de la velocidad de convergencia del proceso a la fijación de un alelo, calcularemos la varianza de X_n . El cálculo de la varianza permite el cálculo de la heterocigosidad, que estudiaremos en la Subsección 1.1.2, y será de utilidad para obtener los tamaños efectivos en nuestros modelos, los cuales desarrollaremos en el Capítulo 2.

La varianza de X_n viene dada por la siguiente fórmula

$$\text{Var}(X_n) = X_0(N - X_0)(1 - \lambda^n) \quad (1.1.23)$$

donde $\lambda = 1 - \frac{1}{N}$ [33].

Para demostrarlo consideremos la propiedad de la varianza, según la cual, para dos variables X e Y se tiene que

$$\text{Var}(X) = E(\text{Var}(X|Y)) + \text{Var}(E(X|Y)).$$

En este caso usaremos

$$\text{Var}(X_n) = E(\text{Var}(X_n|X_{n-1})) + \text{Var}(E(X_n|X_{n-1})) \quad (1.1.24)$$

y calcularemos $E(\text{Var}(X_n|X_{n-1}))$ y $\text{Var}(E(X_n|X_{n-1}))$. Se tiene por (1.1.2) (distribución condicional binomial) que

$$\text{Var}(X_n|X_{n-1}) = N \frac{X_{n-1}}{N} \left(1 - \frac{X_{n-1}}{N}\right). \quad (1.1.25)$$

Si se aplica esperanza en (1.1.25) se obtiene

$$\begin{aligned} E(\text{Var}(X_n|X_{n-1})) &= E\left(N \frac{X_{n-1}}{N} \left(1 - \frac{X_{n-1}}{N}\right)\right) \\ &= E(X_{n-1}) - \frac{1}{N} E(X_{n-1}^2) \end{aligned} \quad (1.1.26)$$

y por (1.1.4) se tiene que

$$\text{Var}(E(X_n|X_{n-1})) = \text{Var}(X_{n-1}). \quad (1.1.27)$$

Luego reemplazando (1.1.26) y (1.1.27) en (1.1.24) se tiene

$$\begin{aligned} \text{Var}(X_n) &= E(X_{n-1}) - \frac{1}{N} E(X_{n-1}^2) + \text{Var}(X_{n-1}) \\ &= E(X_{n-1}) - \frac{1}{N} \left(\text{Var}(X_{n-1}) + E(X_{n-1})^2\right) + \text{Var}(X_{n-1}) \\ &= \text{Var}(X_{n-1}) \left(1 - \frac{1}{N}\right) + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \quad \text{por (1.1.5)} \end{aligned}$$

$$= Var(X_{n-1})\lambda + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \quad (1.1.28)$$

donde $\lambda = \left(1 - \frac{1}{N}\right)$.

Ahora consideremos la fórmula recursiva (1.1.28) para calcular la $Var(X_n)$:

$$\begin{aligned} Var(X_n) &= \left(Var(X_{n-2})\lambda + E(X_0) \left(1 - \frac{E(X_0)}{N}\right)\right)\lambda + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \\ &= Var(X_{n-2})\lambda^2 + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \sum_{k=0}^1 \lambda^k \\ &= \left(Var(X_{n-3})\lambda + E(X_0) \left(1 - \frac{E(X_0)}{N}\right)\right)\lambda^2 + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \sum_{k=0}^1 \lambda^k \\ &= Var(X_{n-3})\lambda^3 + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \sum_{k=0}^2 \lambda^k \\ &= \dots \\ &= Var(X_0)\lambda^n + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \sum_{k=0}^{n-1} \lambda^k \end{aligned} \quad (1.1.29)$$

$$= Var(X_0)\lambda^n + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \frac{1 - \lambda^n}{1 - \lambda} \quad (1.1.30)$$

$$= Var(X_0)\lambda^n + E(X_0) \left(1 - \frac{E(X_0)}{N}\right) \frac{1 - \lambda^n}{1/N} \quad (1.1.31)$$

$$= Var(X_0)\lambda^n + E(X_0)(N - E(X_0))(1 - \lambda^n) \quad (1.1.32)$$

y se obtiene (1.1.23) como queríamos probar.

Las expresiones (1.1.29) y (1.1.30) son iguales reemplazando $\sum_{k=0}^{n-1} \lambda^k$ por $\frac{1-\lambda^n}{1-\lambda}$. Las expresiones (1.1.31) y (1.1.32) son iguales reemplazando $E(X_0)$ por X_0 y $Var(X_0)$ por 0.

1.1.2. Heterocigosidad

Para estudiar la tasa a la cual la diversidad genética se pierde por deriva se define la heterocigosidad, $h(n)$, como la probabilidad de que dos alelos

tomados al azar sin reemplazo en la generación n sean de distinto tipo. Halle-
mos la fórmula para $h(n)$, la cual describe cómo cambia el sesgo hacia uno de
los dos tipos de alelos luego de n generaciones. Para hallar dicha probabilidad
se considera el siguiente experimento: se toman dos alelos (sin reemplazo) de
la generación n y se define el siguiente evento

$$A_n = \text{los alelos son de distinto tipo.}$$

Consideremos

$$I_{A_n} = \begin{cases} 1 & \text{si ocurre } A_n \\ 0 & \text{si no ocurre } A_n \end{cases}$$

entonces

$$E(I_{A_n}|X_n) = P(A_n|X_n).$$

Ahora como

$$P(A_n|X_n) = 2 \frac{X_n}{N} \frac{N - X_n}{N - 1}$$

resulta

$$E(I_{A_n}|X_n) = 2 \frac{X_n}{N} \frac{N - X_n}{N - 1}.$$

Si tomamos esperanza en esta última igualdad obtenemos la probabilidad
que queríamos calcular, porque

$$h(n) = E(E(I_{A_n}|X_n)) = E(I_{A_n}) = P(A_n).$$

Entonces para calcular la heterocigosidad necesitamos hallar la esperanza
de $X_n(N - X_n)$. Comencemos calculando $E(X_n(N - X_n)|X_{n-1})$. Para ello
debemos considerar (1.1.4) y (1.1.25):

$$\begin{aligned} E(X_n(N - X_n)|X_{n-1}) &= NE(X_n|X_{n-1}) - E(X_n^2|X_{n-1}) \\ &= NX_{n-1} - (Var(X_n|X_{n-1}) + E^2(X_n|X_{n-1})) \quad (1.1.33) \\ &= NX_{n-1} - \left(X_{n-1} \left(1 - \frac{X_{n-1}}{N} \right) + X_{n-1}^2 \right) \\ &= NX_{n-1} \left(1 - \frac{X_{n-1}}{N} \right) - X_{n-1} \left(1 - \frac{X_{n-1}}{N} \right) \end{aligned}$$

$$= X_{n-1} (N - X_{n-1}) \left(1 - \frac{1}{N}\right). \quad (1.1.34)$$

Si tomamos esperanza en (1.1.34) obtenemos

$$\begin{aligned} E(X_n (N - X_n)) &= E(X_{n-1} (N - X_{n-1})) \left(1 - \frac{1}{N}\right) & (1.1.35) \\ &= E(X_{n-2} (N - X_{n-2})) \left(1 - \frac{1}{N}\right)^2 \\ &= \dots \\ &= E(X_0 (N - X_0)) \left(1 - \frac{1}{N}\right)^n \\ &= (NE(X_0) - E(X_0^2)) \left(1 - \frac{1}{N}\right)^n \\ &= (NE(X_0) - (Var(X_0) + E(X_0)^2)) \left(1 - \frac{1}{N}\right)^n \\ &= E(X_0) (N - E(X_0)) \left(1 - \frac{1}{N}\right)^n - Var(X_0) \left(1 - \frac{1}{N}\right)^n. \end{aligned} \quad (1.1.36)$$

Nuevamente si se considera que $E(X_0) = X_0$ y $Var(X_0) = 0$, resulta

$$h(n) = h(0) \left(1 - \frac{1}{N}\right)^n \quad (1.1.37)$$

donde $h(0) = 2 \frac{X_0}{N} \frac{N-X_0}{N-1}$ es la probabilidad condicional a X_0 de que dos alelos elegidos (sin reemplazo) aleatoriamente de la población inicial sean diferentes.

De esta última expresión se sigue que $h(n)$ disminuirá a una tasa de $1 - 1/N$, donde N es el tamaño poblacional. En el modelo de Wright-Fisher N coincide con el llamado tamaño poblacional efectivo [38], como se verá más adelante. Observemos que la tasa de pérdida de heterocigosidad por deriva será mayor en poblaciones pequeñas, mientras que la probabilidad de que un alelo se fije en la población disminuye a medida que el tamaño poblacional aumenta. En consecuencia, mediante la tasa a la cual disminuye la heterocigosidad en la población, $1 - 1/N$, es posible tener una idea acerca

de cuánto tarda la fijación en ocurrir. Dado que $(1 - x) \approx e^{-x}$ cuando x es pequeño, si N es lo suficientemente grande $h(n) \approx h(0)e^{-\frac{n}{N}}$, por lo tanto, la heterocigosidad tiende a cero exponencialmente cuando $\frac{n}{N} \rightarrow \infty$ [9].

Obtención de la tasa $1 - 1/N$ como un autovalor de la matriz de transición

Veremos que $\lambda = 1 - 1/N$ es un autovalor de la matriz de transición del proceso de Wright-Fisher. Considerando que la heterocigosidad, según (1.1.37), viene dada por $h(n) = h(0) \left(1 - \frac{1}{N}\right)^n$, λ es la tasa a la cual se pierde la heterocigosidad en la población, y por lo tanto, dicho autovalor da información acerca del tiempo medio de fijación [11].

A partir de (1.1.34) se tiene

$$E(X_n(N - X_n) | X_{n-1} = i) = i(N - i) \left(1 - \frac{1}{N}\right)$$

y por definición de esperanza condicional se tiene

$$E(X_n(N - X_n) | X_{n-1} = i) = \sum_{j=0}^N j(N - j) p_{ij},$$

donde $p_{ij} = P(X_n = j | X_{n-1} = i)$ es la probabilidad de pasar de i alelos de tipo A a j alelos de tipo A en la generación siguiente. Si igualamos ambas expresiones obtenemos

$$\sum_{j=0}^N j(N - j) p_{ij} = i(N - i) \left(1 - \frac{1}{N}\right),$$

que también se lo puede expresar en forma matricial de la siguiente manera

$$\begin{pmatrix} p_{00} & p_{01} & p_{02} & \dots & p_{0N} \\ p_{10} & p_{11} & p_{12} & \dots & p_{1N} \\ p_{20} & p_{21} & p_{22} & \dots & p_{2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{N0} & p_{N1} & p_{N2} & \dots & p_{NN} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \cdot (N - 1) \\ 2 \cdot (N - 2) \\ \vdots \\ 0 \end{pmatrix} = \left(1 - \frac{1}{N}\right) \begin{pmatrix} 0 \\ 1 \cdot (N - 1) \\ 2 \cdot (N - 2) \\ \vdots \\ 0 \end{pmatrix}.$$

Esto es $E(X_{n+1}(N - X_{n+1}) | X_n = i) = (Pv)_i$ donde P es la matriz de transición de la cadena de Markov.

Luego, hemos hallado un vector v tal que $Pv = (1 - \frac{1}{N})v$, con lo cual

$$\lambda = 1 - \frac{1}{N} \tag{1.1.38}$$

es un autovalor de P , como queríamos probar.

1.1.3. Coalescencia

La teoría de la coalescencia se basa en la genealogía de genes y estudia cómo se van uniendo los linajes hacia atrás en el tiempo hasta alcanzar un antepasado común. La unión de linajes se denomina coalescencia. Una aplicación de la teoría de coalescencia es la estimación de parámetros de los procesos evolutivos que dieron lugar a la variación genética, como el tamaño poblacional efectivo y el tiempo al ancestro en común más reciente (*tmrca*, por sus siglas en inglés). En el modelo de Wright-Fisher algunos individuos pueden no dejar descendencia en la generación siguiente lo que implica la pérdida de linajes genéticos. Es fácil ver que en el modelo de Wright-Fisher la probabilidad de que dos alelos provengan de la misma copia en la generación anterior es $1/N$. En efecto, es la probabilidad de tomar dos veces el mismo alelo de la generación anterior, para cualquiera de los N alelos, es decir, $\frac{1}{N} \frac{1}{N} N$. Calculemos ahora la probabilidad de un evento de coalescencia en la generación n (hacia atrás). Para que ocurra un evento de coalescencia de dos alelos elegidos al azar exactamente en la generación n , dichos alelos no tienen que haber coalescido en las generaciones precedentes. La probabilidad de que no coalezcan en las primeras $n - 1$ generaciones es $(1 - \frac{1}{N})^{n-1}$, mientras que la probabilidad de que coalezcan en la n -ésima generación es $\frac{1}{N}$. Luego la probabilidad de que dos alelos elegidos al azar coalezcan en la n -ésima generación es $(1 - \frac{1}{N})^{n-1} \frac{1}{N}$. Se sigue de este razonamiento que si se siguiera la ancestría de los alelos lo suficientemente hacia atrás, se hallaría que todos descienden del mismo ancestro [9].

1.1.4. Tamaño efectivo

Comprender el proceso de deriva y la pérdida de diversidad en la población requiere información acerca de cómo se distribuyen las probabilidades de dejar descendencia por parte de los individuos de la población. Si por ejemplo es muy alta la probabilidad de que unos pocos individuos sean los que se reproducen, la población se comporta para los fines de la deriva como si fuera pequeña. Se puede cuantificar un efecto de este tipo definiendo el tamaño poblacional efectivo. Para ello se toma como referencia para la comparación el proceso de Wright-Fisher, que modela la transmisión de alelos de dos tipos en una población panmíctica de tamaño N , en la que en cada generación todos los individuos aportan sus tipos alélicos para producir un pool de alelos del cual se mantiene sólo una fracción de tamaño N tomada al azar. Por este motivo se suele decir informalmente que el tamaño efectivo de un proceso en estudio es el número de individuos que produciría la deriva observada si todos ellos fueran capaces de reproducirse y lo hicieran aleatoriamente. Pero pasemos ahora a las precisiones necesarias.

El tamaño poblacional efectivo, que denotaremos por N_e , permite establecer una equivalencia entre el modelo de Wright-Fisher y otros modelos poblacionales. Dicha equivalencia se puede establecer, por ejemplo, respecto a la heterocigosidad y definir el tamaño efectivo como el número de individuos que debería tener una población que sigue el proceso de Wright-Fisher para que su heterocigosidad sea igual que la de la población en estudio, a igual heterocigosidad inicial [20]. En general, el tamaño efectivo de una población para un estadístico dado es el tamaño que debería tener una población que sigue el proceso de Wright-Fisher para tener el mismo valor para el estadístico que el modelo poblacional en consideración. Los estadísticos que se pueden utilizar para calcular el valor del N_e son: la varianza de X_{n+1} condicional a X_n (N_e^v), el mayor autovalor no unitario de la matriz de transición (N_e^a), el tiempo esperado de coalescencia (N_e^c) y la heterocigosidad (N_e^h) [9].

1.2. Procesos de Cannings

Una generalización del modelo de Wright-Fisher fue introducida por C. Cannings (1974)[4], quien desarrolló una familia de cadenas de Markov relevantes para el estudio de la composición genética de poblaciones haploides con tamaño fijo. Supondremos en esta sección que no hay mutación de un tipo de alelo a otro e introduciremos mutación recién en la aplicación que consideramos en la Sección 4.1.

En esta sección nos proponemos presentar los resultados que obtuvo Cannings más algunos que no encontramos en la bibliografía y necesitamos para nuestras aplicaciones, incluyendo las demostraciones de cada uno de ellos.

En el modelo de Cannings se considera una población de genes con dos variantes alélicas, A y a, de tamaño fijo N , que se reproducen a cada instante $n = 0, 1, 2, 3, \dots$. Cada individuo de la población puede sobrevivir o no a la siguiente generación, como también dejar descendencia. Para cada i ($1 \leq i \leq N$) se define por y_i al número de descendientes (incluyendo a él mismo si sobrevive) del i -ésimo individuo en la siguiente generación. Entonces $y = (y_1, y_2, \dots, y_N)$ es un vector aleatorio en el espacio $\{(a_1, a_2, \dots, a_N) \in \mathbb{Z}_{\geq 0}^N : a_1 + a_2 + \dots + a_N = N\}$ tal que su distribución es intercambiable, es decir, la distribución de (y_i, y_j, \dots, y_k) es independiente de (i, j, \dots, k) . Luego resulta que todos los individuos tienen la misma distribución de probabilidad para la descendencia y como el tamaño poblacional se mantiene constante se tiene

$$y_1 + y_2 + \dots + y_N = N. \quad (1.2.1)$$

La distribución de cada variable aleatoria tendrá media 1 porque por (1.2.1) se tiene que $E(\sum_{i=1}^N y_i) = N$ y por el hecho de que todas las variables tienen la misma distribución resulta que $E(\sum_{i=1}^N y_i) = \sum_{i=1}^N E(y_i) = NE(y_1)$ de donde se obtiene que para todo i ($1 \leq i \leq N$):

$$E(y_i) = 1. \quad (1.2.2)$$

Se denotará por σ^2 la varianza de cada variable y_i :

$$\sigma^2 = \text{Var}(y_i). \quad (1.2.3)$$

Siendo X_n el número de alelos A en la generación n , se tiene que $\{X_n\}_{n \geq 0}$ es una cadena de Markov con conjunto de posibles estados $\{0, 1, 2, \dots, N\}$ y matriz de transición $\{p_{ij}\}$ donde $p_{ij} = P(X_{n+1} = j \mid X_n = i)$.

En los modelos de Cannings si $P(y_i = 1) = 1$ para todo i , $p_{ii} = 1$ y la matriz de transición es la identidad, por lo cual estamos en un caso trivial en el que todos los estados son absorbentes. Si consideramos $P(y_i = 1) \neq 1$ para todo i entonces los estados 0 y N son los estados absorbentes de la cadena de Markov porque $p_{00} = p_{NN} = 1$. Excluyendo el caso trivial vamos a demostrar que los estados 0 y N son accesibles desde todo estado i distinto de 0 y N . En el apéndice recordamos algunas definiciones de cadenas de Markov con estados absorbentes, que se usan en la siguiente proposición.

Proposición 1. *Si $P(y_1 = 1) \neq 1$, entonces para todo i distinto de 0 y de N , tenemos que tanto 0 como N son accesibles desde i .*

Demostración. *Dado i distinto de 0 y de N mostraremos que existe $j < i$ tal que $p_{ij} = P(X_1 = j \mid X_0 = i) > 0$. Iterando este resultado se tiene que existe un número de pasos n tal que $P(X_n = 0 \mid X_0 = i) > 0$. La prueba de que N es accesible desde i es simétrica.*

Para $X_0 = i$ numeramos los alelos de tipo A desde 1 hasta i y los de tipo a desde $i+1$ hasta N , siendo y_1, \dots, y_N los respectivos números de descendientes que deja cada uno. Como $P(y_k = 1) \neq 1$ para todo k entre 1 y N , se tiene que hay N valores enteros no negativos z_1, \dots, z_N no todos iguales a 1 tales $P(y_1 = z_1, \dots, y_N = z_N) > 0$. Como las y_k son intercambiables, esto se cumple para cualquier permutación de valores z_1, \dots, z_N . Consideraremos aquella en la que $z_1 \leq z_2 \leq \dots \leq z_N$. Aquí observamos que necesariamente $z_1 = 0$ porque $\sum_{k=1}^N z_k = N$ y no todos los z_k son iguales a 1. Ahora, dado que $\sum_{k=1}^i z_k$ es el número de alelos A en la próxima generación, basta ver que $\sum_{k=1}^i z_k < i$. Tenemos dos casos. Si $z_i \leq 1$, como $z_1 = 0$, resulta $\sum_{k=1}^i z_k < \sum_{k=1}^i 1 = i$. Si $z_i > 1$, se tiene $\sum_{k=i+1}^N z_k > \sum_{k=i+1}^N 1 = N - i$, lo cual implica que $\sum_{k=1}^i z_k < i$ porque debe cumplirse que $N = \sum_{k=1}^N z_k = \sum_{k=1}^i z_k + \sum_{k=i+1}^N z_k$.

En virtud de esta proposición se puede demostrar que los estados 0 y N se alcanzan en tiempo finito con probabilidad 1 (ver Apéndice).

En los procesos de Cannings $E(X_{n+1} | X_n = i) = i$. En efecto, la distribución de $X_{n+1} | X_n = i$ es la de la suma de las i variables y_k que corresponden a alelos A. Agregando la intercambiabilidad y el hecho de que las y_k tiene media 1, se justifican las siguientes igualdades:

$$\begin{aligned} E(X_{n+1} | X_n = i) &= E(y_1 + y_2 + \cdots + y_i) \\ &= iE(y_1) \\ &= i, \end{aligned} \tag{1.2.4}$$

y por lo tanto

$$E(X_{n+1} | X_n) = X_n. \tag{1.2.5}$$

De esto último y el hecho de que el proceso es markoviano, con el mismo procedimiento que utilizamos para el modelo de Wright-Fisher, se obtiene que el valor medio de X_n se mantiene constante en el tiempo y que la cadena de Markov es martingala, como también que la probabilidad de fijación de alguno de los alelos es igual a su frecuencia inicial en la población.

1.2.1. La varianza en los procesos de Cannings

En esta subsección obtendremos la varianza del número de alelos A para el modelo de Cannings. Como para el modelo de Wright-Fisher, esta medida de dispersión es de interés para evaluar cómo interviene la deriva en la evolución de las frecuencias génicas a lo largo de las generaciones [11].

Cálculo de la varianza de X_{n+1} condicional a X_n

Calculemos la $Var(X_{n+1} | X_n = i)$ en términos de σ^2 . Nuevamente por el hecho de que las variables y_k son intercambiables se tiene

$$\begin{aligned} Var(X_{n+1} | X_n = i) &= Var(y_1 + y_2 + \cdots + y_i) \\ &= i\sigma^2 + i(i-1)Cov(y_1, y_2). \end{aligned} \tag{1.2.6}$$

Entonces resta calcular la $Cov(y_1, y_2)$ en términos de σ^2 . Considerando (1.2.1) se tiene

$$Var(y_1 + y_2 + \dots + y_N) = 0, \quad (1.2.7)$$

y por otro lado, como

$$Var(y_1 + y_2 + \dots + y_N) = N\sigma^2 + N(N-1)Cov(y_1, y_2) \quad (1.2.8)$$

si se iguala (1.2.7) y (1.2.8) y se despeja la covarianza se obtiene

$$Cov(y_1, y_2) = -\sigma^2/(N-1). \quad (1.2.9)$$

Luego combinando (1.2.6) y (1.2.9) se tiene que

$$\begin{aligned} Var(X_{n+1} | X_n = i) &= i\sigma^2 - i(i-1)\sigma^2/(N-1) \\ &= \frac{i\sigma^2}{(N-1)}(N-1 - (i-1)) \\ &= \sigma^2 \frac{i(N-i)}{(N-1)}, \end{aligned} \quad (1.2.10)$$

de donde resulta que

$$Var(X_{n+1} | X_n) = \sigma^2 \frac{X_n(N - X_n)}{(N-1)}. \quad (1.2.11)$$

Cálculo de la varianza de X_{n+1}

Ahora, la $Var(X_{n+1})$ se puede calcular como se hizo en la Sección 1.1:

$$\begin{aligned} Var(X_{n+1}) &= E(Var(X_{n+1} | X_n)) + Var(E(X_{n+1} | X_n)) \\ &= \frac{\sigma^2 N E(X_n)}{(N-1)} - \frac{\sigma^2 (Var(X_n) + E(X_n)^2)}{(N-1)} + Var(X_n) \\ &\hspace{15em} \text{por (1.2.11) y (1.2.5)} \\ &= \frac{\sigma^2 E(X_0)}{(N-1)} (N - E(X_0)) + \left(1 - \frac{\sigma^2}{(N-1)}\right) Var(X_n). \end{aligned}$$

A partir de la fórmula recursiva tenemos:

$$Var(X_{n+1}) = \frac{\sigma^2 E(X_0)}{(N-1)} (N - E(X_0)) + \left(1 - \frac{\sigma^2}{(N-1)}\right) Var(X_n) \quad (1.2.12)$$

$$\begin{aligned} &= \frac{\sigma^2 E(X_0)}{(N-1)} (N - E(X_0)) \\ &\quad + \left(1 - \frac{\sigma^2}{(N-1)}\right) \left(\frac{\sigma^2 E(X_0)}{(N-1)} (N - E(X_0)) + \left(1 - \frac{\sigma^2}{(N-1)}\right) Var(X_{n-1}) \right) \end{aligned}$$

$$= \frac{\sigma^2 E(X_0)}{(N-1)} (N - E(X_0)) \left(1 + \left(1 - \frac{\sigma^2}{(N-1)}\right)\right) + \left(1 - \frac{\sigma^2}{(N-1)}\right)^2 Var(X_{n-1})$$

= ...

$$= \frac{\sigma^2 E(X_0)}{(N-1)} (N - E(X_0)) \left(\sum_{k=0}^n \left(1 - \frac{\sigma^2}{(N-1)}\right)^k \right) + \left(1 - \frac{\sigma^2}{(N-1)}\right)^{n+1} Var(X_0)$$

$$= \frac{\sigma^2 E(X_0)}{(N-1)} (N - E(X_0)) \left(\frac{1 - \left(1 - \frac{\sigma^2}{(N-1)}\right)^{n+1}}{1 - \left(1 - \frac{\sigma^2}{(N-1)}\right)} \right) + \left(1 - \frac{\sigma^2}{(N-1)}\right)^{n+1} Var(X_0) \quad (1.2.13)$$

$$= E(X_0) (N - E(X_0)) \left(1 - \left(1 - \frac{\sigma^2}{(N-1)}\right)^{n+1}\right) + \left(1 - \frac{\sigma^2}{(N-1)}\right)^{n+1} Var(X_0) \quad (1.2.14)$$

donde (1.2.13) se obtiene reemplazando $\sum_{k=0}^n \left(1 - \frac{\sigma^2}{(N-1)}\right)^k$ por $\frac{1 - \left(1 - \frac{\sigma^2}{(N-1)}\right)^{n+1}}{1 - \left(1 - \frac{\sigma^2}{(N-1)}\right)}$.

Reemplazando en (1.2.14) $E(X_0)$ por X_0 y $Var(X_0)$ por 0 se obtiene la siguiente expresión para la varianza en términos de σ^2 :

$$Var(X_{n+1}) = X_0 (N - X_0) \left(1 - \left(1 - \frac{\sigma^2}{(N-1)}\right)^{n+1}\right). \quad (1.2.15)$$

1.2.2. Autovalores de la matriz de transición

Cannings calculó los autovalores de la matriz de transición, P , de la cadena de Markov $\{X_n\}_{n \geq 0}$ [4]. El mayor autovalor distinto de 1 es importante en general para obtener la velocidad de convergencia de procesos de Markov. En nuestro estudio veremos su rol en distintas medidas de disminución de la diversidad. A continuación detallamos resultados generales sobre los autovalores de los procesos de Cannings y desarrollamos la demostración de los

mismos.

Teorema 1.2.1. *Si para $i, j = 0, 1, \dots, N$ definimos*

$$p_{ij} = P(X_{n+1} = j \mid X_n = i)$$

entonces los autovalores de la matriz de transición $P = \{p_{ij}\}$ son

$$\begin{aligned}\lambda_0 &= 1 \\ \lambda_j &= E(y_1 y_2 \dots y_j)\end{aligned}$$

Demostración. *Vamos a demostrar que existe una matriz $A = \{a_{ij}\}$, triangular superior, que tiene los mismos autovalores que P y éstos son $\lambda_0 = 1$ y $\lambda_j = E(y_1 y_2 \dots y_j)$ para todo j tal que $1 \leq j \leq N$.*

Para ello veremos que existen Z una matriz no singular y A una matriz triangular superior tales que

$$PZ = ZA$$

o sea

$$P = ZAZ^{-1}.$$

A partir de esta ecuación se tiene que $\det(\lambda I - P) = \det(Z(\lambda I - A)Z^{-1}) = \det(Z) \det(\lambda I - A) \det(Z^{-1}) = \det(\lambda I - A)$, por lo cual los autovalores de P coinciden con los autovalores de A . Además como A es una matriz triangular superior se tiene que los autovalores de A son los elementos de su diagonal.

Para probar la existencia de Z y A se define $Z_{kj} = k^j$ con $0^0 = 1$, es decir:

$$Z = \begin{pmatrix} 0^0 & 0^1 & \dots & 0^j & \dots & 0^N \\ 1^0 & 1^1 & \dots & 1^j & \dots & 1^N \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ k^0 & k^1 & \dots & k^j & \dots & k^N \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ N^0 & N^1 & \dots & N^j & \dots & N^N \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ 1 & 1 & \dots & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & k & \dots & k^j & \dots & k^N \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & N & \dots & N^j & \dots & N^N \end{pmatrix}.$$

Esta matriz Z es la matriz de Vandermonde y por lo tanto no singular

CAPÍTULO 1. DERIVA GENÉTICA

(ver apéndice).

Con esta elección de Z se tiene que

$$(PZ)_{ij} = \sum_{k=0}^N p_{ik} k^j = E(X_{n+1}^j | X_n = i)$$

y

$$(ZA)_{ij} = \sum_{k=0}^N i^k a_{kj}.$$

Luego para que $PZ = ZA$ se satisfaga, $(PZ)_{ij}$ y $(ZA)_{ij}$ tienen que ser iguales para cada par i, j , o sea

$$\sum_{k=0}^N p_{ik} k^j = \sum_{k=0}^N i^k a_{kj}.$$

Veamos columna por columna la existencia de la matriz A , triangular superior y con los elementos diagonales deseados.

Para $j = 0$ y todo i ($0 \leq i \leq N$) tenemos $\sum_{k=0}^N p_{ik} k^j = \sum_{k=0}^N p_{ik} k^0 = \sum_{k=0}^N p_{ik} = 1$. Ahora, la sumatoria $\sum_{k=0}^N i^k a_{k0} = i^0 a_{00} + i^1 a_{10} + \dots + i^N a_{N0}$ es igual a 1 para todo i si tomamos $a_{00} = 1$ y para $k \geq 1$, $a_{k0} = 0$.

Para $j \geq 1$ obtendremos una expresión polinómica para $\sum_{k=0}^N p_{ik} k^j$ que igualada con el polinomio $\sum_{k=0}^N i^k a_{kj}$ nos mostrará la existencia de valores a_{kj} para los cuales se verifica la igualdad y donde los a_{jj} son los elementos diagonales de A deseados. Para eso partimos de $\sum_{k=0}^N p_{ik} k^j = E(X_{n+1}^j | X_n = i) = E(y_1 + \dots + y_i)^j$.

Si $j \geq 1$ y $1 \leq i \leq N$, por un desarrollo del multinomio alternativo al del Teorema Multinomial y utilizando la intercambiabilidad al tomar la esperanza podemos expresar $E(y_1 + \dots + y_i)^j = E[(y_1 + \dots + y_i) \dots (y_1 + \dots + y_i)]$ mediante

$$\sum_{k=1}^j (i)_k \left[\sum_{h_1 + \dots + h_k = j, h_1 \geq \dots \geq h_k > 0} e_{h_1, h_2, \dots, h_k} E(y_1^{h_1} y_2^{h_2} \dots y_k^{h_k}) \right].$$

Aquí $(i)_k = i(i-1) \dots (i-k+1)$ es la cantidad de maneras de tomar k de

los subíndices de $1, \dots, i$ en orden, valor que resulta 0 si $i < k$. A su vez, para k subíndices ordenados, e_{h_1, h_2, \dots, h_k} es la cantidad de esperanzas de productos de las variables y_i tales que los subíndices aparecen por primera vez (de izquierda a derecha) en el orden dado, y tales que todas estas esperanzas son iguales a $E\left(y_1^{h_1} y_2^{h_2} \dots y_k^{h_k}\right)$, donde $h_1 + \dots + h_k = j$ y $h_1 \geq \dots \geq h_k > 0$. Si $i = 0$ también se cumple la igualdad

$$\sum_{k=0}^N p_{ik} k^j = \sum_{k=1}^j (i)_k \left[\sum_{h_1 + \dots + h_k = j, h_1 \geq \dots \geq h_k > 0} e_{h_1, h_2, \dots, h_k} E\left(y_1^{h_1} y_2^{h_2} \dots y_k^{h_k}\right) \right]$$

porque ambos miembros son iguales a cero.

Para ilustrar el desarrollo del multinomio, veamos el caso $j = 4$ y cualquier i entre 1 y N :

$$\begin{aligned} E(y_1 + \dots + y_i)^4 &= E[(y_1 + \dots + y_i)(y_1 + \dots + y_i)(y_1 + \dots + y_i)(y_1 + \dots + y_i)] \\ &= (i)_1 E(y_1^4) + (i)_2 [4E(y_1^3 y_2) + 3E(y_1^2 y_2^2)] + (i)_3 6E(y_1^2 y_2 y_3) \\ &\quad + (i)_4 E(y_1 y_2 y_3 y_4). \end{aligned}$$

Para más detalle, expliquemos por ejemplo el coeficiente $6 = e_{2,1,1}$ que multiplica a $E(y_1^2 y_2 y_3)$. Los tres subíndices en orden ya están elegidos de $(i)_3$ maneras y van a ser reemplazados por los subíndices 1, 2 y 3 por la intercambiabilidad. Debemos contar entonces de cuántas maneras se puede conseguir el producto de las tres variables y_1, y_2 e y_3 (en orden para su primera aparición de izquierda a derecha), donde una de ellas se la toma dos veces por estar elevada al cuadrado. Las maneras son seis: $y_1 y_1 y_2 y_3$, $y_1 y_2 y_1 y_3$, $y_1 y_2 y_3 y_1$, $y_1 y_2 y_2 y_3$, $y_1 y_2 y_3 y_2$ y $y_1 y_2 y_3 y_3$, cuyas esperanzas son todas iguales a $E(y_1^2 y_2 y_3)$ por la intercambiabilidad. Cabe señalar que no necesitaremos calcular los coeficientes e_{h_1, h_2, \dots, h_k} en general, salvo el de $E(y_1 y_2 \dots y_j)$ que es uno. Lo importante es mostrar el desarrollo alternativo del multinomio, que dado $j \geq 1$, da la misma expresión para cualquier i entre 1 y N .

Ahora, separando el último sumando de la expresión obtenida para $\sum_{k=0}^N p_{ik} k^j$

tenemos

$$\sum_{k=1}^{j-1} (i)_k \left[\sum_{h_1+\dots+h_k=j, h_1 \geq \dots \geq h_k > 0} e_{h_1, h_2, \dots, h_k} E \left(y_1^{h_1} y_2^{h_2} \dots y_k^{h_k} \right) \right] \\ + (i)_j \sum_{h_1+\dots+h_j=j, h_1 \geq \dots \geq h_j > 0} e_{h_1, h_2, \dots, h_j} E \left(y_1^{h_1} y_2^{h_2} \dots y_j^{h_j} \right)$$

donde $\sum_{h_1+\dots+h_j=j, h_1 \geq \dots \geq h_j > 0} e_{h_1, h_2, \dots, h_j} E \left(y_1^{h_1} y_2^{h_2} \dots y_j^{h_j} \right)$ se reduce a $E(y_1 y_2 \dots y_j)$. Es decir,

$$\sum_{k=0}^N p_{ik} k^j = \sum_{k=1}^{j-1} (i)_k \left[\sum_{h_1+\dots+h_k=j, h_1 \geq \dots \geq h_k > 0} e_{h_1, h_2, \dots, h_k} E \left(y_1^{h_1} y_2^{h_2} \dots y_k^{h_k} \right) \right] \\ + (i)_j E(y_1 y_2 \dots y_j).$$

Como $(i)_j = i(i-1)\dots(i-j+1)$, distribuyendo este producto vemos que $\sum_{k=0}^N p_{ik} k^j$ es un polinomio de grado j en i igual a $i^j E(y_1 y_2 \dots y_j)$ más términos de grado menor a j cuyos coeficientes involucran sumas de términos $e_{h_1, h_2, \dots, h_k} E \left(y_1^{h_1} y_2^{h_2} \dots y_k^{h_k} \right)$.

Luego, la condición $\sum_{k=0}^N p_{ik} k^j = \sum_{k=0}^N i^k a_{kj}$ termina siendo una igualdad de polinomios en i , y se cumple para cualesquiera $j \geq 1$ e i ($0 \leq i \leq N$) si para $k \leq j$, a_{kj} es igual al coeficiente de i^k en la expresión polinómica de grado j en i de $\sum_{k=0}^N p_{ik} k^j$, y si $k > j$, $a_{kj} = 0$. Con esta condición nos queda que $a_{jj} = E(y_1 y_2 \dots y_j)$ y que A es triangular superior, como queríamos.

Propiedades del conjunto de autovalores

El siguiente teorema agrega propiedades al conjunto de autovalores. Aborda el caso trivial en que todos los individuos tienen exactamente un descendiente y aparte de esto, establece un orden en el espectro de autovalores, que va de 1 para λ_0 y λ_1 hasta cero para tantos autovalores como individuos que no dejen descendencia.

Teorema 1.2.2. *Los autovalores dados en el Teorema 1.2.1 cumplen que $\lambda_j = 1$ para cada $0 \leq j \leq N$ si y sólo si $P(y_l = 1) = 1$ para cada $1 \leq l \leq N$,*

de lo contrario

$$1 = \lambda_0 = \lambda_1 > \lambda_2 > \lambda_3 > \dots > \lambda_k = \lambda_{k+1} = \lambda_{k+2} = \dots = \lambda_N = 0$$

si y sólo si $N - k + 1$ del conjunto de N individuos cumplen que $y_l = 0$ en cada generación, es decir, debe haber exactamente $N - k + 1$ individuos que mueren y no se reproducen.

Demostración. Por el Teorema 1.2.1 sabemos que $\lambda_0 = 1$ y que $\lambda_1 = E(y_1)$. Luego, como demostramos en (1.2.2) que $E(y_1) = 1$, resulta $\lambda_1 = 1$.

Si $P(y_l = 1) = 1$ para cada $1 \leq l \leq N$ se tiene que $y_1 \dots y_j = 1$ para cada $1 \leq j \leq N$, entonces como $\lambda_j = E(y_1 \dots y_j)$, según el Teorema 1.2.1, resulta que $\lambda_j = 1$ para $1 \leq j \leq N$. Veamos que el resultado recíproco también es verdadero.

Asumimos que $\lambda_j = 1$ para todo j tal que $1 \leq j \leq N$. Supongamos que existe l entre 1 y N tal que $P(y_l = 1)$ no es uno, entonces $P(y_1 = 1, y_2 = 1, \dots, y_N = 1)$ es distinta de uno. Como $\lambda_N = E(y_1 y_2 \dots y_N) = P(y_1 = 1, y_2 = 1, \dots, y_N = 1)$ (porque $\sum_{j=1}^N y_j = N$ y las y_i sólo pueden tomar valores enteros no negativos) esto implicaría que λ_N es distinto de uno, absurdo.

Ahora supongamos que exactamente $N - k + 1$ individuos en cada generación cumplen que $y_l = 0$, entonces cada k -upla de variables y_l seleccionada de las variables y_1, \dots, y_N debe tener alguna coordenada nula, por lo tanto $E(y_1 \dots y_m) = 0$ para todo $m \geq k$, y entonces $\lambda_k = \dots = \lambda_N = 0$. Veamos que el resultado recíproco también es verdadero.

Supongamos que $\lambda_k = \dots = \lambda_N = 0$, entonces para todo $m \geq k$ se tiene que $E(y_1 \dots y_m) = 0$, lo cual implica que con probabilidad uno tenemos $y_1 \dots y_m = 0$ y por lo tanto cada m -upla de variables y_l seleccionada de las variables y_1, \dots, y_N debe tener alguna coordenada nula, que es equivalente a decir que exactamente $N - k + 1$ individuos en cada generación cumplen que $y_l = 0$.

Ahora veamos las desigualdades entre autovalores. Si se considera λ_j siendo $j < k$, existe al menos una j -upla de variables y_l con todas sus coordenadas distintas de cero, entonces $\lambda_j = E(y_1 \dots y_j) > 0$. Probemos que $\lambda_{j-1} > \lambda_j$

para $j < k$.

Si reemplazamos y_j por $N - y_1 - \dots - y_{j-1} - y_{j+1} - \dots - y_N$ en $\lambda_j = E(y_1 \dots y_j)$ se tiene

$$\lambda_j = E(y_1 \dots y_{j-1} (N - y_1 - \dots - y_{j-1} - y_{j+1} - \dots - y_N)),$$

que se puede reescribir de la siguiente manera

$$\lambda_j = N\lambda_{j-1} - (j-1)E(y_1 \dots y_{j-2}y_{j-1}^2) - (N-j)\lambda_j$$

y reagrupando se obtiene

$$\lambda_j = (N\lambda_{j-1} - (j-1)E(y_1 \dots y_{j-2}y_{j-1}^2)) / (N-j+1). \quad (1.2.16)$$

Es fácil ver a partir de (1.2.16) que la igualdad entre λ_j y λ_{j-1} ocurre si y sólo si $E(y_1 \dots y_{j-2}y_{j-1}^2) = \lambda_{j-1}$. Como también que $\lambda_{j-1} > \lambda_j$ si y sólo si $E(y_1 \dots y_{j-2}y_{j-1}^2) > \lambda_{j-1}$. En esta última desigualdad si agrupamos todos los términos del lado izquierdo y reemplazamos a λ_{j-1} por $E(y_1 \dots y_{j-2}y_{j-1})$ obtenemos que $E(y_1 \dots y_{j-2}y_{j-1}(y_{j-1} - 1)) > 0$. Finalmente se tiene que $\lambda_{j-1} > \lambda_j$ si y sólo si $E(y_1 \dots y_{j-2}y_{j-1}(y_{j-1} - 1)) > 0$. Esto último siempre se cumple dada la intercambiabilidad de las y_i y que existe al menos una $(j-1)$ -upla de y_i con todas sus coordenadas distintas de cero y que además alguna de las variables debe ser mayor que uno puesto que $\sum_{r=1}^{k-1} y_r = N$ y $N > k-1$.

Cálculo del autovalor λ_2

Ahora calculemos λ_2 , el autovalor no unitario más grande de la matriz de transición de la cadena de Markov según establece el Teorema 1.2.2, excluyendo el caso trivial en el que $P(y_i = 1) = 1$ para $1 \leq i \leq N$. El cálculo de dicho autovalor es importante porque da información acerca de la velocidad de convergencia hacia la fijación de alguno de los alelos.

Vamos a calcular λ_2 en función de σ^2 :

$$\lambda_2 = E(y_1 y_2) = Cov(y_1, y_2) + E(y_1)E(y_2),$$

y como $E(y_1) = E(y_2) = 1$ y la covarianza fue calculada en (1.2.9) si se reemplaza en esta última expresión se obtiene

$$\lambda_2 = 1 - \sigma^2/(N - 1). \quad (1.2.17)$$

Finalmente para hallar λ_2 podemos reemplazar σ^2 en (1.2.17).

El valor que tome λ_2 depende de cuál sea el proceso de Cannings que se esté considerando. En la Subsección 1.2.5 veremos en particular el proceso de Wright-Fisher.

1.2.3. Heterocigosidad

Calculemos la heterocigosidad de los procesos de Cannings, es decir, la probabilidad de que dos alelos elegidos aleatoriamente sin reemplazo en la generación n sean de distinto tipo, como se definió en la Sección 1.1.

Veremos que la heterocigosidad de los procesos de Cannings es

$$h(n) = h(0)\lambda_2^n, \quad (1.2.18)$$

donde $h(0) = 2 \frac{X_0}{N} \frac{N-X_0}{N-1}$ es la probabilidad condicional a X_0 de que dos alelos elegidos (sin reemplazo) aleatoriamente de la población inicial sean diferentes, y λ_2 es el autovalor no unitario más grande de la matriz de transición.

A partir de (1.2.18), como ya hemos estudiado en el modelo de Wright-Fisher, se puede ver que λ_2 se corresponde con la tasa a la cual se pierde la heterocigosidad en la población, es por eso que dicho autovalor es de interés a la hora de estudiar la velocidad de convergencia a los estados absorbentes.

La heterocigosidad, según vimos en la Sección 1.1, se obtiene hallando el valor de la esperanza de $X_n(N - X_n)$. Para ello comencemos calculando $E(X_n(N - X_n) | X_{n-1})$, que según (1.1.33) es:

$$E(X_n(N - X_n) | X_{n-1}) = NX_{n-1} - (Var(X_n | X_{n-1}) + E^2(X_n | X_{n-1})).$$

Primero expresemos la $Var(X_n | X_{n-1})$ en función de λ_2 para luego escribir a la heterocigosidad como una función de λ_2 .

Escribimos $\sigma^2 = (1 - \lambda_2)(N - 1)$, que se obtiene despejando σ^2 de (1.2.17). El siguiente paso consiste en reemplazar σ^2 por $(1 - \lambda_2)(N - 1)$ en $Var(X_n | X_{n-1}) = \sigma^2 \frac{X_{n-1}(N - X_{n-1})}{(N-1)}$, que fue calculada en (1.2.11), y obtenemos que

$$Var(X_n | X_{n-1}) = (1 - \lambda_2)X_{n-1}(N - X_{n-1}).$$

Luego reemplazando esta última expresión en (1.1.33) y considerando que $E(X_n | X_{n-1}) = X_{n-1}$ según (1.2.5) se tiene

$$\begin{aligned} E(X_n(N - X_n) | X_{n-1}) &= NX_{n-1} - (1 - \lambda_2)X_{n-1}(N - X_{n-1}) - X_{n-1}^2 \\ &= \lambda_2 X_{n-1}(N - X_{n-1}). \end{aligned} \quad (1.2.19)$$

Tomando esperanza en (1.2.19) obtenemos

$$\begin{aligned} E(X_n(N - X_n)) &= \lambda_2 E(X_{n-1}(N - X_{n-1})) \\ &= \lambda_2^2 E(X_{n-2}(N - X_{n-2})) \\ &= \dots \\ &= \lambda_2^n E(X_0(N - X_0)) \\ &= \lambda_2^n E(X_0)(N - E(X_0)) - \lambda_2^n Var(X_0). \end{aligned} \quad (1.2.20)$$

Finalmente dado que $E(X_0) = X_0$, $Var(X_0) = 0$ y considerando que la heterocigosidad viene dada por $h(n) = E\left(2\frac{X_n}{N}\frac{N-X_n}{N-1}\right)$, según se definió en la Subsección 1.1.2, se obtiene (1.2.18).

1.2.4. Tamaño efectivo

Calcularemos para el modelo de Cannings los tamaños efectivos con respecto a la varianza condicional, al mayor autovalor no unitario de la matriz de transición y a la heterocigosidad.

Recordando que el tamaño efectivo de una población para un estadístico dado es el tamaño que debería tener una población que sigue el proceso de Wright-Fisher para tener el mismo valor para el estadístico que el modelo poblacional en estudio, calculemos N_e^v , N_e^a y N_e^h , introducidos en la Subsección 1.1.4.

Para calcular

- N_e^v : tenemos que igualar las expresiones de la varianza condicional suponiendo que en la generación $n - 1$ la proporción de alelos de tipo A, p_{n-1} , es la misma en ambas poblaciones.
- N_e^a : tenemos que igualar las expresiones de λ_2 .
- N_e^h : tenemos que igualar las fórmulas de la heterocigosidad suponiendo que inicialmente la proporción de alelos A en ambas poblaciones es la misma.

Si consideramos el Cuadro 1.1, podemos observar en la tercera columna que los tamaños efectivos respecto a la varianza condicional, a λ_2 y a la heterocigosidad resultan iguales [10]:

$$N_e^v = N_e^a = N_e^h = (N - 1)/\sigma^2. \quad (1.2.21)$$

En general esto ocurre en modelos de genética poblacional de tiempo discreto con tamaño poblacional constante en los cuales el paso de una generación a la siguiente es a partir de un muestreo aleatorio [2].

De la fórmula (1.2.21) podemos observar que cuanto más grande es la dispersión del número de hijos que tiene cada individuo la población en términos efectivos será más pequeña, y por lo tanto la deriva será mayor que la esperada en función de su tamaño poblacional. Esto se debe a que hay más probabilidades de que unos tengan más hijos que otros.

Wright-Fisher	Cannings	Cálculo: Tamaño Efectivo
$Var(\frac{X_n}{N} X_{n-1}) = \frac{1}{N}p_{n-1}(1-p_{n-1})$	$Var(\frac{X_n}{N} X_{n-1}) = \frac{\sigma^2}{N-1}p_{n-1}(1-p_{n-1})$	$\frac{1}{N_e^v}p_{n-1}(1-p_{n-1}) = \frac{\sigma^2}{N-1}p_{n-1}(1-p_{n-1})$ Luego $\frac{1}{N_e^v} = \frac{\sigma^2}{N-1} \therefore N_e^v = \frac{N-1}{\sigma^2}$
$\lambda_2 = 1 - \frac{1}{N}$	$\lambda_2 = 1 - \frac{\sigma^2}{N-1}$	$1 - \frac{1}{N_e^a} = 1 - \frac{\sigma^2}{N-1}$ Luego $\frac{1}{N_e^a} = \frac{\sigma^2}{N-1} \therefore N_e^a = \frac{N-1}{\sigma^2}$
$h(n) = h(0)\lambda_2^n$	$h(n) = h(0)\lambda_2^n$	$h(0)\left(1 - \frac{1}{N_e^h}\right) = h(0)\left(1 - \frac{\sigma^2}{N-1}\right)$ Luego $\frac{1}{N_e^h} = \frac{\sigma^2}{N-1} \therefore N_e^h = \frac{N-1}{\sigma^2}$

Cuadro 1.1: En este cuadro se presentan las expresiones de la varianza condicional, de λ_2 y de la heterocigosidad del modelo de Wright-Fisher y del de Cannings para luego calcular los tamaños efectivos respecto de cada estadístico.

1.2.5. El modelo de Wright-Fisher como caso particular del modelo de Cannings

El modelo de Wright-Fisher es un caso particular del modelo de Cannings, en el que $y = (y_1, \dots, y_N)$ tiene una distribución multinomial simétrica, y por lo tanto, para cada $1 \leq i \leq N$,

$$y_i \sim Bi(N, 1/N)$$

con media 1 y varianza

$$\sigma^2 = N \frac{1}{N} \left(1 - \frac{1}{N} \right) = 1 - \frac{1}{N}.$$

Las variables y_i son intercambiables por ser las coordenadas de la multinomial simétrica. Luego, se tiene que $E(X_{n+1} | X_n = i) = i$ procediendo de la misma manera que se hizo en (1.2.4):

$$\begin{aligned} E(X_{n+1} | X_n = i) &= E(y_1 + y_2 + \dots + y_i) \\ &= iE(y_1) \\ &= i. \end{aligned}$$

Observemos que para calcular la varianza condicional, la varianza no condicional y λ_2 para el modelo de Wright-Fisher basta con reemplazar σ^2 por $1 - \frac{1}{N}$ en (1.2.11), en (1.2.15) y en (1.2.17), respectivamente. Luego se obtiene

$$\begin{aligned} Var(X_{n+1} | X_n) &= \frac{(N-1) X_n (N - X_n)}{N(N-1)} \\ &= X_n (1 - X_n/N), \end{aligned}$$

$$Var(X_{n+1}) = X_0 (N - X_0) (1 - \lambda^{n+1}) \quad \text{con } \lambda = 1 - \frac{1}{N}$$

y

$$\begin{aligned}\lambda_2 &= 1 - \frac{\frac{N-1}{N}}{(N-1)} \\ &= 1 - \frac{1}{N}\end{aligned}$$

el autovalor obtenido en (1.1.38) que coincide con la tasa a la cual disminuye la heterocigosidad en la población como se expresa en (1.1.37). En la Subsección 1.2.2, titulada *Autovalores de la matriz de transición*, se presentaron dos teoremas enunciados por C. Cannings (1974) [4] en los cuales se calcularon los autovalores de las matrices de transición de las cadenas de Markov para modelos poblacionales que generalizan el modelo de Wright-Fisher. A partir de dichos resultados se deduce que $\lambda = 1 - \frac{1}{N}$ es el autovalor no unitario más grande de la matriz de transición asociada al modelo de Wright-Fisher.

Para calcular la heterocigosidad tenemos que reemplazar λ_2 por $1 - \frac{1}{N}$ en (1.2.18) y obtenemos el mismo resultado de la Sección 1.1:

$$h(n) = h(0) \left(1 - \frac{1}{N}\right)^n.$$

A partir del Teorema 1.2.2 se tiene que los autovalores de la matriz de transición en el modelo de Wright-Fisher cumplen

$$1 = \lambda_0 = \lambda_1 > \lambda_2 > \lambda_3 > \dots > \lambda_N$$

porque no ocurre que todos los individuos dejan un descendiente con probabilidad 1 ni que hay individuos que con probabilidad 1 mueren sin reproducirse.

Feller (1951) [12] demostró que los autovalores de la matriz asociada al modelo de Wright-Fisher vienen dados por

$$\lambda_j = \frac{N!}{(N-j)!} N^{-j} \quad \forall 0 \leq j \leq N. \quad (1.2.22)$$

Son fáciles de verificar los autovalores que ya calculamos como $\lambda_0 = \lambda_1 = 1$ y $\lambda_2 = 1 - 1/N$.

CAPÍTULO 1. DERIVA GENÉTICA

Capítulo 2

Generalización y variante del modelo de Wright-Fisher

Propondremos y estudiaremos dos modelos de deriva genética que se caracterizarán por tener renovación parcial de la población y superposición de generaciones, con el objetivo de utilizarlos en distintas aplicaciones biológicas. Dichas aplicaciones serán abordadas en el Capítulo 4. Como veremos en este capítulo, se pueden estudiar con los métodos propuestos por Cannings ([3] y [4]), que hemos desarrollado en el Capítulo 1, Sección 1.2.

2.1. Expansión determinística y expansión aleatoria

El objetivo de esta sección será estudiar dos modelos de deriva genética en los cuales la generación $n + 1$ se obtiene de la generación n a partir de un muestreo aleatorio, por el cual algunos individuos serán removidos, otros permanecerán, mientras que otros se reproducirán para restaurar el tamaño original de la población. La renovación en las sucesivas generaciones es parcial, a diferencia del modelo de Wright-Fisher en el que la renovación es total. Esto último, para ciertos valores de parámetros, da lugar a superposición de generaciones.

En particular estudiaremos dos modelos de deriva que se diferencian en

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

cómo se genera la descendencia de los individuos elegidos en cada generación para reproducirse. En uno de los modelos la expansión que restaura el tamaño original es determinística mientras que en el otro es aleatoria. El modelo de renovación parcial con expansión aleatoria es una generalización del modelo de Wright-Fisher de genética poblacional, y refleja el azar que puede existir en una población en el proceso de restitución de su tamaño original. En cambio el de expansión determinística corresponde a un crecimiento más rígido en el que hay azar para establecer qué individuos van a reproducirse pero luego estos arrojan el mismo número de descendientes cada uno, pudiendo representar el crecimiento por expansión clonal en ciertos tejidos con células de dos tipos.

Comencemos describiendo el muestreo involucrado en el paso de una generación a la siguiente para ambos modelos. Inicialmente se tiene una población de individuos haploides de tamaño N con dos variantes alélicas, A y a . Por lo tanto escribiremos individuo y alelo de manera indistinta. El muestreo aleatorio que tiene lugar en la generación n para construir la generación $n + 1$ es el resultado de cuatro etapas que aquí detallamos. Cabe aclarar que siempre que nos referimos a elección aleatoria de individuos, entendemos que todos los individuos tienen la misma probabilidad de ser incluidos en el grupo elegido. Más aún, las variables aleatorias que dan el número de descendientes de cada uno de los N individuos serán intercambiables.

1. Se eliminan $r < N$ alelos de la generación n , elegidos aleatoriamente y sin reemplazo.
2. De los $N - r$ restantes se eligen s ($0 < s \leq N - r$) alelos aleatoriamente y sin reemplazo, a partir de los cuales se va a restaurar el tamaño original de la población. Llamaremos Y_n a la variable aleatoria que representa el número de alelos de tipo A entre los s .
3. En esta etapa se produce la renovación parcial, que consiste en la expansión de los s alelos elegidos al azar para reproducirse.
 - **Expansión aleatoria:** Se produce un muestreo con reemplazo de tamaño $r + s$ de los s alelos producto de la segunda etapa, con probabilidad Y_n/s de elegir un alelo de tipo A . Es decir, el número

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

de alelos A entre los $r + s$ nuevos individuos tiene distribución binomial con parámetros $r + s$ y Y_n/s , por lo cual llamamos a este muestreo binomial. Entonces se renuevan en cada generación $r + s$ alelos y $N - r - s$ quedan como estaban. Siempre que $N - r - s > 0$ habrá superposición de generaciones.

- **Expansión determinística:** Los s alelos se reproducen por clonación, es decir, el número de alelos de tipo A entre los r nuevos es proporcional al número de alelos de tipo A entre los s . Esto es, la cantidad de alelos de tipo A de los r alelos producto de esta etapa será rY_n/s . Suponemos que r es múltiplo de s para asegurar que el número de alelos se mantiene entero, pero más adelante abandonamos este supuesto para extender el uso de las fórmulas con valores continuos. En este proceso los s alelos que dejan descendencia permanecen en la población y se renuevan r alelos a cada paso. También tenemos aquí que si $N - r - s > 0$ habrá superposición de generaciones salvo en el caso en el que $r = 0$, que descartaremos por trivial en nuestros cálculos correspondientes a la expansión determinística.

4. Finalmente la generación $n + 1$ se forma con los $r + s$ y los r alelos, según corresponda, producto de la tercera etapa, y los $N - r - s$ y $N - r$ alelos que restan, respectivamente.

En ambos modelos de renovación parcial, a diferencia del de Wright-Fisher, puede haber superposición de generaciones.

En las Figuras 2.1 y 2.2 se representa el muestreo involucrado en el paso de la generación n a la generación $n + 1$ para el modelo con restauración de la población a partir de una expansión aleatoria y de una expansión determinística respectivamente. En cada caso la urna con bolillas de dos colores distintos representa una población de individuos de dos tipos de tamaño constante $N = 16$. En ambos casos consideramos $s = 3$ y $r = 9$. Los alelos de tipo A están representados por las bolillas blancas y los de tipo a por las bolillas negras. Nuevamente se define por X_n el número de alelos de tipo A (o bolillas blancas) en la generación n .

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

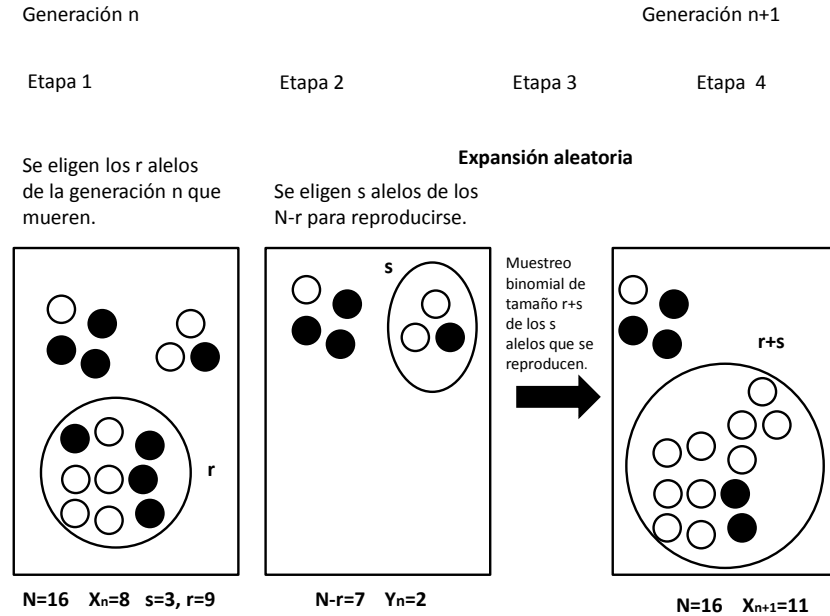


Figura 2.1: En la generación n hay $X_n = 8$ alelos de tipo A y de los $s = 3$ alelos elegidos para reproducirse dos tercios son alelos de tipo A. Luego se realiza un muestreo con reemplazo de tamaño $r + s = 12$ de los $s = 3$ alelos que se reproducen. Este ejemplo representa uno de los posibles resultados del muestreo, en el cual se sacaron dos veces bolillas negras y 10 veces bolillas blancas. El número de alelos A se incrementó en 3 en el paso de la generación n a la $n + 1$, es decir, $X_{n+1} = 11$.

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

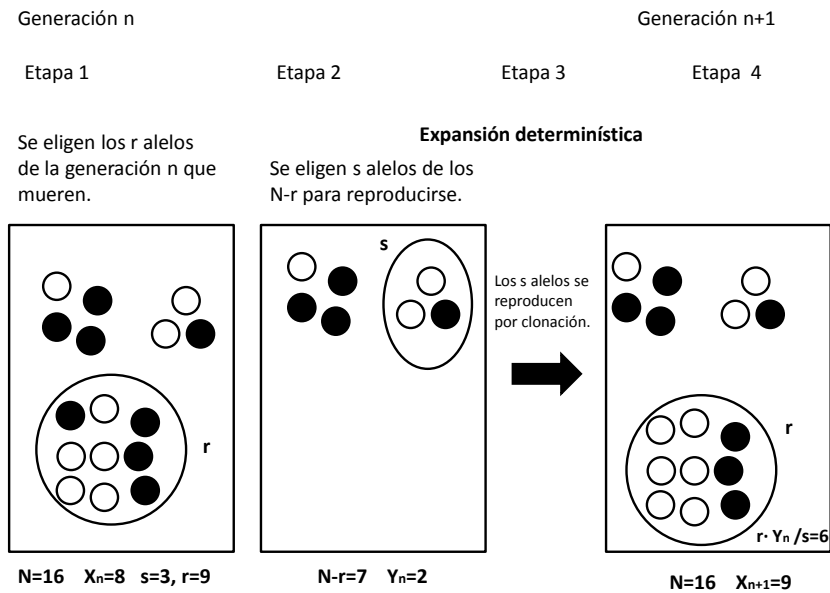


Figura 2.2: En la generación n hay $X_n = 8$ alelos de tipo A y de los $s = 3$ alelos elegidos para reproducirse dos tercios son alelos de tipo A. Estos alelos se expanden clonalmente y dejan $r = 9$ descendientes, de los cuales dos tercios son alelos de tipo A como consecuencia de la expansión determinística. En este ejemplo el número de alelos A se incrementó en 1 en el paso de la generación n a la $n + 1$, es decir, $X_{n+1} = 9$.

2.1.1. El número de descendientes por individuo

Analizaremos nuestros modelos con las herramientas desarrolladas en el Capítulo 1, Sección 1.2. Siendo y_i el número de descendientes del i -ésimo individuo, se cumple que $y_1 + y_2 + \dots + y_N = N$. Aquí los descendientes y_i son el número total de individuos de su mismo tipo que deja el individuo i en un paso del proceso. Por ejemplo $y_i = 1$ se aplica tanto al caso de un individuo que pertenece al grupo de los $N - r - s$ que sobrevive pero no da origen a un nuevo individuo, como al caso de uno del grupo de los s en la expansión aleatoria, el cual es elegido una vez para reproducirse pero él mismo muere. Se cumple que todos los individuos tienen la misma distribución de probabilidad para la descendencia y la distribución de (y_i, y_j, \dots, y_k) es independiente de

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

(i, j, \dots, k) porque todos los individuos están en las mismas condiciones para ser elegidos para reproducirse, morir o permanecer en cada generación. Es decir, las variables y_i son intercambiables y estamos en las condiciones del modelo general estudiado en la Sección 1.2.

A partir de lo desarrollado en el párrafo anterior se tiene que en ambos modelos las variables y_i tienen media 1 porque $E(\sum_{i=1}^N y_i) = N$ y como todas las variables tienen la misma distribución resulta que $E(\sum_{i=1}^N y_i) = \sum_{i=1}^N E(y_i) = NE(y_1)$ de donde se deduce que para todo i ($1 \leq i \leq N$):

$$E(y_i) = 1. \quad (2.1.1)$$

Calculemos la distribución de probabilidad de y_i para el modelo en el que la expansión que restaura el tamaño original de la población es determinística y para el modelo en el que la expansión es aleatoria. Para ello es necesario tener en cuenta que en el muestreo aleatorio involucrado en el paso de una generación a la siguiente cada individuo puede morir, sobrevivir o dejar descendencia. En función de esto último y considerando la fórmula de la probabilidad total se tiene que

$$\begin{aligned} P(y_i = j) &= P(y_i = j \mid \text{Pertenece a los } N-r-s)P(\text{Pertenece a los } N-r-s) \\ &\quad + P(y_i = j \mid \text{Pertenece a los } r)P(\text{Pertenece a los } r) \\ &\quad + P(y_i = j \mid \text{Pertenece a los } s)P(\text{Pertenece a los } s), \end{aligned} \quad (2.1.2)$$

donde el grupo de los $N - r - s$ es el de los individuos que sobreviven en el evento de renovación. El grupo de los r es el de aquellos que mueren y el grupo de los s es el de los que pueden dejar descendencia.

Renovación parcial de la población con expansión aleatoria

Veremos que la distribución de probabilidad de y_i ($1 \leq i \leq N$) en el modelo en el que el tamaño de la población se restaura a partir de una expansión aleatoria es:

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

$$P(y_i = j) = \begin{cases} \frac{r}{N} + \left(1 - \frac{1}{s}\right)^{r+s} \frac{s}{N} & \text{si } j = 0 \\ \frac{N-r-s}{N} + \binom{r+s}{1} \frac{1}{s} \left(1 - \frac{1}{s}\right)^{r+s-1} \frac{s}{N} & \text{si } j = 1 \\ \binom{r+s}{j} \frac{1}{s}^j \left(1 - \frac{1}{s}\right)^{r+s-j} \frac{s}{N} & \text{si } j > 1. \end{cases} \quad (2.1.3)$$

Es necesario tener en cuenta que la renovación parcial en este modelo consiste en un muestreo con reposición de tamaño $r+s$ de los s alelos elegidos al azar para reproducirse. Por lo tanto,

$$y_i \mid \text{Pertenece a los } s \sim Bi\left(r+s, \frac{1}{s}\right). \quad (2.1.4)$$

Observemos que un individuo a pesar de pertenecer al grupo de los s podría no dejar descendientes, es decir, podría no ser elegido en el muestreo de tamaño $r+s$ de los s .

Ahora calculemos $P(y_i = j)$.

Si $j = 0$, entonces el i -ésimo individuo pertenece al grupo de los r , lo cual ocurre con probabilidad r/N o bien pertenece al grupo de los s , lo cual ocurre con probabilidad s/N , pero no deja ningún descendiente. Por (2.1.4) resulta que $P(y_i = 0 \mid \text{Pertenece a los } s) = \binom{r+s}{0} \left(\frac{1}{s}\right)^0 \left(1 - \frac{1}{s}\right)^{r+s} = \left(1 - \frac{1}{s}\right)^{r+s}$. Además $P(y_i = 0 \mid \text{Pertenece a los } N-r-s) = 0$. Reemplazando en (2.1.2) tenemos que

$$\begin{aligned} P(y_i = 0) &= 0 \frac{N-r-s}{N} + 1 \frac{r}{N} + \left(1 - \frac{1}{s}\right)^{r+s} \frac{s}{N} \\ &= \frac{r}{N} + \left(1 - \frac{1}{s}\right)^{r+s} \frac{s}{N}. \end{aligned} \quad (2.1.5)$$

Si $j = 1$ resulta $P(y_i = 1 \mid \text{Pertenece a los } r) = 0$ porque tenemos sólo dos posibilidades para el i -ésimo individuo, o permanece en la población, es decir, pertenece a los $N-r-s$, evento que tiene probabilidad $(N-r-s)/N$ o deja exactamente un descendiente, es decir, pertenece al grupo de los s , lo cual ocurre con probabilidad s/N . Como $P(y_i = 1 \mid \text{Pertenece a los } s) =$

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

$\binom{r+s}{1} \frac{1}{s} \left(1 - \frac{1}{s}\right)^{r+s-1}$ por (2.1.4), reemplazando en (2.1.2) tenemos que

$$\begin{aligned} P(y_i = 1) &= 1 \frac{N-r-s}{N} + 0 \frac{r}{N} + \binom{r+s}{1} \frac{1}{s} \left(1 - \frac{1}{s}\right)^{r+s-1} \frac{s}{N} \\ &= \frac{N-r-s}{N} + \binom{r+s}{1} \frac{1}{s} \left(1 - \frac{1}{s}\right)^{r+s-1} \frac{s}{N}. \end{aligned} \quad (2.1.6)$$

Finalmente, si $j > 1$,

$$P(y_i = j \mid \text{Pertenece a los } s) = \binom{r+s}{j} \frac{1^j}{s} \left(1 - \frac{1}{s}\right)^{r+s-j}.$$

Mientras que $P(\text{Pertenece a los } s) = s/N$, $P(y_i = j \mid \text{Pertenece a los } r) = 0$ y $P(y_i = j \mid \text{Pertenece a los } N-r-s) = 0$. Luego reemplazando en (2.1.2) tenemos que para $j > 1$

$$\begin{aligned} P(y_i = j) &= 0 \frac{N-r-s}{N} + 0 \frac{r}{N} + \binom{r+s}{j} \frac{1^j}{s} \left(1 - \frac{1}{s}\right)^{r+s-j} \frac{s}{N} \\ &= \binom{r+s}{j} \frac{1^j}{s} \left(1 - \frac{1}{s}\right)^{r+s-j} \frac{s}{N}. \end{aligned} \quad (2.1.7)$$

Ahora calculemos la varianza de y_i , que denotamos por σ^2 como en la Sección 1.2:

Proposición 2. *Si la renovación parcial de la población es con expansión aleatoria se tiene que la varianza de y_i ($1 \leq i \leq N$) es*

$$\sigma^2 = \frac{(r+s)(r+s-1)}{sN}. \quad (2.1.8)$$

Demostración. *Para calcular la $\text{Var}(y_i)$ basta con calcular $E(y_i^2)$ porque ya vimos que $E(y_i) = 1$ según (2.1.1). Considerando la distribución de probabilidad de y_i dada por (2.1.3) se puede calcular su segundo momento:*

$$\begin{aligned} E(y_i^2) &= \sum_{j=0}^{r+s} j^2 P(y_i = j) \\ &= P(y_i = 1) + \sum_{j=2}^{r+s} j^2 P(y_i = j) \end{aligned}$$

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

$$= \frac{N - r - s}{N} + \frac{s}{N} \sum_{j=1}^{r+s} j^2 \binom{r+s}{j} (1/s)^j (1 - 1/s)^{r+s-j} \quad (2.1.9)$$

$$= \frac{N - r - s}{N} + \frac{s}{N} ((r+s)(1/s)(1 - 1/s) + ((r+s)(1/s))^2) \quad (2.1.10)$$

$$= 1 - \frac{r+s}{N} + \frac{s}{N} \left(\frac{r+s}{s} - \frac{r+s}{s^2} + \frac{(r+s)^2}{s^2} \right)$$

$$= 1 - \frac{r+s}{N} + \frac{r+s}{N} - \frac{r+s}{sN} + \frac{(r+s)^2}{sN}$$

$$= 1 + \frac{(r+s)(r+s-1)}{sN}.$$

Los segundos términos de las expresiones (2.1.9) y (2.1.10) son iguales si se considera el segundo momento de una binomial de parámetros $r+s$ y $1/s$. Luego se obtiene

$$\begin{aligned} \text{Var}(y_i) &= E(y_i^2) - E^2(y_i) \\ &= 1 + \frac{(r+s)(r+s-1)}{sN} - 1 \\ &= \frac{(r+s)(r+s-1)}{sN} \end{aligned}$$

como queríamos probar.

El modelo de Wright-Fisher corresponde a $s = N$ y $r = 0$. En dicho caso y_i tiene una distribución binomial con parámetros N y $1/N$ como vimos en el Capítulo 1, Subsección 1.2.5.

Renovación parcial de la población con expansión determinística

Resta hallar la distribución de probabilidad de y_i para el modelo en el que el tamaño poblacional se restaura a partir de una expansión determinística, en la cual todos los individuos candidatos a reproducirse dejan igual número de descendientes.

Veamos que la distribución de probabilidad de y_i ($1 \leq i \leq N$) es:

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

$$P(y_i = j) = \begin{cases} \frac{r}{N} & \text{si } j = 0 \\ \frac{N-r-s}{N} & \text{si } j = 1 \\ \frac{s}{N} & \text{si } j = r/s + 1. \end{cases} \quad (2.1.11)$$

El individuo i pertenece al grupo de los r si y sólo si $j = 0$. Luego $P(y_i = 0 \mid \text{Pertenece a los } r) = 1$, $P(y_i = 0 \mid \text{Pertenece a los } N - r - s) = 0$ y $P(y_i = 0 \mid \text{Pertenece a los } s) = 0$. Reemplazando en (2.1.2) tenemos que

$$\begin{aligned} P(y_i = 0) &= 0 \frac{N-r-s}{N} + 1 \frac{r}{N} + 0 \frac{s}{N} \\ &= \frac{r}{N}. \end{aligned} \quad (2.1.12)$$

El individuo i pertenece al grupo de los $N - r - s$ si y sólo si $j = 1$. Luego $P(y_i = 1 \mid \text{Pertenece a los } N - r - s) = 1$, $P(y_i = 1 \mid \text{Pertenece a los } r) = 0$ y $P(y_i = 1 \mid \text{Pertenece a los } s) = 0$. Reemplazando en (2.1.2) tenemos que

$$\begin{aligned} P(y_i = 1) &= 1 \frac{N-r-s}{N} + 0 \frac{r}{N} + 0 \frac{s}{N} \\ &= \frac{N-r-s}{N}. \end{aligned} \quad (2.1.13)$$

Finalmente el individuo i pertenece al grupo de los s si y sólo si $j = r/s + 1$. Luego $P(y_i = j \mid \text{Pertenece a los } s) = 1$, $P(y_i = j \mid \text{Pertenece a los } r) = 0$ y $P(y_i = j \mid \text{Pertenece a los } N - r - s) = 0$. Reemplazando en (2.1.2) tenemos que

$$\begin{aligned} P(y_i = \frac{r}{s} + 1) &= 0 \frac{N-r-s}{N} + 0 \frac{r}{N} + 1 \frac{s}{N} \\ &= \frac{s}{N}. \end{aligned} \quad (2.1.14)$$

Calculemos la varianza de las y_i , que también denotamos por σ^2 :

Proposición 3. *Si la renovación parcial de la población es con expansión determinística se tiene que la varianza de y_i ($1 \leq i \leq N$) es*

$$\sigma^2 = \frac{(r+s)r}{sN} \quad (2.1.15)$$

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

Demostración. Para calcular la $Var(y_i)$ basta con calcular $E(y_i^2)$ porque ya vimos que $E(y_i) = 1$ según (2.1.1). Considerando la distribución de probabilidad de y_i dada por (2.1.11) se pueden calcular su segundo momento:

$$\begin{aligned}
 E(y_i^2) &= \frac{N - r - s}{N} + \left(\frac{r}{s} + 1\right)^2 \frac{s}{N} \\
 &= 1 - \frac{r + s}{N} + \frac{(r + s)^2}{sN} \\
 &= 1 + \frac{r + s}{N} \left(\frac{(r + s)}{s} - 1\right) \\
 &= 1 + \frac{(r + s)r}{sN}.
 \end{aligned}$$

Luego

$$\begin{aligned}
 Var(y_1) &= E(y_1^2) - E^2(y_1) \\
 &= 1 + \frac{(r + s)r}{sN} - 1 \\
 &= \frac{(r + s)r}{sN}
 \end{aligned} \tag{2.1.16}$$

como queríamos probar.

En ambos modelos $\{X_n\}_{n \geq 0}$ es una cadena de Markov con conjunto de estados $\{0, 1, 2, \dots, N\}$. El número de alelos A podría pasar a ser 0 indicando la pérdida de dicho alelo de la población, o N , indicando la pérdida del alelo a de la población. Cuando uno de los dos tipos de alelos se pierde de la población, nunca retorna, es por eso que 0 y N son estados absorbentes. Luego, como consecuencia de lo que probamos en el Capítulo 1, $\{X_n\}_{n \geq 0}$ es una cadena de Markov absorbente y la probabilidad de que algún alelo se fije en la población es 1.

En ambos modelos se cumple que $E(X_{n+1} | X_n = i) = i$. En efecto, como las variables y_i tienen la misma distribución y media 1 resulta que

$$\begin{aligned}
 E(X_{n+1} | X_n = i) &= E(y_1 + y_2 + \dots + y_i) \\
 &= iE(y_1) \\
 &= i,
 \end{aligned}$$

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

luego

$$E(X_{n+1} | X_n) = X_n. \quad (2.1.17)$$

A partir de este último resultado y la propiedad de Markov, como ya hemos visto para el modelo de Wright-Fisher en la Sección 1.1, se demuestra por inducción utilizando $E(X_{n+1} | X_n) = X_n$ que la esperanza de X_n se mantiene constante en el tiempo. También se puede ver que el proceso es martingala. Además se puede probar de la misma manera que se hizo en el Capítulo 1, Subsección 1.1.1 que la probabilidad de fijación para cada uno de los alelos es igual a su frecuencia inicial en la población, es decir,

$$P(X_\tau = N | X_0 = i) = i/N.$$

2.1.2. Autovalores

El cálculo del mayor autovalor no unitario, λ_2 , de la matriz de transición P de la cadena de Markov $\{X_n\}_{n \geq 0}$ será importante para estudiar la velocidad de convergencia de cada proceso a los estados absorbentes. A partir de λ_2 podemos calcular la probabilidad de que dos individuos tomados al azar sean de distinto tipo en cada generación y el tamaño efectivo respecto a dicho autovalor (ver modelo de Cannings en el Capítulo 1, Sección 1.2) para luego analizar cómo interviene la deriva en el cambio de la frecuencia de uno u otro alelo en la población a lo largo de las generaciones.

Calculemos los autovalores λ_1 y λ_2 para cada uno de los modelos. El Teorema 1.2.1 en el Capítulo 1, Sección 1.2 establece que $\lambda_j = E(y_1 y_2 \dots y_j)$ para cada $1 \leq j \leq N$.

Así tenemos que $\lambda_1 = 1$ porque, como vimos, $E(y_1) = 1$ tanto para el modelo con expansión aleatoria como para el modelo con expansión determinística. Para calcular λ_2 en ambos modelos tenemos que considerar la expresión para dicho autovalor (en términos de σ^2) que se obtuvo en la Sección 1.2:

$$\lambda_2 = 1 - \sigma^2 / (N - 1) \quad (1.2.17).$$

Luego reemplazando σ^2 por $\frac{(r+s)(r+s-1)}{sN}$ si la expansión es aleatoria o por $\frac{(r+s)r}{sN}$

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

si la expansión es determinística, se obtiene que si la expansión es aleatoria

$$\lambda_2 = 1 - \frac{(r+s)(r+s-1)}{s(N-1)N}, \quad (2.1.18)$$

y si la expansión es determinística

$$\lambda_2 = 1 - \frac{r(r+s)}{s(N-1)N}. \quad (2.1.19)$$

Veremos en la siguiente subsección que el cálculo de λ_2 es importante para estudiar la tasa a la cual se pierde la diversidad en la población.

2.1.3. Deriva genética y heterocigosidad

El muestreo al azar involucrado en la formación de cada generación posibilita que la composición genética de la población cambie a lo largo del tiempo. Pretendemos estudiar a qué velocidad se pierde la heterocigosidad en la población por deriva y para ello calcularemos la varianza y la heterocigosidad en ambos modelos desarrollados.

Primero calculemos la varianza de X_{n+1} condicional a X_n y la varianza (no condicional) de X_{n+1} para cada modelo. Es de interés el cálculo de la sucesión de varianzas en ambos procesos porque cuanto más pequeña sea la varianza, menos dispersión tendrá la distribución del número de individuos de cada tipo, y por lo tanto la fijación en un estado absorbente será más lenta. Al igual que lo hicimos en el Capítulo 1 suponemos que el número de alelos A en la generación cero, X_0 , no es aleatorio, y por lo tanto se tiene que $E(X_0) = X_0$ como también que la varianza de X_0 es cero, $Var(X_0) = 0$.

Proposición 4. ■ *Si la expansión es aleatoria se tiene*

$$Var(X_{n+1} | X_n) = X_n(N - X_n) \left(\frac{(r+s)(r+s-1)}{s(N-1)N} \right) \quad (2.1.20)$$

y

$$Var(X_{n+1}) = X_0(N - X_0) \left(1 - \left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N} \right)^{n+1} \right). \quad (2.1.21)$$

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

- Si la expansión es determinística se tiene

$$\text{Var}(X_{n+1} | X_n) = X_n(N - X_n) \frac{r(r+s)}{s(N-1)N} \quad (2.1.22)$$

y

$$\text{Var}(X_{n+1}) = X_0(N - X_0) \left(1 - \left(1 - \frac{r(r+s)}{s(N-1)N} \right)^{n+1} \right). \quad (2.1.23)$$

Demostración. Tanto para calcular la varianza de X_{n+1} condicional a X_n como para calcular la varianza de X_{n+1} retomemos las expresiones para la varianza (en términos de σ^2) que se obtuvieron cuando estudiamos el modelo de Cannings. Luego para calcular la $\text{Var}(X_{n+1} | X_n)$ y $\text{Var}(X_{n+1})$ tenemos que reemplazar a σ^2 en las siguientes expresiones, que hallamos en el Capítulo 1, Sección 1.2:

$$\text{Var}(X_{n+1} | X_n) = \sigma^2 \frac{X_n(N - X_n)}{(N-1)} \quad (1.2.11)$$

y

$$\text{Var}(X_{n+1}) = X_0(N - X_0) \left(1 - \left(1 - \frac{\sigma^2}{(N-1)} \right)^{n+1} \right) \quad (1.2.15).$$

El reemplazo en cada una de estas expresiones de σ^2 por $\frac{(r+s)(r+s-1)}{sN}$ si la expansión es aleatoria nos da (2.1.20) y (2.1.21), mientras que si reemplazamos σ^2 por $\frac{(r+s)r}{sN}$ si la expansión es determinística obtenemos (2.1.22) y (2.1.23).

Ahora calculemos la heterocigosidad

- Proposición 5.** ▪ Si la expansión es aleatoria la heterocigosidad es

$$h(n) = h(0) \left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N} \right)^n. \quad (2.1.24)$$

- Si la expansión es determinística la heterocigosidad es

$$h(n) = h(0) \left(1 - \frac{r(r+s)}{N(N-1)s} \right)^n. \quad (2.1.25)$$

Demostración. En el Capítulo 1, Sección 1.2 hallamos una expresión para la

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

heterocigosidad expresada en términos de λ_2 , que es

$$h(n) = h(0)\lambda_2^n \quad (1.2.18)$$

donde $h(0) = 2\frac{X_0}{N}\frac{N-X_0}{N-1}$ es la probabilidad condicional a X_0 de que dos alelos elegidos aleatoriamente de la población inicial sean diferentes. Entonces reemplazando λ_2 por (2.1.18) si la expansión es aleatoria y por (2.1.19) si la expansión es determinística en (1.2.18) se obtiene (2.1.24) y (2.1.25), respectivamente.

2.1.4. Tamaño efectivo y número efectivo de generaciones

Cálculo de los tamaños efectivos

Entender el proceso de deriva y la pérdida de diversidad en la población requiere información acerca de su tamaño poblacional efectivo, que es una estimación comparativa del número de individuos que dejarán descendencia. Éstos son los que contribuyen a la generación siguiente, y por lo tanto su número describe cómo una población puede perder diversidad genética más rápidamente de lo predicho por el tamaño poblacional. El modelo de Wright-Fisher es el que se toma como referencia para las definiciones. Como vimos en el Capítulo 1, secciones 1.1 y 1.2, los tamaños efectivos de la población para los procesos de Cannings respecto a la heterocigosidad (N_e^h), respecto a λ_2 (N_e^a) y respecto a la varianza condicional (N_e^v) son iguales:

$$N_e^h = N_e^a = N_e^v = (N-1)/\sigma^2 \quad (1.2.21).$$

Para calcular los tamaños efectivos de los modelos propuestos debemos reemplazar, si la expansión es aleatoria, a σ^2 por $\frac{(r+s)(r+s-1)}{sN}$ en (1.2.21) y se obtiene:

$$N_e^h = N_e^a = N_e^v = \frac{s(N-1)N}{(r+s)(r+s-1)}. \quad (2.1.26)$$

Mientras que si la expansión es determinística se reemplaza a σ^2 por $\frac{(r+s)r}{sN}$ en (1.2.21) y se obtiene:

$$N_e^h = N_e^a = N_e^v = \frac{s(N-1)N}{(r+s)r}. \quad (2.1.27)$$

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

Tamaño efectivo y heterocigosidad

Si consideramos la fórmula de la heterocigosidad para el modelo con expansión aleatoria

$$h(n) = h(0) \left(1 - \frac{1}{\frac{s(N-1)N}{(r+s)(r+s-1)}} \right)^n$$

y reemplazamos $\frac{s(N-1)N}{(r+s)(r+s-1)}$ por N_e^h tenemos

$$h(n) = h(0) \left(1 - \frac{1}{N_e^h} \right)^n .$$

De la misma manera si en la fórmula para la heterocigosidad del modelo con expansión determinística

$$h(n) = h(0) \left(1 - \frac{1}{\frac{N(N-1)s}{r(r+s)}} \right)^n$$

reemplazamos $\frac{N(N-1)s}{r(r+s)}$ por N_e^h se tiene que

$$h(n) = h(0) \left(1 - \frac{1}{N_e^h} \right)^n .$$

A partir de estas últimas expresiones tenemos que la tasa a la cual se reduce la diversidad en la población es

$$1 - 1/N_e^h . \tag{2.1.28}$$

Si $1 - 1/N_e^h$ es pequeño hay gran disminución de la heterocigosidad, lo cual implica mucha deriva.

En particular, si N_e^h vale 1, 2, 3, ..., tenemos que $1 - \frac{1}{N_e^h}$ vale $0, \frac{1}{2}, \frac{2}{3}, \dots$ convergiendo a 1 para $N_e^h \rightarrow \infty$, lo cual corresponde a ausencia de deriva, o sea ausencia de disminución de la diversidad. Esto no quita que pueden tener sentido valores intermedios de la tasa $1 - \frac{1}{N_e^h}$ (como por ejemplo $\frac{1}{3}$) debido al cálculo del cual surge N_e^h , el cual podría no dar entero. Pero la tasa debería estar en $[0, \infty)$ y no dar valores menores que 0 porque tendríamos $h(n) < 0$, que no tiene sentido dada la definición de heterocigosidad.

Con el fin de determinar cuándo la deriva es máxima, estudiamos el compor-

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

tamiento de la tasa de disminución de la heterocigosidad cuando el número de individuos que mueren más el número de individuos que participan en la reproducción es $r + s = cN$ donde c es una constante tal que $0 < c \leq 1$.

Proposición 6. *Sea $r + s$ suficientemente grande, de manera que $r + s \approx r + s - 1$ y $N \approx N - 1$. Entonces la tasa $1 - 1/N_e^h$ se minimiza para cada c fijo con $0 < c \leq 1$ cuando $s = 1$, y en ese caso se tiene*

$$1 - 1/N_e^h \approx 1 - c^2. \quad (2.1.29)$$

Demostración. *Veamos que la proposición se cumple tanto en el modelo con reproducción aleatoria como en el de reproducción determinística.*

Consideremos un c fijo ($0 < c \leq 1$) y escribamos al tamaño efectivo en función de $r + s = cN$. Luego obtenemos:

Si la expansión es aleatoria

$$\begin{aligned} N_e^h &= \frac{s(N-1)N}{(r+s)(r+s-1)} && (r+s \approx r+s-1 \text{ y } N \approx N-1) \\ &\approx \frac{sN^2}{(r+s)^2} && (r+s = cN) \\ &= \frac{sN^2}{(cN)^2} \\ &= \frac{s}{c^2}. \end{aligned}$$

Si la expansión es determinística

$$\begin{aligned} N_e^h &= \frac{s(N-1)N}{(r+s)r} && (N \approx N-1) \\ &\approx \frac{sN^2}{(r+s)r} && (r+s = cN) \\ &= \frac{sN^2}{cN(cN-s)}. \end{aligned} \quad (2.1.30)$$

De estas expresiones podemos ver que para c fijo N_e^h es una función creciente

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

de s . Luego, la tasa $1 - 1/N_e^h$ es una función creciente de s y por lo tanto se minimiza cuando s es mínimo, es decir, cuando $s = 1$, y en dicho caso se tiene que

Si la expansión es aleatoria

$$\begin{aligned} 1 - 1/N_e^h &\approx 1 - 1/\frac{s}{c^2} \\ &= 1 - 1/\frac{1}{c^2} \\ &= 1 - c^2. \end{aligned} \tag{2.1.31}$$

Si la expansión es determinística

$$\begin{aligned} 1 - 1/N_e^h &\approx 1 - 1/\frac{sN^2}{cN(cN - s)} \\ &= 1 - 1/\frac{N^2}{cN(cN - 1)} \quad (cN \approx cN - 1). \\ &\approx 1 - 1/\frac{1}{c^2} \\ &= 1 - c^2, \end{aligned} \tag{2.1.32}$$

como queríamos probar.

De esta proposición tenemos que para ambos modelos si $c = 1$ la mínima tasa de disminución de diversidad es $1 - 1/1 = 0$. Lo que ocurre es que toda la población se restaura a partir de un único alelo ($s = 1$), es decir, en la siguiente generación todos los alelos serán de un solo tipo. Además, dado que N_e^h puede interpretarse como el tamaño de una población de Wright-Fisher con deriva equivalente, obtenemos de la proposición el siguiente corolario para nuestros modelos.

Corolario 1. *La situación de máxima deriva (cualquiera sea N) es la de una población de $1/c^2$ individuos que sigue el proceso de Wright-Fisher.*

Número efectivo de generaciones

Proponemos aquí un nuevo parámetro para describir la magnitud de la deriva a lo largo del proceso de evolución de las poblaciones.

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

Definición 1. *El número efectivo de generaciones con respecto a la heterocigosidad, que notamos por n_e^h , es el número de generaciones de los procesos en estudio para el cual el cambio de la heterocigosidad por deriva coincide con la del modelo de Wright-Fisher en una generación.*

El número efectivo de generaciones con respecto a la varianza, que notamos por n_e^v , es el número de generaciones de los procesos en estudio para el cual la varianza del número de alelos de tipo A en la generación n coincide con la del modelo de Wright-Fisher en una generación.

Ahora definiremos el número efectivo de generaciones respecto a λ_2 recordando que dicho autovalor es la tasa a la cual se pierde la heterocigosidad en la población en cada generación.

Definición 2. *El número efectivo de generaciones con respecto a λ_2 , que notamos por n_e^a , es el número de generaciones de los procesos en estudio para el cual la tasa a la cual se pierde la heterocigosidad por deriva coincide con la del modelo de Wright-Fisher en una generación.*

Proposición 7. *El número efectivo de generaciones con respecto a la heterocigosidad, a la varianza y a λ_2 son iguales en ambos modelos:*

- *Expansión aleatoria*

$$n_e^h = n_e^a = n_e^v = \frac{\ln\left(1 - \frac{1}{N}\right)}{\ln\left(1 - \frac{(r+s)(r+s-1)}{sN(N-1)}\right)}. \quad (2.1.33)$$

- *Expansión determinística*

$$n_e^h = n_e^a = n_e^v = \frac{\ln\left(1 - \frac{1}{N}\right)}{\ln\left(1 - \frac{r(r+s)}{N(N-1)}\right)}. \quad (2.1.34)$$

Demostración. Para calcular

- n_e^v : tenemos que igualar las expresiones de la varianza suponiendo que inicialmente ambas poblaciones tienen la misma proporción de alelos A, p_0 .
- n_e^a : tenemos que igualar las tasas a la cual se pierde la heterocigosidad en la población en cada generación.

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

- n_e^h : tenemos que igualar las fórmulas de la heterocigosidad suponiendo que inicialmente la proporción de alelos A en ambas poblaciones es la misma.

Entonces consideremos las expresiones para la varianza, λ_2 y la heterocigosidad: Si la expansión es aleatoria (generalización de Wright-Fisher) las expresiones que igualaremos son:

Wright-Fisher	Generalización de Wright-Fisher
$Var(\frac{X_n}{N}) = p_0(1-p_0)\left(1 - \left(1 - \frac{1}{N}\right)^n\right)$	$Var(\frac{X_n}{N}) = p_0(1-p_0)\left(1 - \left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)^n\right)$
$\lambda_2 = 1 - \frac{1}{N}$	$\lambda_2 = 1 - \frac{(r+s)(r+s-1)}{s(N-1)N}$
$h(n) = h(0)\lambda_2^n$	$h(n) = h(0)\lambda_2^n$

De manera que obtenemos

n_e^v	$p_0(1-p_0)\left(1 - \left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)^{n_e^v}\right) = p_0(1-p_0)\left(1 - \left(1 - \frac{1}{N}\right)^{n_e^v}\right)$ <p>Luego $\left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)^{n_e^v} = \left(1 - \frac{1}{N}\right)^{n_e^v} \therefore n_e^v = \frac{\ln\left(1 - \frac{1}{N}\right)}{\ln\left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)}$</p>
n_e^a	$\left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)^{n_e^a} = \left(1 - \frac{1}{N}\right)^{n_e^a}$ <p>Luego $n_e^a = \frac{\ln\left(1 - \frac{1}{N}\right)}{\ln\left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)}$</p>
n_e^h	$h(0)\left(1 - \left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)^{n_e^h}\right) = h(0)\left(1 - \left(1 - \frac{1}{N}\right)^{n_e^h}\right)$ <p>Luego $\left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)^{n_e^h} = \left(1 - \frac{1}{N}\right)^{n_e^h} \therefore n_e^h = \frac{\ln\left(1 - \frac{1}{N}\right)}{\ln\left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)}$</p>

De esta última tabla se puede observar que los números efectivos de generaciones respecto a la varianza, a λ_2 y a la heterocigosidad son iguales:

$$n_e^v = n_e^a = n_e^h = \frac{\ln\left(1 - \frac{1}{N}\right)}{\ln\left(1 - \frac{(r+s)(r+s-1)}{s(N-1)N}\right)}. \quad (2.1.35)$$

Mientras que si la expansión es determinística (variante de Wright-Fisher) las expresiones que igualaremos son:

Wright-Fisher	Variante de Wright-Fisher
$Var(\frac{X_n}{N}) = p_0(1-p_0)\left(1 - \left(1 - \frac{1}{N}\right)^n\right)$	$Var(\frac{X_n}{N}) = p_0(1-p_0)\left(1 - \left(1 - \frac{r(r+s)}{N(N-1)}\right)^n\right)$
$\lambda_2 = 1 - \frac{1}{N}$	$\lambda_2 = 1 - \frac{r(r+s)}{N(N-1)}$
$h(n) = h(0)\lambda_2^n$	$h(n) = h(0)\lambda_2^n$

En este caso obtenemos

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

n_e^v	$p_0 (1 - p_0) \left(1 - \left(1 - \frac{r(r+s)}{N(N-1)} \right)^{n_e^v} \right) = p_0 (1 - p_0) \left(1 - \left(1 - \frac{1}{N} \right) \right)$ <p style="text-align: center;">Luego $\left(1 - \frac{r(r+s)}{N(N-1)} \right)^{n_e^v} = \left(1 - \frac{1}{N} \right) \therefore n_e^v = \frac{\ln \left(1 - \frac{1}{N} \right)}{\ln \left(1 - \frac{r(r+s)}{N(N-1)} \right)}$</p>
n_e^a	$\left(1 - \frac{r(r+s)}{N(N-1)} \right)^{n_e^a} = \left(1 - \frac{1}{N} \right)$ <p style="text-align: center;">Luego $n_e^a = \frac{\ln \left(1 - \frac{1}{N} \right)}{\ln \left(1 - \frac{r(r+s)}{N(N-1)} \right)}$</p>
n_e^h	$h(0) \left(1 - \left(1 - \frac{r(r+s)}{N(N-1)} \right)^{n_e^h} \right) = h(0) \left(1 - \left(1 - \frac{1}{N} \right) \right)$ <p style="text-align: center;">Luego $\left(1 - \frac{r(r+s)}{N(N-1)} \right)^{n_e^h} = \left(1 - \frac{1}{N} \right) \therefore n_e^h = \frac{\ln \left(1 - \frac{1}{N} \right)}{\ln \left(1 - \frac{r(r+s)}{N(N-1)} \right)}$</p>

Nuevamente se tiene que los números efectivos de generaciones respecto a la varianza, a λ_2 y a la heterocigocidad son iguales:

$$n_e^v = n_e^a = n_e^h = \frac{\ln \left(1 - \frac{1}{N} \right)}{\ln \left(1 - \frac{r(r+s)}{N(N-1)} \right)}. \quad (2.1.36)$$

Luego como n_e^h , n_e^a y n_e^v coinciden, los notaremos por n_e . □

El número efectivo de generaciones toma valores en el intervalo $(0, +\infty)$. Si n_e está en el intervalo $(0, 1)$ implica que en una generación nuestros modelos producen más deriva que el modelo de Wright-Fisher. Si $n_e = 1$ nuestros modelos en una generación producen la misma deriva que el de Wright-Fisher. Mientras que si $n_e > 1$ el modelo de Wright-Fisher produce más deriva en una generación en comparación con nuestros modelos.

El número efectivo de generaciones será retomado en el Capítulo 4, en el cual se busca describir los cambios por deriva de una frecuencia alélica en una plaga con el modelo en el cual la reproducción de los individuos que dejan descendencia es aleatoria.

Tamaño efectivo versus tamaño poblacional

Como vimos en el Capítulo 1, Subsección 1.1.2 en el modelo de Wright-Fisher la heterocigocidad, según (1.1.37), es

$$h(n) = h(0) \left(1 - \frac{1}{N} \right)^n$$

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

donde $h(0) = 2 \frac{X_0}{N} \frac{N-X_0}{N-1}$ es la probabilidad de que dos alelos elegidos aleatoriamente de la población inicial sean de distinto tipo. A partir de esta expresión se puede concluir que la heterocigosidad disminuirá a una tasa de $1 - 1/N$, donde N es el tamaño poblacional. Por lo tanto, cuanto mayor sea el tamaño de la población menor será el efecto de la deriva. Nos proponemos cuantificar este efecto en los casos en los que en cada generación sólo una fracción de individuos contribuye con su tipo a la siguiente.

En los modelos que estamos estudiando la tasa a la cual se reduce la diversidad en la población es, según (2.1.28), $1 - 1/N_e^h$, es decir, no depende del tamaño de la población sino del tamaño efectivo. Esto motiva el estudio de la deriva como fuerza evolutiva incluso en poblaciones grandes.

En función de esto último comparamos el tamaño efectivo con respecto a la heterocigosidad, N_e^h , y el tamaño de la población, N . Retomamos las expresiones que hallamos para N_e^h y estandarizamos el número de individuos que van a ser reemplazados y el número de individuos elegidos para reproducirse mediante $r' = r/N$ y $s' = s/N$ respectivamente en las expresiones (2.1.26) y (2.1.27). Entonces si la expansión es aleatoria obtenemos

$$N_e^h = N \left(\frac{s'(1 - 1/N)}{(r' + s')(r' + s' - 1/N)} \right). \quad (2.1.37)$$

Mientras que si la expansión es determinística se obtiene

$$N_e^h = N \left(\frac{s'(1 - 1/N)}{(r' + s') r'} \right). \quad (2.1.38)$$

Ahora veamos para qué valores de los parámetros r' y s' se tiene que el tamaño poblacional efectivo respecto a la heterocigosidad coincide con el tamaño de la población o es sólo una fracción del mismo. Para ello consideremos las expresiones (2.1.37) y (2.1.38) y nuevamente como lo hicimos en la Proposición 6 suponemos $r + s$ suficientemente grande como para que $r + s \approx r + s - 1$, y por lo tanto también $N \approx N - 1$ y $1/N \approx 0$. Estos supuestos son coherentes con los tamaños poblacionales de los escenarios biológicos del Capítulo 4.

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

Luego tenemos que los r' y s' satisfacen $N_e^h = kN$ con $0 < k \leq 1$ si cumplen:

$$\begin{array}{ll}
 \text{Si la expansión es aleatoria} & \text{Si la expansión es determinística} \\
 \frac{s'}{(r' + s')^2} = k & \frac{s'}{(r' + s')r'} = k.
 \end{array}
 \tag{2.1.39}$$

Haciendo algunas operaciones algebraicas se obtiene que estas últimas expresiones se pueden reescribir de la siguiente manera:

Si la expansión es aleatoria:

$$0 = kr'^2 + 2ks'r' + s'(ks' - 1).$$

Si la expansión es determinística:

$$0 = kr'^2 + ks'r' - s'.$$

Luego para cada posible valor de s' aquellos valores de r' que satisfacen dichas igualdades son, si la expansión es aleatoria

$$r' = \frac{-2ks' + \sqrt{(2ks')^2 - 4ks'(ks' - 1)}}{2k}$$

y si la expansión es determinística

$$r' = \frac{-ks' + \sqrt{(ks')^2 + 4ks'}}{2k}.$$

Finalmente para representar los s' y r' que satisfacen $N_e^h = kN$, para distintos valores de k , graficaremos las siguientes funciones: $r'_k : [0, 1] \rightarrow [0, 1]$ tal que

$$r'_k(s') = \frac{-2ks' + \sqrt{(2ks')^2 - 4ks'(ks' - 1)}}{2k} \tag{2.1.40}$$

para el modelo en el que la restauración de la población es a partir de una expansión

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

aleatoria, y $r'_k : [0, 1] \rightarrow [0, 1]$ tal que

$$r'_k(s') = \frac{-ks' + \sqrt{(ks')^2 + 4ks'}}{2k} \quad (2.1.41)$$

para el modelo en el que la restauración de la población es a partir de una expansión determinística.

Primero consideremos la curva que representa los puntos de la forma (s', r') tales que $N_e^h = N$ para ambos modelos, es decir, $k = 1$. Entonces obtenemos el conjunto de puntos para los cuales $N_e^h < N$ que se corresponde con el área del gráfico que se encuentra por encima de cada curva y el conjunto de puntos para los cuales $N_e^h > N$ que se corresponde con el área del gráfico que se encuentra por debajo de cada curva (ver Figura 2.3).

Para cualquier población, cuanto más pequeña es, más afectada se ve por el fenómeno de muestreo aleatorio que caracteriza a la deriva. Así es que la deriva es más significativa en la región donde $N_e^h < N$ porque esto significa que la población se comporta como si efectivamente tuviera menos de N individuos. Veamos entonces para qué valores de r' y s' el número efectivo resulta menor que el total de individuos, N , estableciendo relaciones con el fenómeno de la genética poblacional conocido como “cuello de botella”.

Analicemos la proporción de alelos que se eliminan (r') en función de la proporción de alelos que se reproducen (s') que satisfacen $N_e^h = kN$ para $k = 1$, $k = 1/2$, $k = 1/2^2$, $k = 1/2^3$ y $k = 1/2^4$, tanto en el modelo con renovación aleatoria como en el de renovación determinística a partir de las figuras 2.4 y 2.5 respectivamente.

Si comparamos en las figuras 2.4 y 2.5 los números efectivos entre los modelos para los mismos valores de parámetros observamos que cuando la expansión es aleatoria la deriva es mayor, es decir, las curvas punteadas se encuentran por encima de las curvas continuas (ver también Figura 2.3). Esto quiere decir que para que ambos modelos den lugar a la misma deriva para un mismo valor de s , en el modelo con renovación determinística tiene que ser mayor el número de alelos que serán reemplazados, r , respecto al modelo con expansión aleatoria. En efecto, en la expansión aleatoria la deriva tiene lugar tanto en la elección de los s individuos que dejarán descendencia como en el muestreo aleatorio de tamaño $r + s$ que restaura el tamaño poblacional, mientras que si la expansión es determinística, los $r + s$ individuos producto de dicha expansión tendrán las mismas proporciones de alelos A y a que los s , por lo cual la deriva sólo se produce en la elección de los

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

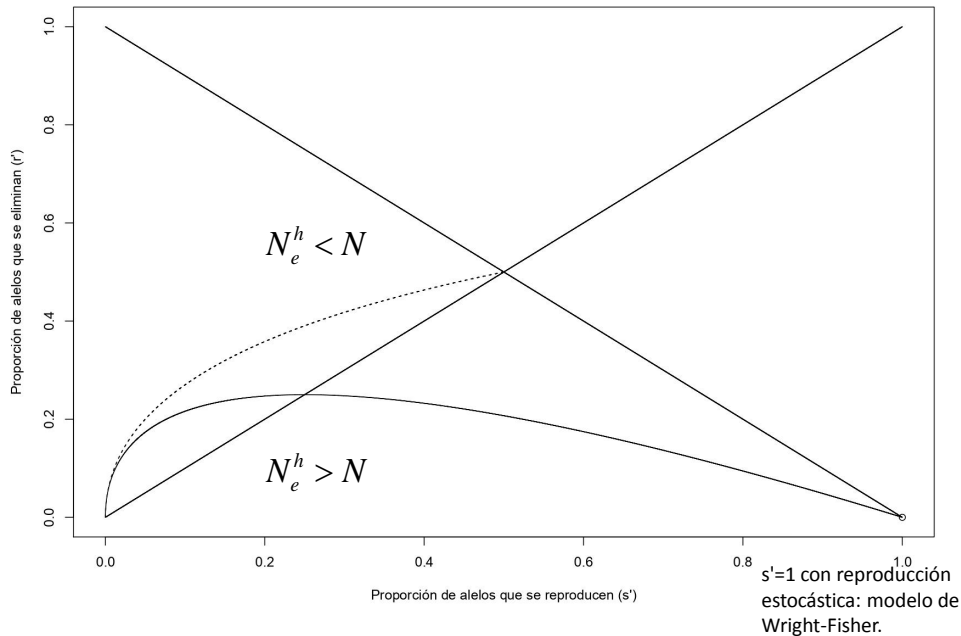


Figura 2.3: Proporción de alelos que se eliminan en función de la proporción de los que se reproducen bajo la condición tamaño efectivo igual al tamaño poblacional. La función que se representa con línea punteada corresponde al modelo con expansión determinística y la función que se representa con línea continua al modelo con expansión aleatoria. Además graficamos la recta $r' = 1 - s'$ y la recta $r' = s'$.

s individuos que dejarán descendencia.

Las figuras 2.4 y 2.5 permiten distinguir distintas situaciones de interés para el paso de una generación a la siguiente:

- Si $r' + s' = 1$ (corresponde a la diagonal del gráfico en la que $r' + s' = 1$) en el paso de una generación a la siguiente se renueva toda la población. Luego si $s' \approx 1/N$ las sucesivas generaciones se originan de unos pocos individuos, que no necesariamente van a tener una composición genética semejante a la de la población de la que fueron elegidos. La proporción de alelos o tipos de cada variante dentro de los s podría ser muy diferente a la de toda la población. Esto es lo que suele llamarse un evento de cuello de botella.

CAPÍTULO 2. GENERALIZACIÓN Y VARIANTE DEL MODELO DE WRIGHT-FISHER

- Dado un valor fijo de $r' + s'$ o sea un valor fijo también de individuos que permanecen sin participar en la muerte ni en la reproducción (que son $1 - (r' + s')$), podemos tener una alta deriva para s' pequeño, aún si $1 - (r' + s')$ es distinto de cero. Los gráficos de las curvas que cortan las subdiagonales correspondientes a valores fijos de $r' + s'$ sugieren que N_e^h se puede hacer arbitrariamente pequeño si se achica s' , pero esta reducción está acotada por el hecho de que $s \geq 1$ y por lo tanto $s' \geq 1/N$. Además excluimos del análisis valores de $r' + s'$ muy cercanos a cero porque estamos suponiendo que $r + s \approx r + s - 1$, lo cual es consistente con las aplicaciones biológicas que estudiaremos en el Capítulo 4.

Tamaño efectivo en distintos escenarios de renovación parcial si la expansión es aleatoria

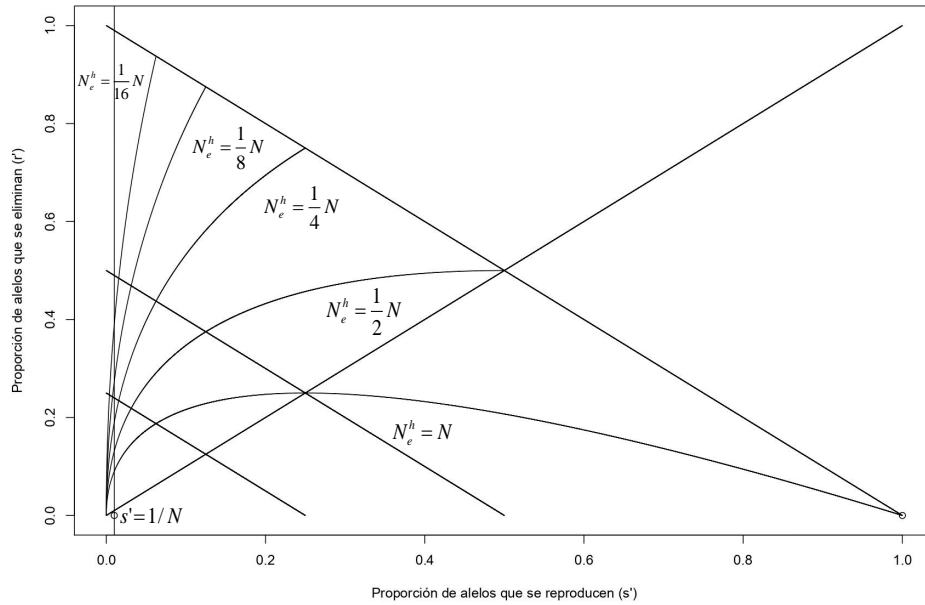


Figura 2.4: Gráfico para el modelo con restauración aleatoria de la población de la proporción de alelos que se eliminan en función de la proporción de los que se reproducen bajo la condición $N_e^h = kN$ para $k = 1, 1/2, 1/2^2, 1/2^3, 1/2^4$. Además graficamos las rectas $r' = 1 - s'$, $r' = 1/2 - s'$, $r' = 1/4 - s'$, $r' = s'$ y la recta vertical $s' = 1/N$.

2.1. EXPANSIÓN DETERMINÍSTICA Y EXPANSIÓN ALEATORIA

Tamaño efectivo en distintos escenarios de renovación parcial si la expansión es determinística

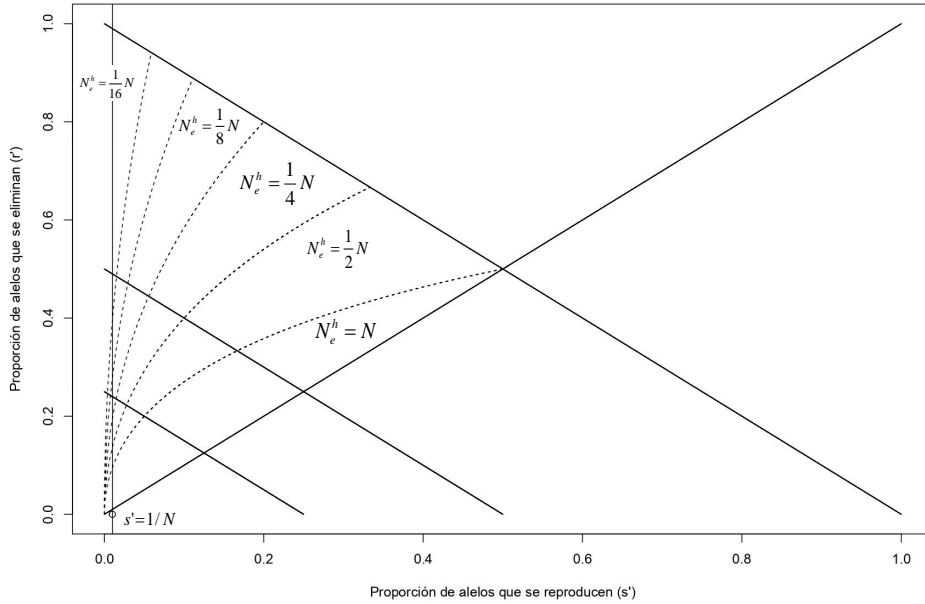


Figura 2.5: Gráfico para el modelo con restauración determinística de la proporción de alelos que se eliminan en función de la proporción de los que se reproducen bajo la condición $N_e^h = kN$ para $k = 1, 1/2, 1/2^2, 1/2^3, 1/2^4$. Además graficamos las rectas $r' = 1 - s'$, $r' = 1/2 - s'$, $r' = 1/4 - s'$, $r' = s'$ y la recta vertical $s' = 1/N$.

Finalmente concluimos que en nuestros modelos de renovación parcial la población no necesariamente tiene que atrevesar un cuello de botella para que los efectos de la deriva sean significativos (ver Cuadro 2.1).

	$r + s = N$	$r + s < N$
Reproducción Aleatoria	$s \approx N$ Modelo de Wright-Fisher $s \approx 1$ Cuello de botella	$s \approx 1$ Alta deriva
Reproducción Determinística	$s \approx 1$ Cuello de botella	$s \approx 1$ Alta deriva

Cuadro 2.1: En este cuadro se resumen para ambos modelos distintos escenarios para los cuales la deriva es significativa.

2.2. Distribuciones extremas de las variables y_i en nuestros modelos

En los modelos con renovación parcial que hemos propuesto se pueden alcanzar distribuciones extremas de las variables y_i según el valor de los parámetros s y r . Estas corresponden a mínima y máxima deriva posible dentro de todos los modelos de Cannings y las describimos a continuación.

Mínima deriva:

Si para todo i tenemos $P(y_i = 1) = 1$, entonces la varianza σ^2 de las y_i es nula, la heterocigosidad se mantiene constante y $\lambda_0 = \lambda_1 = \lambda_2 = \dots = \lambda_N = 1$. En virtud del Teorema 1.2.2, este espectro domina (tiene valores mayores o iguales, componente a componente) al que se obtendría con cualquier otra distribución de las y_i . En el modelo de expansión determinística este caso extremo se realiza con $s = N$ y $r = 0$.

Máxima deriva:

Si para todo i tenemos $P(y_i = N) = 1/N$ y $P(y_i = 0) = 1 - 1/N$, entonces la varianza σ^2 de las y_i es $N - 1$, la heterocigosidad se anula en un paso, $N_e = 1$ y $\lambda_2 = 0$. Según el Teorema 1.2.2 tenemos entonces que $1 = \lambda_0 = \lambda_1 > \lambda_2 = \lambda_3 = \dots = \lambda_N = 0$ y que este espectro es dominado por el de cualquier otra distribución de las y_i . Tanto en el modelo de expansión aleatoria como en el de expansión determinística, este caso extremo se realiza con $s = 1$ y $r = N - 1$. Es decir, todos los individuos de la generación n son descendientes del mismo padre de la generación $n - 1$.

Capítulo 3

Variante del modelo de urna de Polya-Eggenberger

En este capítulo desarrollaremos una variante del modelo de urna de Polya-Eggenberger para complementar las propuestas anteriores con un proceso en el que el tamaño poblacional aumenta gradualmente desde valores pequeños, como ocurre en una población en crecimiento.

3.1. El modelo de Polya-Eggenberger

Se considera una urna que inicialmente tiene N bolillas, de las cuales α son rojas y $\beta = N - \alpha$ negras de manera que $\alpha + \beta = N$. Cada vez que se extrae una bolilla se observa su color y se introduce de nuevo a la urna, junto con c bolillas del mismo color [35]. Para este procedimiento, se estudia cómo evoluciona la frecuencia de bolillas rojas (o negras) luego de las sucesivas extracciones.

La probabilidad de haber elegido l bolillas rojas y $m = n - l$ bolillas negras en las n primeras extracciones es:

$$P(X_n = \alpha + cl) = \binom{n}{l} \frac{\alpha(\alpha + c)\dots(\alpha + (l-1)c)\beta(\beta + c)\dots(\beta + (m-1)c)}{N(N + c)\dots(N + (n-1)c)} \quad (3.1.1)$$

que se puede reescribir en términos de la función beta

$$P(X_n = \alpha + cl) = \binom{n}{l} \frac{\beta(\alpha/c + l, \beta/c + m)}{\beta(\alpha/c, \beta/c)}, \quad (3.1.2)$$

CAPÍTULO 3. VARIANTE DEL MODELO DE URNA DE POLYA-EGGENBERGER

donde X_n es la variable aleatoria que representa el número de bolillas rojas en la urna en el paso n , $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ para cada $a, b \in \mathbb{R}$ y $\Gamma(n) = (n-1)!$ para cada $n \in \mathbb{N}$. En la Sección 3.2 deduciremos la versión de esta fórmula que se aplica a nuestra variante y en la que se toma $c = 1$.

3.1.1. Convergencia con $c = 1$

En este caso particular se considera una urna que inicialmente tiene N bolillas, de las cuales α son rojas y $\beta = N - \alpha$ negras, y cada vez que se extrae una bolilla se observa su color y se introduce de nuevo a la urna, junto con otra bolilla de su mismo color.

Tenemos que X_n es una variable aleatoria que representa el número de bolillas rojas en la urna en el paso n tal que $\frac{X_n}{n+N}$ converge c.s., y en particular en distribución, a una variable aleatoria $\beta_{\alpha, \beta}$ que tiene distribución beta con parámetros α y β , es decir,

$$\lim_{n \rightarrow \infty} P\left(\frac{X_n}{n+N} \leq x\right) = P(\beta_{\alpha, \beta} \leq x),$$

donde $\beta_{\alpha, \beta} \sim \beta(\alpha, \beta)$. Este es un resultado clásico del modelo de urna de Polya-Eggenberger ([35] y [19]).

3.2. Variante del modelo de Polya-Eggenberger

Consideremos el caso de urna de Polya-Eggenberger para $c = 1$, es decir, en cada iteración se agrega a la urna una bolilla. En la variante que proponemos la proporción inicial de bolillas rojas (o negras) es aleatoria [8]. En este sentido, generalizamos el modelo conocido en un aspecto que no encontramos en trabajos sobre el tema.

Entonces definimos la variable aleatoria X_0 como el número de bolillas rojas que hay inicialmente en la urna. Vamos a denotar por N_0 el número inicial de bolillas en la urna. En nuestra variante X_0 tiene distribución binomial con parámetros N_0 y $1/2$, lo cual corresponde a la situación biológica que estudiaremos en el Capítulo 4, Subsección 4.3.2.

El proceso de deriva que proponemos consiste en elegir aleatoriamente una bolilla e ir agregando exactamente una bolilla del mismo color que la que fue

3.2. VARIANTE DEL MODELO DE POLYA-EGGENBERGER

elegida a cada paso. Este procedimiento se repite una y otra vez. Esto es, en la primera generación se elige una bolilla al azar del total de N_0 . Si se elige una roja (la probabilidad según X_0 de que esto ocurra es $\frac{X_0}{N_0}$), en la siguiente generación habrá $N_0 + 1$ bolillas, $X_0 + 1$ rojas y $N_0 - X_0$ negras. De lo contrario, si se eligió una negra (la probabilidad según X_0 de que esto ocurra es $1 - \frac{X_0}{N_0}$) en la siguiente generación habrá $N_0 + 1$ bolillas, X_0 rojas y $N_0 - X_0 + 1$ negras.

En general, la generación n tendrá $N_0 + n$ bolillas, donde n representa el número de bolillas que se agregaron a la urna hasta el paso n . La variable aleatoria X_n representa el número de bolillas rojas en la generación n , es decir, luego de la incorporación de n nuevas bolillas a las N_0 iniciales.

Función de probabilidad condicional del número de bolillas rojas en cada generación

Procederemos al cálculo de la probabilidad $P(X_n = j + k \mid X_0 = j)$ para $0 \leq j \leq N_0$ y $0 \leq k \leq n$, que notaremos por $p_{j,j+k}^n$ y será

$$p_{j,j+k}^n = \begin{cases} 0 & \text{si } j = 0 \text{ y } k \neq 0 \text{ o si } j = N_0 \text{ y } k \neq n \\ 1 & \text{si } j = 0 \text{ y } k = 0 \text{ o si } j = N_0 \text{ y } k = n \\ \binom{n}{k} \frac{\beta(j+k, N_0-j+n-k)}{\beta(j, N_0-j)} & \text{si } 1 \leq j \leq N_0 - 1 \text{ y } 0 \leq k \leq n \end{cases} \quad (3.2.1)$$

donde β representa la función beta, definida por $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ para cada $a, b \in \mathbb{R}$.

Justifiquemos primero los casos extremos, en los que la probabilidad da 0 o 1. Si inicialmente todas las bolillas son rojas ($X_0 = N_0$), en cada generación siempre se elegirá una bolilla roja. Luego con probabilidad uno el número de bolillas rojas se incrementará en n , es decir, $p_{N_0, N_0+n}^n = 1$ y con probabilidad cero $X_n \neq N_0 + n$. En cambio, si inicialmente todas las bolillas son negras ($X_0 = 0$), no es posible elegir una bolilla roja, lo cual implica que con probabilidad uno $X_n = 0$, es decir, $p_{0,0}^n = 1$, mientras que la probabilidad de que $X_n \neq 0$ es cero.

Ahora calculemos $P(X_n = j + k \mid X_0 = j)$ para $1 \leq j \leq N_0 - 1$ y $0 \leq k \leq n$, es decir, la probabilidad de que el número inicial de bolillas rojas después de n pasos se incremente en k cuando partimos de una población en la que ambos tipos están representados. Para ello definimos y_i , una variable aleatoria que toma el valor 1 si en la generación i se eligió una bolilla roja, y 0 en caso contrario. Por lo tanto,

CAPÍTULO 3. VARIANTE DEL MODELO DE URNA DE POLYA-EGGENBERGER

$\sum_{i=1}^n y_i$ se corresponde con el número de bolillas rojas que se agregaron hasta el paso n . En el caso que nos ocupa, $\sum_{i=1}^n y_i = k$.

Esto puede darse de distintas maneras. Por ejemplo, podría ocurrir que las primeras k veces se seleccionen bolillas rojas, y por lo tanto, el resto de las veces se seleccionen bolillas negras. En este caso tendríamos que $y_1 = 1, y_2 = 1, \dots, y_k = 1, y_{k+1} = 0, \dots, y_n = 0$ (se corresponde con la siguiente combinación de k unos y $n - k$ ceros: $(1 \dots 10 \dots 0)$). Otra posibilidad sería que las primeras $n - k$ veces se seleccionen bolillas negras, y en las últimas k veces, se seleccionen todas rojas. Entonces tendríamos que $y_1 = 0, \dots, y_{n-k} = 0, y_{n-k+1} = 1, \dots, y_n = 1$ (se corresponde con la siguiente combinación de k unos y $n - k$ ceros: $(0 \dots 01 \dots 1)$). Entonces el conjunto de resultados posibles de las variables aleatorias y_1, \dots, y_n tales que $\sum_{i=1}^n y_i = k$ tendrá tantos elementos como maneras de combinar k unos y $n - k$ ceros, y por lo tanto su cardinal es $\binom{n}{k}$.

Ahora demostremos que

$$\begin{aligned} & P(y_1 = i_1, \dots, y_n = i_n | X_0 = j) \\ &= \frac{j(j+1)(j+2) \dots (j+k-1)(N_0-j)(N_0-j+1) \dots (N_0-j+n-k-1)}{N_0(N_0+1) \dots (N_0+n-1)} \end{aligned} \quad (3.2.2)$$

para $(i_1, i_2, \dots, i_n) \in A_k$, siendo $A_k = \{(i_1, i_2, \dots, i_n) \in \{0, 1\}^n : \sum_{j=1}^n i_j = k\}$.

En palabras, queremos ver que todas las posibles maneras de que ocurra el evento $\{X_n = j + k | X_0 = j\}$ (que como vimos representa que el número inicial de bolillas rojas después de n pasos se incrementa en k) tienen la misma probabilidad, la cual sigue la expresión (3.2.2). Esta fórmula se puede justificar, cualquiera sea $(i_1, i_2, \dots, i_n) \in A_k$, considerando el número de casos favorables al evento $\{y_1 = i_1, \dots, y_n = i_n | X_0 = j\}$ dividido por el número de casos posibles.

Los casos posibles son igualmente probables porque en cada paso todas las bolillas tienen la misma probabilidad de ser elegidas, en efecto, en cualquier generación h la probabilidad de elegir una bolilla es $\frac{1}{N_0+h-1}$ cualquiera sea la composición de la urna en ese momento. El número total de casos posibles es la productoria que aparece en el denominador de (3.2.2) y se calcula en base al número de bolillas que hay en el momento de los sucesivos muestreos. Primero se elige al azar una bolilla de las N_0 , a continuación habrá $N_0 + 1$ bolillas, entre las cuales se elegirá nuevamente una bolilla al azar y así se sigue con el proceso hasta que haya $N_0 + n - 1$

3.2. VARIANTE DEL MODELO DE POLYA-EGGENBERGER

bolillas, que es el total de bolillas al realizar la última de las n elecciones.

El cómputo de la productoria del numerador de (3.2.2), que tiene n factores al igual que el denominador y corresponde a los casos favorables, se basa en que, cualquiera sea el orden en que aparecen los k unos y $n - k$ ceros, se trata de los mismos factores en distinto orden. Para justificar esto observamos en primer lugar que para cualquier caso en el que $\sum_{j=1}^n i_j = k$ existirá un r tal que es el primer subíndice que cumple $y_r = 1$. En esa circunstancia habrá j posibilidades de elegir una bolilla roja en la r -ésima generación sobre un total de $N_0 + r - 1$ bolillas. Más aún, notemos que haya habido o no elecciones previas de bolillas negras, los casos favorables en el momento de elegir una roja son j y después habrá $j + 1$ rojas cuando se elige la segunda bolilla roja, independientemente de la cantidad de bolillas negras que se hayan elegido previamente. Los factores continúan así hasta $j + k - 1$, que es la cantidad de bolillas rojas cuando se elige la última bolilla de dicho color.

Lo mismo ocurre con la elección de las bolillas negras. En la primera elección de una bolilla negra habrá $N_0 - j$ bolillas negras, en el momento de elegir la segunda negra habrá $N_0 - j + 1$ y así hasta elegir la $n - k$ -ésima bolilla negra. Entonces $N_0 - j, N_0 - j + 1, \dots, N_0 - j + n - k - 1$ son las cantidades de bolillas negras al realizar la primera extracción de una negra, la segunda, y así hasta la última, respectivamente. También análogamente al caso de las bolillas rojas, las elecciones de bolillas negras pudieron haber tenido elecciones de rojas intercaladas pero esto no modifica el número de bolillas negras presentes a cada paso, que son las que cuentan para el número de maneras de elegir la primera, la segunda, y hasta la última bolilla que se eligió.

Entonces para cualesquiera $(i_1, i_2, \dots, i_n), (i'_1, i'_2, \dots, i'_n) \in A_k$ se tiene que $p = P(y_1 = i_1, \dots, y_n = i_n | X_0 = j)$ y $p' = P(y_1 = i'_1, \dots, y_n = i'_n | X_0 = j)$ son un producto de n factores que tienen el mismo denominador y los numeradores de los factores de p son los mismos que los de p' , pero están permutados, y por lo tanto, $p = p'$. Esto prueba (3.2.2).

Considerando (3.2.2) y el hecho de que el cardinal de A_k es $\binom{n}{k}$, se tiene

$$P(X_n = j + k | X_0 = j) = \binom{n}{k} \frac{j \dots (j + k - 1)(N_0 - j) \dots (N_0 - j + n - k + 1)}{N_0(N_0 + 1) \dots (N_0 + n - 1)}. \quad (3.2.3)$$

Para terminar la prueba de la fórmula que dimos para esta probabilidad, vea-

CAPÍTULO 3. VARIANTE DEL MODELO DE URNA DE POLYA-EGGENBERGER

mos que esta expresión se puede escribir de manera más compacta mediante la siguiente expresión:

$$P(X_n = j + k \mid X_0 = j) = \binom{n}{k} \frac{\beta(j + k, N_0 - j + n - k)}{\beta(j, N_0 - j)}. \quad (3.2.4)$$

Queremos ver que $\frac{j \dots (j+k-1)(N_0-j) \dots (N_0-j+n-k-1)}{N_0(N_0+1) \dots (N_0+n-1)} = \frac{\beta(j+k, N_0-j+n-k)}{\beta(j, N_0-j)}$.

$$\frac{\prod_{i=j}^{j+k-1} i \prod_{i=0}^{n-k-1} (N_0 - j + i)}{\prod_{i=N_0}^{N_0+n-1} i} = \frac{\frac{(j+k-1)! (N_0+n-j-k-1)!}{(j-1)! (N_0-j-1)!}}{(N_0+n-1)!} \quad (3.2.5)$$

$$= \frac{\Gamma(j+k)\Gamma(N_0+n-j-k)}{\Gamma(j)\Gamma(N_0-j)} \frac{\Gamma(N_0)}{\Gamma(N_0+n)} \quad (3.2.6)$$

$$= \frac{\Gamma(j+k)\Gamma(N_0+n-j-k)}{\Gamma(N_0+n)} \frac{\Gamma(N_0)}{\Gamma(j)\Gamma(N_0-j)} \quad (3.2.7)$$

$$= \frac{\beta(j+k, N_0-j+n-k)}{\beta(j, N_0-j)}. \quad (3.2.8)$$

Las expresiones (3.2.5) y (3.2.6) son iguales considerando que $\Gamma(n) = (n-1)!$ para cada $n \in \mathbb{N}$, mientras que (3.2.7) y (3.2.8) son iguales dado que $\beta(j, N_0-j) = \frac{\Gamma(j)\Gamma(N_0-j)}{\Gamma(N_0)}$ y $\beta(j+k, N_0-j+n-k) = \frac{\Gamma(j+k)\Gamma(N_0-j+n-k)}{\Gamma(N_0+n)}$. Esto último resulta del hecho de que $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ para cada $a, b \in \mathbb{R}$.

Función de probabilidad del número de bolillas rojas en cada generación

Ahora procederemos al cálculo de la probabilidad no condicional, es decir, de $P(X_n = i)$ para $0 \leq i \leq N_0 + n$. Veremos que

$$P(X_n = i) = \begin{cases} \left(\frac{1}{2}\right)^{N_0} & \text{si } i = 0 \text{ o si } i = N_0 + n \\ \left(\frac{1}{2}\right)^{N_0} \sum_{j=\max\{1, i-n\}}^{\min\{i, N_0-1\}} \binom{n}{i-j} \frac{\beta(i, N_0+n-i)}{\beta(j, N_0-j)} \binom{N_0}{j} & \text{si } 1 \leq i \leq N_0 + n - 1 \end{cases} \quad (3.2.9)$$

3.2. VARIANTE DEL MODELO DE POLYA-EGGENBERGER

Por la fórmula de la probabilidad total se tiene que

$$P(X_n = i) = \sum_{j=0}^{N_0} P(X_n = i | X_0 = j) \cdot P(X_0 = j). \quad (3.2.10)$$

Si $i = 0$ entonces para $j \neq 0$ resulta $P(X_n = 0 | X_0 = j) = 0$ y si $j = 0$ $P(X_n = 0 | X_0 = 0) = 1$ por (3.2.1). Luego

$$P(X_n = 0) = P(X_n = 0 | X_0 = 0) \cdot P(X_0 = 0) = P(X_0 = 0) = \left(\frac{1}{2}\right)^{N_0}$$

por el hecho de que X_0 tiene distribución binomial con parámetros N_0 y $\frac{1}{2}$. De la misma manera se puede ver que si $i = N_0 + n$ resulta $P(X_n = N_0 + n) = \left(\frac{1}{2}\right)^{N_0}$.

Ahora calculemos $P(X_n = i)$ para $1 \leq i \leq N_0 + n - 1$ a partir de (3.2.10):

Si $i < j$ se tiene $P(X_n = i | X_0 = j) = 0$ porque el número de bolillas rojas a cada paso o se mantiene igual o se incrementa en uno, nunca disminuye.

Si $i > n + j$ se tiene $P(X_n = i | X_0 = j) = 0$ porque luego de n generaciones el número de bolillas rojas podría incrementarse en a lo sumo $n + j$ unidades.

Si $j \leq i \leq n + j$ entonces $i - n \leq j \leq i$, pero a su vez como j representa el número de bolillas rojas que hay inicialmente se tiene que $0 \leq j \leq N_0$. Los casos $j = 0$ y $j = N_0$ ya fueron estudiados. Luego $P(X_n = i | X_0 = j) > 0$ para cada j tal que $\max\{1, i - n\} \leq j \leq \min\{i, N_0 - 1\}$ (ver Figura 3.1). Finalmente reemplazando en (3.2.10) resulta que

$$\begin{aligned} P(X_n = i) &= \sum_{j=\max\{1, i-n\}}^{\min\{i, N_0-1\}} P(X_n = i | X_0 = j) P(X_0 = j) \\ &= \sum_{j=\max\{1, i-n\}}^{\min\{i, N_0-1\}} \binom{n}{i-j} \frac{\beta(i, N_0 + n - i)}{\beta(j, N_0 - j)} \binom{N_0}{j} \left(\frac{1}{2}\right)^{N_0} \\ &= \left(\frac{1}{2}\right)^{N_0} \sum_{j=\max\{1, i-n\}}^{\min\{i, N_0-1\}} \binom{n}{i-j} \frac{\beta(i, N_0 + n - i)}{\beta(j, N_0 - j)} \binom{N_0}{j} \end{aligned}$$

para cada i ($1 \leq i \leq N_0 + n - 1$).

CAPÍTULO 3. VARIANTE DEL MODELO DE URNA DE POLYA-EGGENBERGER

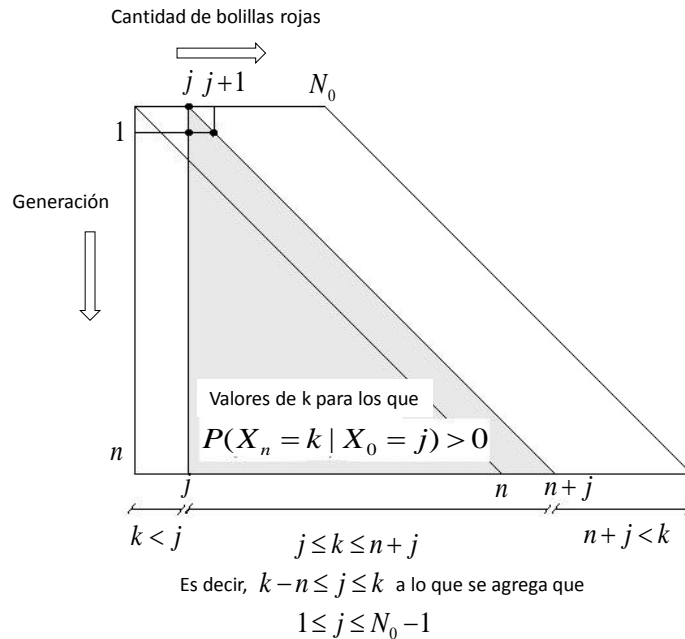


Figura 3.1: En una generación el número de bolillas rojas a cada paso o se mantiene igual o se incrementa en uno, nunca disminuye. Además luego de n generaciones el número inicial de bolillas rojas, j , podría incrementarse en a lo sumo n unidades. Luego la región sombreada representa para cada $1 \leq j \leq N_0 - 1$ aquellos valores de i tales que $P(X_n = i | X_0 = j) > 0$.

Cálculo de la probabilidad de que por lo menos el 90 % de las bolillas en la urna en el paso n sean rojas

Como veremos en el Capítulo 4, Subsección 4.3.2 para la aplicación biológica que consideraremos es de interés la fórmula de $P(X_n \geq 0,9 \cdot (N_0 + n))$, es decir, la probabilidad de que por lo menos el 90 % de las bolillas en la urna en el paso n sean rojas:

Proposición 8.

$$P(X_n \geq 0,9 \cdot (N_0 + n)) = \begin{cases} \left(\sum_{i=0,9 \cdot (N_0+n)}^{N_0+n-1} \left(\frac{1}{2}\right)^{N_0} \sum_{j=\max\{1,i-n\}}^{\min\{i,N_0-1\}} \binom{n}{i-j} \frac{\beta(i,N_0+n-i)}{\beta(j,N_0-j)} \binom{N_0}{j} \right) \\ \quad + \left(\frac{1}{2}\right)^{N_0} \\ \quad \text{si } 0,9 \cdot (N_0 + n) \in \mathbb{N} \\ \left(\sum_{i=[0,9 \cdot (N_0+n)]+1}^{N_0+n-1} \left(\frac{1}{2}\right)^{N_0} \sum_{j=\max\{1,i-n\}}^{\min\{i,N_0-1\}} \binom{n}{i-j} \frac{\beta(i,N_0+n-i)}{\beta(j,N_0-j)} \binom{N_0}{j} \right) \\ \quad + \left(\frac{1}{2}\right)^{N_0} \\ \quad \text{si } 0,9 \cdot (N_0 + n) \notin \mathbb{N} \end{cases} \quad (3.2.11)$$

donde $[0,9 \cdot (N_0 + n)]$ representa la parte entera de $0,9 \cdot (N_0 + n)$.

Demostración. Se trata de sumar las probabilidades puntuales que corresponden al evento en cuestión. Luego se tiene

$$P(X_n \geq 0,9 \cdot (N_0 + n)) = \begin{cases} \sum_{k=0,9 \cdot (N_0+n)}^{N_0+n-1} P(X_n = k) + P(X_n = N_0 + n) \\ \quad \text{si } 0,9 \cdot (N_0 + n) \in \mathbb{N} \\ \sum_{k=[0,9 \cdot (N_0+n)]+1}^{N_0+n-1} P(X_n = k) + P(X_n = N_0 + n) \\ \quad \text{si } 0,9 \cdot (N_0 + n) \notin \mathbb{N} \end{cases} \quad (3.2.12)$$

y reemplazando a $P(X_n = k)$ por (3.2.9) se obtiene (3.2.11). \square

Fórmula asintótica de la probabilidad de que por lo menos el 90 % de las bolillas en la urna en el paso n sean rojas

Es importante considerar algunas propiedades del proceso estocástico que estamos estudiando, $\{X_n\}_{n \in \mathbb{N}}$:

En primer lugar, la esperanza de la proporción de bolillas rojas se mantendrá constante a lo largo de las generaciones. Esto se debe a que la variable aleatoria

CAPÍTULO 3. VARIANTE DEL MODELO DE URNA DE POLYA-EGGENBERGER

$\frac{X_{n+1}}{N_0+n+1}$ es martingala. En efecto, $E(X_n) < \infty$ y para cada $n \in \mathbb{N}$

$$\begin{aligned} E\left(\frac{X_{n+1}}{N_0+n+1} \mid \frac{X_n}{N_0+n}\right) &= \frac{1}{N_0+n+1} E\left(X_n + y_n \mid \frac{X_n}{N_0+n}\right) \\ &= \frac{1}{N_0+n+1} \left(E\left(X_n \mid \frac{X_n}{N_0+n}\right) + E\left(y_n \mid \frac{X_n}{N_0+n}\right) \right) \end{aligned} \quad (3.2.13)$$

$$\begin{aligned} &= \frac{1}{N_0+n+1} \left(X_n + \frac{X_n}{N_0+n} \right) \quad (3.2.14) \\ &= \frac{1}{N_0+n+1} X_n \frac{N_0+n+1}{N_0+n} \\ &= \frac{X_n}{N_0+n} \end{aligned}$$

donde las expresiones (3.2.13) y (3.2.14) son iguales porque $y_n | X_n = x_n \sim Bi(1, x_n/(N_0+n))$.

En segundo lugar, de acuerdo a lo que vimos en la Subsección 3.1.1 se tiene que $\frac{X_n}{n+N_0} | X_0 = j$ converge en distribución a una variable aleatoria β_{j, N_0-j} que tiene distribución beta con parámetros j y N_0-j , es decir, $\beta_{j, N_0-j} \sim \beta(j, N_0-j)$. A partir de este resultado vamos a poder estimar $P(X_n \geq 0,9 \cdot (N_0+n))$ cuando el número de bolillas que se agregaron es grande.

Para poder calcular el límite de $P(X_n \geq 0,9 \cdot (N_0+n))$ cuando $n \rightarrow \infty$, demostremos la siguiente proposición:

Proposición 9.

$$\lim_{n \rightarrow \infty} P\left(\frac{X_n}{n+N_0} \leq x\right) = \sum_{j=0}^{N_0} P(\beta_{j, N_0-j} \leq x) P(X_0 = j) \quad (3.2.15)$$

siendo β_{j, N_0-j} una variable aleatoria tal que $\beta_{j, N_0-j} \sim \beta(j, N_0-j)$ y donde extendemos la definición de la distribución beta a los casos extremos $j=0$ y $j=N_0$ de la siguiente manera

$$\beta_{0, N_0} \sim \beta(0, N_0) \text{ si } P(\beta_{0, N_0} = x) = \begin{cases} 1 & \text{si } x = 0 \\ 0 & \text{en caso contrario} \end{cases}$$

y

$$\beta_{N_0, 0} \sim \beta(N_0, 0) \text{ si } P(\beta_{N_0, 0} = x) = \begin{cases} 1 & \text{si } x = 1 \\ 0 & \text{en caso contrario} \end{cases}$$

3.2. VARIANTE DEL MODELO DE POLYA-EGGENBERGER

Demostración. Si inicialmente todos las bolillas son del mismo color, es decir, X_0 vale cero o N_0 , la situación inicial no va a cambiar a lo largo de las generaciones:

$$P(X_{n+1} = 0 \mid X_0 = 0) = P(X_n = 0 \mid X_0 = 0)$$

como también

$$P(X_{n+1} = N_0 \mid X_0 = N_0) = P(X_n = N_0 \mid X_0 = N_0)$$

para cada n . Por lo tanto, $\lim_{n \rightarrow \infty} P\left(\frac{X_n}{n+N_0} \leq x \mid X_0 = 0\right) = \begin{cases} 1 & \text{si } x = 0 \\ 0 & \text{en caso contrario} \end{cases}$

mientras que $\lim_{n \rightarrow \infty} P\left(\frac{X_n}{n+N_0} \leq x \mid X_0 = N_0\right) = \begin{cases} 1 & \text{si } x = 1 \\ 0 & \text{en caso contrario} \end{cases}$. Luego si extendemos la definición de la distribución beta para $j = 0$ y $j = N_0$ como se propone en el enunciado de la proposición se obtiene (3.2.16) para $j = 0$ y $j = N_0$.

La proposición utiliza ahora un resultado clásico del modelo de urna de Polya-Eggenberger ([35] y [19]) que es el siguiente:

$$\lim_{n \rightarrow \infty} P\left(\frac{X_n}{n+N_0} \leq x \mid X_0 = j\right) = P(\beta_{j, N_0-j} \leq x), \quad (3.2.16)$$

siendo β_{j, N_0-j} una variable aleatoria tal que $\beta_{j, N_0-j} \sim \beta(j, N_0 - j)$. Por la fórmula de la probabilidad total se tiene

$$P\left(\frac{X_n}{n+N_0} \leq x\right) = \sum_{j=0}^{N_0} P\left(\frac{X_n}{n+N_0} \leq x \mid X_0 = j\right) P(X_0 = j). \quad (3.2.17)$$

Luego si tomamos límite cuando $n \rightarrow \infty$ en (3.2.17), considerando (3.2.16) y el hecho de que el límite de una suma finita es igual a la suma de los límites resulta:

$$\begin{aligned} \lim_{n \rightarrow \infty} P\left(\frac{X_n}{n+N_0} \leq x\right) &= \sum_{j=0}^{N_0} \lim_{n \rightarrow \infty} P\left(\frac{X_n}{n+N_0} \leq x \mid X_0 = j\right) P(X_0 = j) \\ &= \sum_{j=0}^{N_0} P(\beta_{j, N_0-j} \leq x) P(X_0 = j) \end{aligned}$$

con $\beta_{j, N_0-j} \sim \beta(j, N_0 - j)$ como queríamos probar. □

CAPÍTULO 3. VARIANTE DEL MODELO DE URNA DE POLYA-EGGENBERGER

Entonces se tiene que $\frac{X_n}{N_0+n}$ converge a una mezcla de betas cuyos coeficientes son las probabilidades de la distribución inicial. En nuestra variante la distribución inicial es una distribución binomial con parámetros N_0 y $p = 1/2$.

Luego de la proposición se obtiene que

$$\lim_{n \rightarrow \infty} P(X_n \geq 0,9 \cdot (n+N_0)) = \sum_{j=1}^{N_0-1} P(\beta_{j,N_0-j} \geq 0,9 | X_0 = j) \binom{N_0}{j} \left(\frac{1}{2}\right)^{N_0} + \left(\frac{1}{2}\right)^{N_0} \quad (3.2.18)$$

con $\beta_{j,N_0-j} \sim \beta(j, N_0 - j)$.

Capítulo 4

Aplicaciones

4.1. Evolución por deriva en plagas

El control químico de plagas en cultivos y sus alternativas de control biológico son objeto de intenso estudio en los últimos años debido a su importancia para la salud humana y sus consecuencias sobre el medio ambiente [32]. Existe un consenso generalizado en cuanto a que para evitar el control químico indiscriminado, basado solamente en un cronograma fijo de aplicaciones periódicas de pesticidas de amplio espectro, se debe poseer un conocimiento amplio de la dinámica y la genética poblacional de las plagas y sus enemigos naturales [30]. Mediante herramientas matemáticas, esto puede conseguirse tanto a través del abordaje estadístico, por ejemplo mediante el desarrollo y aplicación de métodos de muestreo [14], como mediante modelos que buscan describir los cambios en el número de individuos y en la composición genética de las poblaciones involucradas [29].

En las poblaciones plaga hay individuos que tienen en su composición genética alelos que les confieren características que los hacen resistentes a los plaguicidas. Luego de una fumigación la mayoría de los individuos susceptibles mueren mientras que los resistentes sobreviven. A su vez los descendientes de estos últimos heredan la resistencia. De esta manera las plagas se vuelven resistentes a los plaguicidas rápidamente como consecuencia de la aplicación reiterada de pesticidas por el conocido proceso de selección. Ésta es la fuerza evolutiva más estudiada a la hora de describir la velocidad a la que se vuelven resistentes las plagas, pero no es necesariamente la única que opera.

La diversidad genética en una población de insectos o ácaros plaga cambia de

una generación a otra. Los agentes que cambian las frecuencias génicas de dichas poblaciones son la deriva genética, la mutación y la selección. En esta sección analizaremos la evolución de un alelo de una plaga, por ejemplo de insectos, bajo eventos periódicos de fumigación por deriva y mutación en ausencia de selección.

4.1.1. Deriva

Para estudiar la evolución por deriva de un alelo de una plaga con tamaño poblacional constante, bajo eventos periódicos de fumigación, utilizaremos el modelo de renovación parcial y restauración aleatoria del tamaño poblacional, que desarrollamos en la Sección 2.1, Capítulo 2 [18], porque al ser una generalización del modelo de Wright-Fisher sigue la dinámica de reproducción aleatoria más aceptada en genética de poblaciones.

Nuestra generalización del modelo de Wright-Fisher describe la evolución genética sin selección. Suponemos que una fracción de la población se elimina mediante un pesticida y luego hay una restauración del tamaño poblacional debida a la reproducción de una parte de la población restante, que encuentra condiciones favorables para su multiplicación. El proceso describe la repetición de n de estos episodios.

Los insectos pueden ser haploides o diploides. En cualquier caso el número N se refiere al número total en la población de aquellos alelos que pueden ser de tipo A o a. En el caso en que todos los individuos sean diploides, la cantidad de ellos será $N/2$, y N el tamaño poblacional de alelos. Las etapas involucradas en el paso de la generación n a la $n + 1$ luego de un episodio de fumigación son las descritas en la Sección 2.1 que ahora interpretamos para el caso de una plaga:

- 1 Se eliminan $r < N$ alelos de la generación n , elegidos aleatoriamente y sin reemplazo que corresponden a los eliminados por un pesticida.
- 2 De los $N - r$ restantes se eligen s ($0 < s \leq N - r$) alelos aleatoriamente sin reemplazo que corresponden a los alelos que participarán en la reproducción.
- 3 Se realiza un muestreo al azar con reposición de tamaño $r + s$ de los s alelos producto de la segunda etapa, que corresponde a la cantidad de alelos con los cuales se restaura el tamaño poblacional.
- 4 Finalmente la generación $n + 1$ se forma con los $r + s$ alelos producto de la tercera etapa y los $N - r - s$ alelos que restan.

4.1. EVOLUCIÓN POR DERIVA EN PLAGAS

Según las etapas 1 y 3 tenemos que el parámetro r representa la intensidad del efecto del pesticida y el valor de s indica cuántos de los insectos sobrevivientes van a reestablecer la población, donde s no es el número de insectos sino el número de alelos. Como vimos en la Sección 2.1 la población pasará en el largo plazo, pero en un número finito de generaciones, a estar formada por todos alelos A o a, que son los estados absorbentes del proceso de Markov.

Tamaño efectivo de la población y número efectivo de generaciones.

En particular nos interesa determinar la tasa de cambio de la diversidad por deriva, que se puede describir en función de la pérdida de la frecuencia de alelos A de la población. Según las subsecciones 2.1.3 y 2.1.4 del Capítulo 2, las fórmulas para la heterocigosidad y el correspondiente número poblacional efectivo son:

$$h(n) = h(0) \left(1 - \frac{(r+s)(r+s-1)}{sN(N-1)} \right)^n \quad (2.1.24)$$

y

$$N_e^h = N \frac{s(N-1)}{(r+s)(r+s-1)} \quad (2.1.26).$$

A partir de la expresión del N_e^h se puede reescribir la heterocigosidad de la siguiente manera

$$h(n) = h(0) \left(1 - \frac{1}{N_e^h} \right)^n \quad (2.1.28).$$

Mientras que el número efectivo de generaciones es:

$$n_e = \frac{\ln \left(1 - \frac{1}{N} \right)}{\ln \left(1 - \frac{(r+s)(r+s-1)}{sN(N-1)} \right)} \quad (2.1.33).$$

A partir de estas expresiones estudiaremos para qué valores de los parámetros s y r la deriva debería ser un factor a tener en cuenta cuando se estudia el cambio de la frecuencia de un alelo de una plaga bajo eventos periódicos de fumigación.

Deriva en la evolución de un alelo de una plaga

La deriva genética es una fuerza evolutiva que afecta a las poblaciones pequeñas. Si bien las poblaciones de insectos son grandes, en el modelo que proponemos en cada generación sólo una fracción de individuos contribuye con sus genes

a la siguiente. Por lo tanto, la tasa de cambio por deriva y la pérdida de heterocigosidad podrían ser mayores que las esperadas en función del tamaño poblacional. En dicho caso, desde un punto de vista efectivo, la población podría evolucionar como si fuera más pequeña de lo que es, motivando así el estudio de la deriva como fuerza evolutiva que interviene en la velocidad a la que se extienden o desaparecen los genes resistentes a los pesticidas [29].

Los alelos de la generación $n+1$ provienen de una muestra de los alelos presentes en la generación anterior, ya que, de los individuos de la población no alcanzados por el pesticida en una fumigación, una parte se mantiene y otra participa en la restauración del tamaño poblacional. El cambio por deriva en las frecuencias alélicas se produce porque un tipo de alelo que por azar deja más descendencia aumentará su frecuencia en la población. El resultado a largo plazo es la fijación de uno de los dos tipos de alelos en la población.

Tamaño poblacional efectivo

La tasa a la cual se pierde la diversidad en la población está determinada por el tamaño efectivo. En función de esto último comparamos el tamaño efectivo (N_e^h) y el tamaño de la población (N) graficando la proporción de alelos que se eliminan por el uso de un pesticida ($r' = r/N$) en función de la proporción de alelos que se reproducen ($s' = s/N$) para los cuales $N_e^h = kN$ para $0 < k < 1$ (ver en la Figura 2.4, Subsección 2.1.4, Capítulo 2). Son de interés aquellos valores de parámetros para los cuales $N_e^h < N$ dado que la deriva podría ser significativa a la hora de explicar un cambio en la frecuencia de alelos en una población plaga.

Cuello de botella

Consideremos un evento de fumigación en el que los individuos que mueren y los que se reproducen son aproximadamente N , es decir, $r + s \approx N$. Sabemos que los r' y s' que satisfacen $N_e^h = kN$ donde $0 < k < 1$ cumplen, según (2.1.39)

$$\frac{s'(1 - 1/N)}{(r' + s')(r' + s' - 1/N)} = k.$$

Ahora si reemplazamos $r' + s'$ por uno y $1/N$ por cero en esta última expresión tenemos que $s' \approx k$. Entonces el tamaño poblacional efectivo es $1/2N$ o $1/2^2N$ o $1/2^3N$ o $1/2^4N$ si s' es $1/2$ o $1/2^2$ o $1/2^3$ o $1/2^4$, respectivamente (ver en la Figura

4.1. EVOLUCIÓN POR DERIVA EN PLAGAS

2.4).

Luego si $r' + s' = 1$ para s' pequeño (el mínimo valor que puede tomar s' es $1/N$ porque 1 es el valor mínimo para s) hay mucha deriva porque en el paso de una generación a la siguiente se renueva toda o la mayor parte de la población a partir de una pequeña parte que sobrevive a un evento de fumigación. Esto corresponde a lo que en genética de poblaciones se denomina atravesar un cuello de botella.

Deriva en ausencia de un evento de cuello de botella

Podemos preguntarnos si es posible que se produzca una deriva similar a la de un cuello de botella, aún si sobrevive sin cambios una parte de la población. De acuerdo con lo desarrollado en la Subsección 2.1.4, Capítulo 2, para este modelo de renovación parcial aleatoria los cuellos de botella no son los únicos eventos que podrían cambiar las frecuencias de alelos por puro azar en las poblaciones plaga. Consideremos los ejemplos del siguiente cuadro

	$r' + s' = 1/2$	$r' + s' = 1/4$
$N_e^h = 1/16N$	$s' \approx 1/N$	$s' < 1/N$
$N_e^h = 1/8N$	$s' \approx 1/N$	$s' < 1/N$
$N_e^h = 1/4N$	$s' \approx 1/N$	$s' \approx 1/N$

Cuadro 4.1: Situaciones de deriva significativa. Valores de s' para los cuales el tamaño efectivo resulta $N_e^h = 1/2^l N$ para $l = 2, 3, 4$ y $r' + s' = 1/2$ o $r' + s' = 1/4$.

En este cuadro se observa que

- Si de la mitad de la población que fue afectada por el pesticida, la mayoría murió como consecuencia de la fumigación y los que sobrevivieron son los que restauraron el tamaño poblacional, entonces la deriva será significativa para aquellos valores de s' tales que s' es cercano a $1/N$.
- Si las tres cuartas partes de la población ($1 - (r' + s') = 3/4$) permance sin cambios a pesar del evento de fumigación y $s' \approx 1/N$, el tamaño efectivo se puede reducir de tal manera que sea $1/4$ del tamaño poblacional. Sin embargo, dado que s' está acotado inferiormente por $1/N$ el tamaño efectivo de la población no se puede reducir de tal manera que represente $1/8$ o $1/16$ de N .

Número efectivo de generaciones

La importancia del número efectivo de generaciones reside en que representa la cantidad de eventos de fumigación que producen la misma deriva que una fumigación única si el proceso fuera el clásico de Wright-Fisher.

Retomemos la expresión (2.1.33), que hallamos para n_e , estandarizemos el número alelos que van a ser reemplazados y el número de alelos elegidos para reproducirse mediante $s' = s/N$, $r' = r/N$ y reemplacemos x por $1/N$. Entonces obtenemos

$$n_e = \frac{\ln(1-x)}{\ln\left(1 - \frac{(r'+s')(r'+s'-x)}{s'} \frac{x}{1-x}\right)}. \quad (4.1.1)$$

Nos interesa ver qué ocurre en poblaciones grandes, es decir, cuando $N \rightarrow \infty$, que es equivalente a estudiar el límite cuando $x \rightarrow 0$:

Luego

$$\begin{aligned} \lim_{x \rightarrow 0} n_e &= \lim_{x \rightarrow 0} \frac{\ln(1-x)}{\ln\left(1 - \frac{(r'+s')(r'+s'-x)}{s'} \frac{x}{1-x}\right)} \\ &= \lim_{x \rightarrow 0} \frac{\frac{-1}{1-x}}{\frac{1}{1 - \frac{(r'+s')(r'+s'-x)}{s'} \frac{x}{1-x}} \left(- \left(\frac{(r'+s')(-1)}{s'} \frac{x}{1-x} + \frac{(r'+s')(r'+s'-x)}{s'} \frac{1}{(1-x)^2} \right) \right)}. \end{aligned} \quad (4.1.2)$$

$$(4.1.3)$$

Las expresiones (4.1.2) y (4.1.3) son iguales por la regla de L'Hopital.

Operando algebraicamente podemos reescribir a (4.1.3) como:

$$\frac{1 - \frac{(r'+s')(r'+s'-x)}{s'} x}{\frac{(r'+s')(-1)}{s'} x(1-x) + \frac{(r'+s')(r'+s'-x)}{s'}}.$$

Finalmente si reemplazamos esta última expresión en (4.1.3) obtenemos

$$\begin{aligned} \lim_{x \rightarrow 0} n_e &= \lim_{x \rightarrow 0} \frac{1 - \frac{(r'+s')(r'+s'-x)}{s'} x}{\frac{(r'+s')(-1)}{s'} x(1-x) + \frac{(r'+s')(r'+s'-x)}{s'}} \\ &= \frac{s'}{(r'+s')^2}. \end{aligned} \quad (4.1.4)$$

Notemos además que esta expresión es aproximadamente N_e^h/N . En efecto, pa-

4.1. EVOLUCIÓN POR DERIVA EN PLAGAS

ra N suficientemente grande resulta $N_e^h \approx s/(r' + s')^2$ según (2.1.30) (ver Subsección 2.1.4), entonces basta con reescribir $s'/(r' + s')^2$ como $s'/(r' + s')^2 N = N_e^h/N$.

Número efectivo de generaciones y deriva

Para la aplicación que estamos estudiando son de interés aquellas combinaciones de s' y r' tales que dan lugar a mucha deriva, es decir, tales que $n_e < 1$ (que a su vez implica que $N_e^h < N$).

- Los s' y r' que satisfacen $N_e^h = kN$ para k entre 0 y 1 cumplen $n_e = k$ porque $n_e = N_e^h/N$. Luego cuanto menor sea el valor de k ($0 < k < 1$) mayor va a ser la deriva. Al igual que ocurre con el tamaño efectivo, el número efectivo de generaciones no se podrá hacer arbitrariamente pequeño dado que en la práctica s está acotado inferiormente por 1.
- Si igualamos n_e a 1, nos ubicamos en una deriva igual a la de Wright-Fisher en cada generación y obtenemos la misma curva de la Figura 2.3 para la relación entre r' y s' tal que $N_e^h = N$.
- Si $r' + s' = 1$ vemos que $n_e = s'$. Si s' es pequeño estamos en una situación de cuello de botella y si s' toma su valor máximo de 1, estamos en el modelo de Wright-Fisher.

Comportamiento no monótono de la deriva para una proporción fija de individuos eliminados

Si fijamos un valor de $r' \in (0, 1)$ tenemos que el número efectivo de generaciones es una función de s' definida en el intervalo $(0, 1 - r')$ tal que $n_{e,r'}(s') = s'/(r' + s')^2$.

Calculemos la primera derivada de la función $n_{e,r'}$:

$$\begin{aligned} n_{e,r'}^{(1)}(s') &= \frac{(r' + s')^2 - 2s'(r' + s')}{(r' + s')^4} \\ &= \frac{r'^2 - s'^2}{(r' + s')^4}. \end{aligned}$$

De donde tenemos que $n_{e,r'}^{(1)}(s')$ toma el signo de $r' - s'$, dado que $s' \in (0, 1 - r')$ y $r' \in (0, 1)$.

CAPÍTULO 4. APLICACIONES

Luego si $r' \geq 1/2$, $n_{e,r'}$ es una función creciente de s' en todo su intervalo de definición $(0, 1 - r')$. Pero si $r' < 1/2$, tenemos que es creciente hasta r' , alcanza un máximo en r' y decrece entre r' y $1 - r'$.

Entonces en la medida que aumenta s' aumenta el número efectivo de generaciones, es decir, hay menos deriva, pero si $s' > r'$ (ya que si $s' = r'$ la función alcanza el máximo) el número efectivo disminuye y aumenta la deriva.

Tamaño efectivo en distintos escenarios de renovación parcial

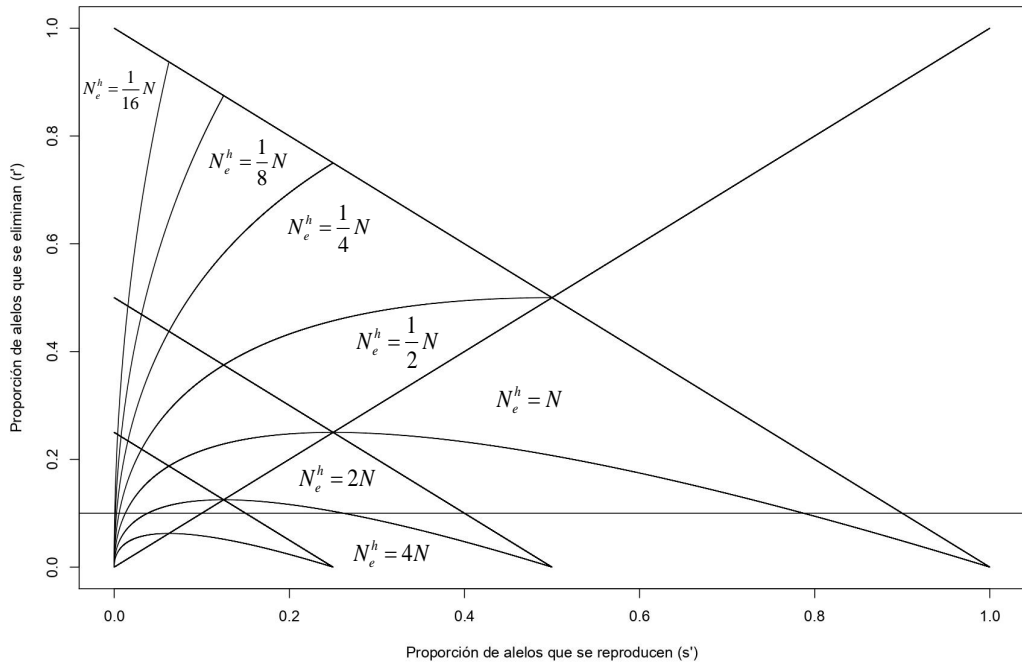


Figura 4.1: Gráfico de la proporción de alelos que se eliminan en función de la proporción de los que se reproducen bajo la condición $N_e^h = kN$ para $k = 2^2, 2, 1, 1/2, 1/2^2, 1/2^3, 1/2^4$. Además graficamos las rectas $r' = 1 - s'$, $r' = 1/2 - s'$, $r' = 1/4 - s'$, $r' = s'$ y la recta horizontal $r' = 0,1$.

Este comportamiento para $r < 1/2$ se puede observar en la Figura 4.1 mediante la recta horizontal $r' = 0,1$ que se interseca con las curvas r'_k para $k = 1/2$, $k = 1$, $k = 2$ y nuevamente para $k = 1$ conforme aumenta el valor de s' . Luego según el valor que tome s' se tiene que $N_e^h < N$, $N_e^h = N$ o $N_e^h > N$. En este caso particular, como r' es una fracción pequeña del total, cuando s' está próximo a

$1 - r'$ el modelo se va pareciendo cada vez más al modelo de Wright-Fisher y es por eso que el tamaño poblacional es aproximadamente igual al tamaño poblacional efectivo.

A partir del modelo propuesto, se puede concluir que la deriva puede ser significativa según el valor de los parámetros que controlan el número de alelos que mueren (r) y el número de alelos que sobreviven y se reproducen (s) luego de cada episodio de fumigación. Por lo tanto, debe ser un factor a tener en cuenta cuando se estudia la evolución de un alelo, incluso para aquellos alelos que confieren resistencia a las plagas. Sólo se encontrarán formas de controlar la aparición de resistencias si se comprende la teoría evolutiva y se conoce cómo influyen las soluciones a corto plazo en la evolución de las poblaciones de insectos ([30] y [34]).

4.1.2. Mutación

Continuando con la posibilidad de evolución sin selección, debemos considerar que adicionalmente a la deriva, los alelos pueden estar afectados por mutación. Por este motivo en esta subsección estudiaremos el modelo de deriva y restauración aleatoria del tamaño poblacional incorporando mutación. Los alelos de tipo A podrán mutar a los de tipo a, y los alelos de tipo a podrán mutar a los alelos de tipo A, en ambos casos con una probabilidad de mutación asociada.

Si bien, según estudiamos en el Capítulo 1, la deriva a largo plazo tiende a eliminar la variación genética, las mutaciones contrarrestan dicho efecto. En particular, deja de haber estados absorbentes porque la posibilidad de mutación impide la fijación de un alelo.

Nuevamente tenemos una población de alelos de tamaño N con dos variantes alélicas, A y a. En este modelo de deriva con renovación parcial se tiene que la generación $n + 1$ se obtiene de la generación n a partir de un muestreo aleatorio, por el cual $r < N$ alelos de la generación n , elegidos aleatoriamente y sin reemplazo serán removidos. De los $N - r$ restantes se eligen $0 < s \leq N - r$ alelos aleatoriamente y sin reemplazo, a partir de los cuales se va a restaurar el tamaño original de la población. Para incluir mutación consideramos que los alelos de tipo a pueden mutar a A con probabilidad u y los de tipo A pueden mutar a a con probabilidad v , siendo $0 < u < 1$ y $0 < v < 1$. Por lo tanto, en la generación $n + 1$ si un alelo elegido al azar es de tipo A puede provenir de la generación anterior de un alelo de tipo A o puede provenir de un alelo de tipo a, que mutó a A. Las mutaciones

CAPÍTULO 4. APLICACIONES

ocurren de modo independiente, es decir, que el hecho de que un alelo haya mutado no cambia la probabilidad de que otro alelo cualquiera lo haga.

Como lo hicimos en el Capítulo 2, Sección 2.1 describiremos la etapa de renovación parcial, pero agregando mutación. Para ello definiremos el muestreo con reposición de tamaño $r + s$ de los s alelos utilizando los siguientes eventos:

A : El tipo del alelo tomado de los s alelos para pasar a la nueva generación es A .

M : Hubo mutación del alelo tomado de los s .

A' : El tipo de alelo que pasa efectivamente a la siguiente generación después de la eventual mutación es A .

Siendo Y_n el número de alelos de tipo A entre los s en la generación n , tendremos las siguientes probabilidades según el valor de Y_n : $P(A) = \frac{Y_n}{s}$ y $P(A') = \frac{Y_n}{s}(1 - v) + \left(\frac{s - Y_n}{s}\right)u$. Esta última fórmula pasa a ser el parámetro de la distribución binomial condicional a Y_n de alelos A en los $r + s$ de la nueva generación y se obtiene calculando la probabilidad total mediante:

$$\begin{aligned} P(A') &= P(A'AM) + P(A'A^cM) + P(A'AM^c) + P(A'A^cM^c) \\ &= P(A' | AM)P(M | A)P(A) + P(A' | A^cM)P(M | A^c)P(A^c) \\ &\quad + P(A' | AM^c)P(M^c | A)P(A) + P(A' | A^cM^c)P(M^c | A^c)P(A^c) \\ &= 0 + 1 \cdot u \cdot \left(\frac{s - Y_n}{s}\right) + 1 \cdot (1 - v) \frac{Y_n}{s} + 0 \\ &= \frac{Y_n}{s}(1 - v) + \left(\frac{s - Y_n}{s}\right)u \end{aligned}$$

El término $\frac{Y_n}{s}(1 - v)$ corresponde a la probabilidad de que se elija un alelo de tipo A y que no mute a a , mientras que $\left(\frac{s - Y_n}{s}\right)u$ corresponde a la probabilidad de que se elija un alelo de tipo a y que mute a A .

Se renuevan en cada generación $r + s$ alelos y $N - r - s$ se mantienen sin cambios. Luego, la generación $n + 1$ se forma con los $r + s$ alelos producto de la etapa de expansión aleatoria y los $N - r - s$ alelos que restan.

Si se define X_n como el número de alelos de tipo A en la generación n , entonces $\{X_n, n = 0, 1, \dots\}$ es una cadena de Markov homogénea con espacio de estados $E = \{0, 1, \dots, N\}$. Veremos que es ergódica y por lo tanto tiene una única distribución estacionaria $\pi = (\pi(0), \pi(1), \dots, \pi(N))$ tal que $\lim_{n \rightarrow \infty} P(X_n = i) = \pi(i)$ para

todo $i \in E$.

Proposición 10. *Existe $n \geq 1$ tal que para cualquier par de estados i y j se tiene $P(X_n = j | X_0 = i) > 0$.*

Demostración. *Basta probar que las siguientes transiciones en un paso tienen probabilidad positiva: (1) $i \rightarrow i$ para todo $i \in E$, (2) $i \rightarrow i + 1$ para todo $i \neq N$ y (3) $i \rightarrow i - 1$ para todo $i \neq 0$. Para $i \rightarrow i$ tenemos que un evento de probabilidad positiva incluido en $\{X_1 = i | X_0 = i\}$ es aquel en el que el muestreo con reposición reproduce los mismos tipos de alelos de los $r + s$ individuos originales y no se produce mutación alguna. La transición $i \rightarrow i + 1$ se da con probabilidad positiva si ese muestreo repite los mismos tipos de alelos y además se produce exactamente una mutación, la cual es de un alelo de tipo a al tipo A . Análogamente, la transición $i \rightarrow i - 1$ se da con probabilidad positiva si se producen los mismos tipos de alelos y además hay exactamente una mutación, la cual es del tipo A al a .*

Al igual que lo hicimos en los Capítulos 1 y 2 suponemos que el número de alelos A en la generación cero, X_0 , no es aleatorio, y por lo tanto se tiene que $E(X_0) = X_0$ como también que la varianza de X_0 es cero, $Var(X_0) = 0$.

Esperanza de la distribución de equilibrio

Calculemos la $E(X_{n+1})$ para luego hallar la esperanza de la distribución estacionaria.

Proposición 11.

$$E(X_{n+1}) = \left(1 - \frac{(r+s)(u+v)}{N}\right)^{n+1} \left(X_0 - \frac{Nu}{u+v}\right) + \frac{Nu}{u+v} \quad (4.1.5)$$

Demostración. *La expansión de los alelos que restauran el tamaño original de la población es aleatoria y se renuevan en cada generación $r + s$ alelos y los $N - r - s$ restantes quedan como estaban. Entonces para calcular la $E(X_{n+1})$ escribimos a X_{n+1} como la suma de dos variables aleatorias:*

$$X_{n+1} = X_{n+1}^{r+s} + X_{n+1}^{N-r-s} \quad (4.1.6)$$

donde X_{n+1}^{r+s} se define como el número de alelos A entre los $r + s$ alelos producto de la renovación parcial con expansión de los s alelos y X_{n+1}^{N-r-s} como el número de alelos A entre los $N - r - s$ alelos restantes.

CAPÍTULO 4. APLICACIONES

Consideremos ahora la variable aleatoria Y_n , el número de alelos A entre los s , y su distribución condicional a $X_n = i$. Como la elección de los s es a partir de un muestreo aleatorio y sin reemplazo de una población de N alelos de dos tipos, resulta

$$Y_n | X_n = i \sim H(N, i, s), \quad (4.1.7)$$

es decir, tiene una distribución hipergeométrica. Además en el muestreo con reemplazo de tamaño $r + s$ tenemos

$$X_{n+1}^{(r+s)} | Y_n \sim Bi\left(r + s, \frac{Y_n}{s}(1 - v) + \left(\frac{s - Y_n}{s}\right)u\right). \quad (4.1.8)$$

Veamos que

$$E_i(X_{n+1}^{r+s}) = (r + s) \left(\frac{i}{N}(1 - v) + \left(\frac{N - i}{N}\right)u \right). \quad (4.1.9)$$

El subíndice i indica que condicionamos respecto a $X_n = i$.

Primero calculemos $E(X_{n+1}^{r+s} | Y_n)$:

Si consideramos (4.1.8) tenemos que

$$E(X_{n+1}^{r+s} | Y_n) = (r + s) \left(\frac{Y_n}{s}(1 - v) + \left(\frac{s - Y_n}{s}\right)u \right). \quad (4.1.10)$$

Luego tomando esperanza en esta última expresión obtenemos:

$$\begin{aligned} E_i(X_{n+1}^{r+s}) &= E_i(E(X_{n+1}^{r+s} | Y_n)) \\ &= E_i\left((r + s) \left(\frac{Y_n}{s}(1 - v) + \left(\frac{s - Y_n}{s}\right)u \right)\right) \quad \text{por (4.1.10)} \\ &= (r + s) \left(\frac{E_i(Y_n)}{s}(1 - v) + \left(\frac{s - E_i(Y_n)}{s}\right)u \right) \\ &= (r + s) \left(\frac{s \frac{i}{N}}{s}(1 - v) + \left(\frac{s - s \frac{i}{N}}{s}\right)u \right) \quad \text{por (4.1.7)} \\ &= (r + s) \left(\frac{i}{N}(1 - v) + \left(\frac{N - i}{N}\right)u \right) \end{aligned} \quad (4.1.11)$$

que es la expresión de (4.1.9) que buscábamos.

4.1. EVOLUCIÓN POR DERIVA EN PLAGAS

Ahora calculemos $E_i(X_{n+1}^{N-r-s})$.

$$X_{n+1}^{N-r-s} | Y_n \sim H(N-s, i-Y_n, N-r-s), \quad (4.1.12)$$

es una distribución hipergeométrica porque la elección de los $N-r-s$ alelos que permanecen sin cambios en el paso de una generación a la siguiente es a partir de un muestreo aleatorio y sin reemplazo de una población de $N-s$ alelos, que son de tipo a o A .

Luego

$$\begin{aligned} E_i(X_{n+1}^{N-r-s}) &= E_i\left(E\left(X_{n+1}^{N-r-s} | Y_n\right)\right) \\ &= E_i\left((N-r-s) \frac{i-Y_n}{N-s}\right) \quad \text{por (4.1.12)} \\ &= \frac{(N-r-s)}{N-s} (i - E_i(Y_n)) \\ &= \frac{(N-r-s)}{N-s} \left(i - s \frac{i}{N}\right) \quad \text{por (4.1.7)} \\ &= \frac{(N-r-s)}{N-s} i \left(\frac{N-s}{N}\right) \\ &= (N-r-s) \frac{i}{N}. \end{aligned}$$

Así tenemos que

$$E_i(X_{n+1}^{N-r-s}) = (N-r-s) \frac{i}{N}. \quad (4.1.13)$$

Finalmente combinando (4.1.11), (4.1.13) y (4.1.6) se tiene

$$\begin{aligned} E_i(X_{n+1}) &= E_i\left(X_{n+1}^{N-r-s} + X_{n+1}^{r+s}\right) \\ &= E_i\left(X_{n+1}^{N-r-s}\right) + E_i\left(X_{n+1}^{r+s}\right) \\ &= (N-r-s) \frac{i}{N} + (r+s) \left(\frac{i}{N}(1-v) + \left(\frac{N-i}{N}\right)u\right) \quad (4.1.14) \\ &= \frac{i}{N} (N - (u+v)(r+s)) + (r+s)u. \end{aligned}$$

Luego

$$E(X_{n+1}|X_n) = \frac{X_n}{N} (N - (u+v)(r+s)) + (r+s)u. \quad (4.1.15)$$

CAPÍTULO 4. APLICACIONES

Si tomamos esperanza en (4.1.15) obtenemos:

$$\begin{aligned}
 E(X_{n+1}) &= E(E(X_{n+1}|X_n)) \\
 &= E\left(\frac{X_n}{N}(N - (u+v)(r+s)) + (r+s)u\right) \\
 &= \frac{(N - (u+v)(r+s))}{N}E(X_n) + (r+s)u. \tag{4.1.16}
 \end{aligned}$$

Por último consideremos la fórmula recursiva (4.1.16) para obtener $E(X_{n+1})$:

$$\begin{aligned}
 E(X_{n+1}) &= \frac{(N - (u+v)(r+s))}{N} \left(\frac{(N - (u+v)(r+s))}{N} E(X_{n-1}) + (r+s)u \right) \\
 &\quad + (r+s)u \\
 &= \left(\frac{(N - (u+v)(r+s))}{N} \right)^2 E(X_{n-1}) \\
 &\quad + \left(\frac{(N - (u+v)(r+s))}{N} \right) (r+s)u + (r+s)u \\
 &= \left(\frac{(N - (u+v)(r+s))}{N} \right)^3 E(X_{n-2}) \\
 &\quad + \left(\frac{(N - (u+v)(r+s))}{N} \right)^2 (r+s)u \\
 &\quad + \left(\frac{(N - (u+v)(r+s))}{N} \right) (r+s)u + (r+s)u \\
 &= \dots \\
 &= \left(\frac{(N - (u+v)(r+s))}{N} \right)^{n+1} E(X_0) + (r+s)u \sum_{l=0}^n \left(1 - \frac{(u+v)(r+s)}{N} \right)^l \\
 &\tag{4.1.17} \\
 &= \left(\frac{(N - (u+v)(r+s))}{N} \right)^{n+1} X_0 + (r+s)u \left(\frac{1 - \left(\frac{(N - (u+v)(r+s))}{N} \right)^{n+1}}{1 - \left(1 - \frac{(u+v)(r+s)}{N} \right)} \right) \\
 &\tag{4.1.18} \\
 &= \left(\frac{(N - (u+v)(r+s))}{N} \right)^{n+1} X_0 + \frac{Nu}{(u+v)} \left(1 - \left(\frac{(N - (u+v)(r+s))}{N} \right)^{n+1} \right) \\
 &= \left(1 - \frac{(u+v)(r+s)}{N} \right)^{n+1} \left(X_0 - \frac{Nu}{(u+v)} \right) + \frac{Nu}{(u+v)},
 \end{aligned}$$

4.1. EVOLUCIÓN POR DERIVA EN PLAGAS

donde (4.1.18) se obtiene si reemplazamos $E(X_0)$ por X_0 y la serie

$$\sum_{l=0}^n \left(1 - \frac{(u+v)(r+s)}{N}\right)^l \text{ por } \left(\frac{1 - \left(\frac{(N-(u+v)(r+s))}{N}\right)^{n+1}}{1 - \left(1 - \frac{(u+v)(r+s)}{N}\right)}\right).$$

Es decir que obtuvimos

$$E(X_{n+1}) = \left(1 - \frac{(u+v)(r+s)}{N}\right)^{n+1} \left(X_0 - \frac{Nu}{(u+v)}\right) + \frac{Nu}{(u+v)}$$

como queríamos demostrar.

Ahora calculemos la esperanza de la distribución límite:

Proposición 12. *La esperanza de la distribución límite de la cadena de Markov, que denotaremos por μ , es*

$$\mu = N \frac{u}{u+v}. \quad (4.1.19)$$

Demostración. *Para calcular μ primero veamos que*

$$\mu = \lim_{n \rightarrow \infty} E(X_n) \quad (4.1.20)$$

En efecto,

$$\begin{aligned} \lim_{n \rightarrow \infty} E(X_n) &= \lim_{n \rightarrow \infty} \sum_{j=0}^N j P(X_n = j) \\ &= \sum_{j=0}^N j \lim_{n \rightarrow \infty} P(X_n = j) \end{aligned} \quad (4.1.21)$$

$$= \sum_{j=0}^N j \pi(j) \quad (4.1.22)$$

$$= \mu.$$

Las expresiones (4.1.21) y (4.1.22) son iguales porque $\{X_n, n = 0, 1, \dots\}$ es irreducible y aperiódica con un número finito de estados. Entonces tiene una única distribución estacionaria $\pi = (\pi(0), \pi(1), \dots, \pi(N))$ tal que $\lim_{n \rightarrow \infty} P(X_n = j) = \pi(j)$.

Ahora probemos (4.1.19):

La fórmula del número esperado de alelos de tipo A en la generación n es:

$$E(X_n) = \left(1 - \frac{(r+s)(u+v)}{N}\right)^n \left(X_0 - \frac{Nu}{u+v}\right) + \frac{Nu}{u+v} \quad (4.1.5).$$

CAPÍTULO 4. APLICACIONES

Si tomamos límite en esta última expresión resulta:

$$\lim_{n \rightarrow \infty} E(X_n) = \frac{Nu}{u+v} + \lim_{n \rightarrow \infty} \left(1 - \frac{(r+s)(u+v)}{N}\right)^n \left(X_0 - \frac{Nu}{u+v}\right) \quad (4.1.23)$$

$$= N \frac{u}{u+v} \quad (4.1.24)$$

como queríamos probar. Las expresiones (4.1.23) y (4.1.24) son iguales porque $-1 < \left(1 - \frac{(r+s)(u+v)}{N}\right) < 1$, luego $\lim_{n \rightarrow \infty} \left(1 - \frac{(r+s)(u+v)}{N}\right)^n = 0$. En efecto, como $0 < u+v < 2$ y $0 < (r+s)/N \leq 1$ se tiene que $0 < (u+v)(r+s)/N < 2$, lo cual implica que $-1 < \left(1 - \frac{(r+s)(u+v)}{N}\right) < 1$.

Alternativamente, también se llega a esta expresión tomando límite en la fórmula recursiva (4.1.16) que obtuvimos para la esperanza. Consideremos

$$E(X_{n+1}) = \frac{(N - (u+v)(r+s))}{N} E(X_n) + (r+s)u \quad (4.1.16).$$

Luego si tomamos límite en ambos miembros de esta última expresión resulta:

$$\begin{aligned} \lim_{n \rightarrow \infty} E(X_{n+1}) &= \frac{(N - (u+v)(r+s))}{N} \lim_{n \rightarrow \infty} E(X_n) + (r+s)u \\ \mu &= \frac{(N - (u+v)(r+s))}{N} \mu + (r+s)u \quad \text{por (4.1.20)}. \end{aligned}$$

Finalmente despejamos μ y obtenemos $\mu = N \frac{u}{u+v}$.

Evolución de la proporción esperada

La fórmula del número esperado de alelos de tipo A en la generación n es:

$$E(X_n) = \left(1 - \frac{(r+s)(u+v)}{N}\right)^n \left(X_0 - \frac{Nu}{u+v}\right) + \frac{Nu}{u+v} \quad (4.1.5).$$

En este modelo de renovación parcial con mutación la esperanza del número de alelos A no es constante como cuando sólo actúa la deriva. Estudiemos la variación por mutación a partir de (4.1.5) y (4.1.16).

La dinámica de la convergencia se puede ilustrar mediante la fórmula recursiva que obtuvimos para el número esperado de alelos de tipo A en la generación $n+1$:

4.1. EVOLUCIÓN POR DERIVA EN PLAGAS

$$E(X_{n+1}) = \frac{(N - (u + v)(r + s))}{N} E(X_n) + (r + s)u \quad (4.1.16).$$

Reordenando resulta:

$$\frac{E(X_{n+1})}{N} = \frac{E(X_n)}{N} - \frac{E(X_n)}{N} \frac{(r + s)}{N} v + \left(1 - \frac{E(X_n)}{N}\right) \frac{(r + s)}{N} u.$$

En base a esta expresión definimos:

$$f_{r+s} : [0, 1] \longrightarrow [0, 1] \text{ con } f_{r+s}(x) = x - x \frac{(r + s)}{N} v + (1 - x) \frac{(r + s)}{N} u \quad (4.1.25)$$

donde $x = \frac{E(X_n)}{N}$.

La evolución de la proporción esperada puede verse gráficamente en la Figura 4.2. En esta figura se representa la recta identidad y la función f_{r+s} que corresponde a la fórmula recursiva de las proporciones esperadas. Si se parte de x_0 se obtiene el valor $x_1 = f_{r+s}(x_0)$ en el cual hay que volver a evaluar la función para obtener $f_{r+s}(x_1) = x_2$ y así sucesivamente para converger a μ . Dado que el punto en el cual la horizontal que pasa por $(x_0, f_{r+s}(x_0))$ corta la identidad tiene abscisa y ordenada $x_1 = f_{r+s}(x_0)$, podemos evaluar la función en x_1 trazando una vertical desde (x_1, x_1) hasta la gráfica de f_{r+s} para obtener x_2 y así sucesivamente. La trayectoria escalonada muestra esta evolución, que es una convergencia al equilibrio ubicado en la intersección de la gráfica de f_{r+s} y la identidad. Análogamente, si se parte de un x'_0 mayor que el punto de equilibrio, se converge a $\mu/N = \frac{u}{u+v}$ desde valores mayores.

Analicemos qué sucede en los casos extremos, es decir, cuando todos los alelos son de tipo A o todos de tipo a. Observemos que si $x = 0$ entonces la proporción esperada de alelos de tipo A en la siguiente generación es $(r + s)u/N$, es decir, se corresponde con la fracción de alelos que mutan de tipo a a A en el paso de una generación a la siguiente. Mientras que si $x = 1$ la proporción esperada de alelos de tipo A en la siguiente generación es $1 - (r + s)v/N$ porque en el paso de una generación a la siguiente la fracción de alelos que mutan de tipo A a a es $(r + s)v/N$.

Por otro lado tenemos que con la población en el estado de equilibrio ($f_{r+s}(x) = x$) los términos $x \frac{(r+s)}{N} v$ y $(1 - x) \frac{(r+s)}{N} u$ son iguales, de donde resulta que la proporción esperada de alelos de tipo A en el equilibrio es $\mu/N = \frac{u}{u+v}$, como vimos

CAPÍTULO 4. APLICACIONES

en (4.1.19).

Concluimos que el equilibrio de proporciones esperadas de alelos de tipo A se alcanza en $u/(u+v)$, que no depende de los parámetros r y s . Éstos últimos sí tienen incidencia en la velocidad de convergencia al equilibrio, al influir en el valor de la tasa $1 - \frac{(r+s)(u+v)}{N}$ de la fórmula cerrada (4.1.16), la cual coincide con la pendiente de la recta f_{r+s} , que es $f_{r+s}(1) - f_{r+s}(0) = 1 - \frac{(r+s)(u+v)}{N}$. Cuanto mayor es $r+s$, menor es la pendiente de la recta y mayor es la velocidad de convergencia al equilibrio. Esto se debe a que $r+s$ es el número de alelos que pueden mutar en cada generación.

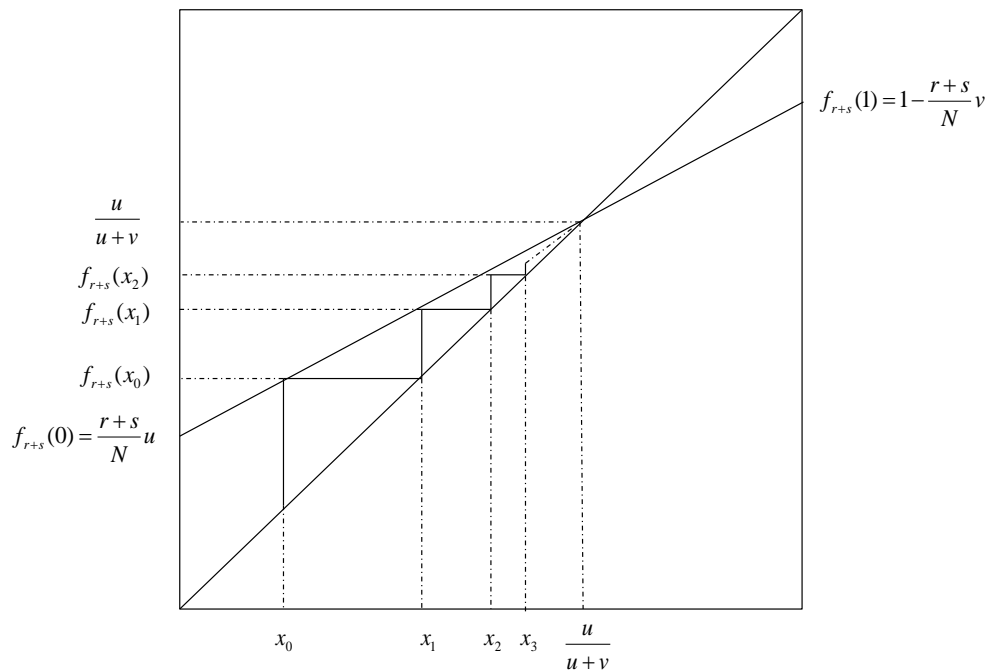


Figura 4.2: Convergencia de los valores de las proporciones esperadas de alelos de tipo A a la proporción esperada de la distribución de equilibrio para $1 - \frac{(r+s)(u+v)}{N} > 0$.

Finalmente señalamos que la tasa $1 - \frac{(r+s)(u+v)}{N}$ está acotada por -1 y 1 , pu-

4.1. EVOLUCIÓN POR DERIVA EN PLAGAS

diendo ser negativa. En este último caso se produciría un comportamiento oscilante alrededor del punto de equilibrio en la convergencia de la sucesión de esperanzas como se puede ver en la Figura 4.3. Un valor negativo requeriría que $u + v$ sea mayor que $N/(r + s)$, que es a su vez mayor que 1. Esto es posible en teoría porque u y v son probabilidades condicionadas respecto de distintos eventos (que el alelo sea de tipo a y que el alelo sea de tipo A) con lo cual $0 < u + v < 2$. Sin embargo las tasas de mutación reales son siempre pequeñas, descartando la posibilidad de que la tasa $1 - \frac{(r+s)(u+v)}{N}$ sea negativa.

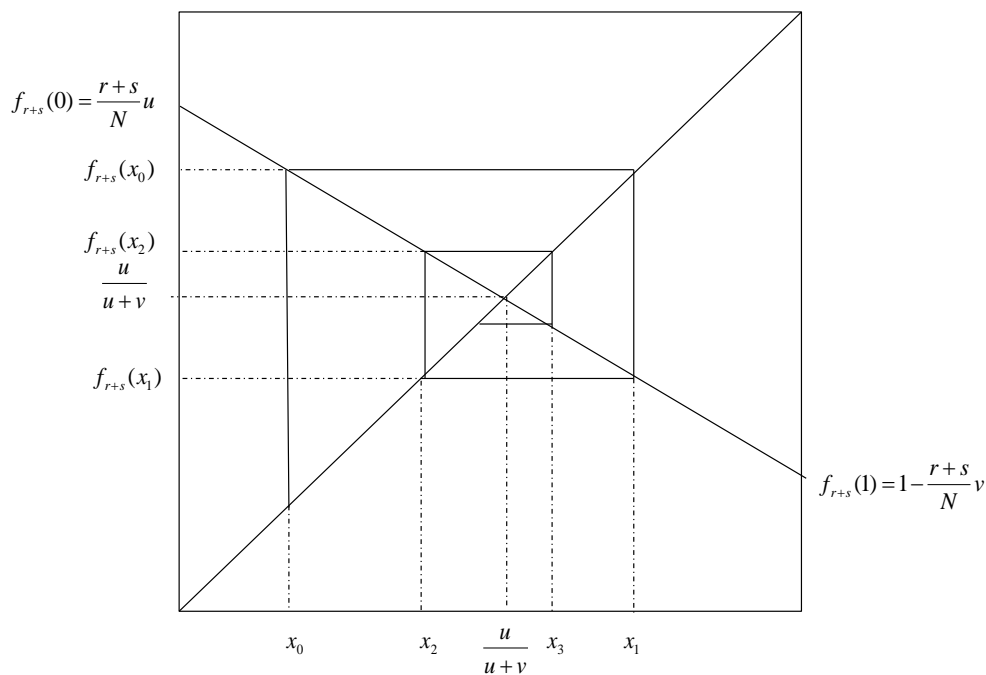


Figura 4.3: Convergencia de los valores de las proporciones esperadas de alelos de tipo A a la proporción esperada de la distribución de equilibrio para $1 - \frac{(r+s)(u+v)}{N} < 0$.

4.2. Inactivación del cromosoma X: Lyonización

El núcleo de una célula humana contiene normalmente 46 cromosomas, 23 correspondientes a la madre y 23 provistos por el padre. Se organizan en pares, siendo 22 pares de autosomas, más un par que determina el sexo. El par sexual consiste en dos cromosomas X en las mujeres (por lo cual su cariotipo, o conjunto de cromosomas, se denota mediante (46, XX)) y un cromosoma X y uno Y en los varones (46, XY). Un hombre hereda un cromosoma X de la madre y el cromosoma Y del padre, y una mujer hereda un cromosoma X de la madre y el cromosoma X del padre.

Si bien una mujer tendrá dos cromosomas X, uno de origen materno (X_m) y otro de origen paterno (X_p), en cada célula sólo uno de estos cromosomas se encontrará activo, permitiendo la expresión de las características de sus genes. Esto se debe a que en una etapa temprana del desarrollo del embrión de las mujeres, uno de los dos cromosomas X de cada célula se inactiva para compensar fisiológicamente la dosis de los genes ligados a dicho cromosoma en ambos sexos. Este fenómeno de compensación de dosis génica ligado al cromosoma X fue descubierto por la citogenetista Mary Lyon en 1961 [17], y se conoce como Lyonización. La inactivación de un cromosoma X materno o paterno se produce en la embriogénesis y ocurre al azar con igual probabilidad e independientemente en cada célula del embrión femenino (de 5-7 días), cuando existen aproximadamente 10 a 30 células precursoras de cada tejido [24]. Por lo tanto, cada individuo resultará un mosaico donde lo más probable es que aproximadamente la mitad de sus células porte inactivo el cromosoma X paterno y la otra mitad tenga inactivo el cromosoma X materno. Una vez ocurrida la inactivación en cada célula somática, dicha inactivación es heredable clonalmente por las células descendientes durante toda la vida de la mujer. Por lo tanto todas las mujeres podrían considerarse hemicigotas en cada célula para uno u otro de los alelos del cromosoma X.

La hipótesis de Lyon se ha demostrado con muchos *loci* ligados al cromosoma X. Un ejemplo es el color del pelaje en algunos mamíferos, como el patrón calicó de los gatos, que se debe a la inactivación del cromosoma X. Los gatos calicó son normalmente hembras heterocigóticas para los alelos naranja y negro de un *locus* ligado al X. Estas hembras muestran áreas de los dos colores, indicando de esta manera que en un cierto estadio del desarrollo uno de los dos cromosomas X se

4.2. INACTIVACIÓN DEL CROMOSOMA X: LYONIZACIÓN

inactivó y que todas las células descendientes de esa línea conservaron inactivo el mismo cromosoma X. Esto da lugar a manchas de color en el pelaje dado que hay regiones de la gata con células que tienen activo el alelo que codifica para el color naranja, y otras células que por el contrario tienen activo el alelo que codifica para el color negro. En lo que respecta al color blanco, dado que este color no está ligado al cromosoma X ni al Y se expresa independientemente de los otros dos colores. Por este motivo una gata puede presentar los tres colores en su pelaje.

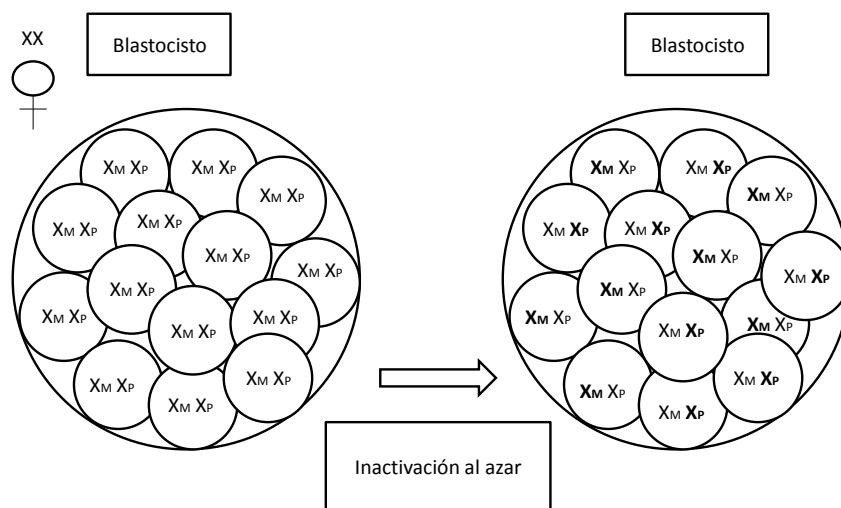


Figura 4.4: Lyonización

4.2.1. Patrón de inactivación del cromosoma X

La inactivación de un cromosoma X materno o paterno en cada una de las células del blastocisto es un proceso estocástico cuyo resultado se puede modelar con una distribución binomial, dado que la probabilidad de inactivar el cromosoma paterno o materno es la misma, igual a $1/2$ (salvo excepciones como mencionaremos más adelante), y ocurre independientemente para cada célula.

Se define como patrón de inactivación del cromosoma X, XIP (X Inactivation Pattern), al porcentaje de células con inactivación del cromosoma X materno (o

paterno). Lo más probable es que aproximadamente la mitad de los clones de una mujer presenten inactivo el cromosoma paterno y la otra mitad presenten inactivo el cromosoma materno, es decir, un XIP del 50%. Sin embargo, no todas las mujeres presentan ese patrón de inactivación. Cuando ocurre un alejamiento significativo del 50%, se dice que hay un sesgo en el patrón de inactivación.

Distribución Binomial del Patrón de Inactivación Inicial

La distribución binomial se puede aplicar para modelar el resultado del proceso de Lyonización. Si se define como Y al número de células que tienen inactivado el cromosoma X materno, resulta que Y es una variable aleatoria con una distribución binomial que tiene como parámetros el número de células precursoras de un tejido, N_0 , el cual se estima entre 10 y 30 según [24], [31] y $p = \frac{1}{2}$, la probabilidad de que en una célula se inactive el cromosoma X materno. Si bien el número más probable de células con inactivación del cromosoma X materno es $N_0/2$ (suponiendo N_0 par), hay una probabilidad positiva de que haya un sesgo en el patrón de inactivación.

Un sesgo en el patrón de inactivación es un apartamiento del 50% por parte del XIP, por ejemplo menor a 35% o mayor a 65%. La probabilidad de un evento tal, es la probabilidad de las correspondientes colas de la distribución de células con el X materno inactivo. En la Figura 4.5 se graficó la función de probabilidad de la variable aleatoria Y tal que $Y \sim Bi(N_0, 1/2)$ para N_0 igual a 10, 15, 20 y 30. Se puede observar que cuanto menor sea el número de células precursoras que darán origen a cada tejido, mayor será la probabilidad de obtener poblaciones celulares con inactivación sesgada. Como ejemplo numérico podemos señalar que para $N_0 = 10$ la probabilidad de tener 80% o más de las células con inactivación materna (8, 9 u 10 de las 10 células) es $7/128 = 0,0546875$, mientras que para $N_0 = 30$ la probabilidad de ese mismo evento (que corresponde a 24, 25, 26, 27, 28, 29 o 30 de las 30 células) es $\approx 0,000715$.

4.2.2. Causas del sesgo en el patrón de inactivación del cromosoma X

Existen mujeres que presentan un sesgo en el patrón de inactivación. Dicho sesgo puede estar asociado a distintas causas [23]:

- (1) Una posible causa es el azar en la inactivación inicial. En este caso la proba-

4.2. INACTIVACIÓN DEL CROMOSOMA X: LYONIZACIÓN

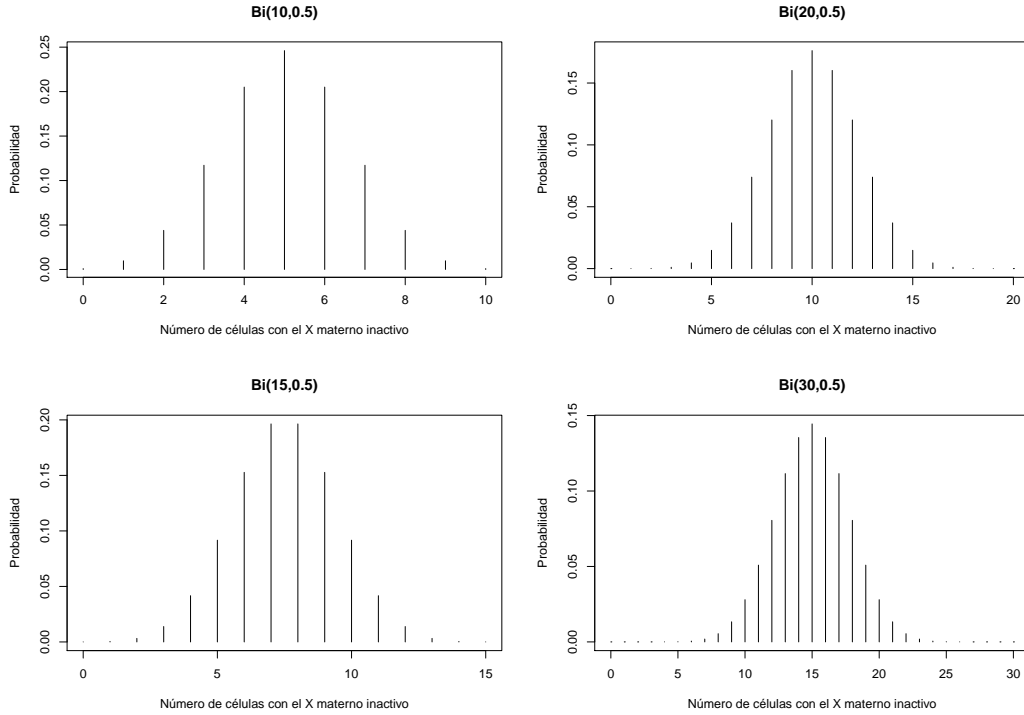


Figura 4.5: Función de probabilidad de la variable Y con $Y \sim Bi(N_0, 1/2)$ variando la cantidad de células precursoras de un tejido, N_0 .

bilidad de que haya un sesgo en el patrón de inactivación es la probabilidad de las colas de la distribución binomial de la variable aleatoria que representa el número de células con inactivación del cromosoma X materno. Como vimos en la subsección anterior la probabilidad de que haya un sesgo en el patrón de inactivación debido a esta causa dependerá del número de células precursoras que darán origen a cada tejido.

- (2) La segunda causa posible es la existencia de una desventaja en la proliferación de las células que presenten activo un cromosoma X portador de una mutación en un gen que es importante para su funcionalidad, maduración, etc., afectando la supervivencia de las células de ese clon de inactivación.
- (3) Por último, por un defecto en el proceso de inactivación en sí mismo. Por ejemplo, mutaciones en el gen XIST (X-inactive specific transcript), que produce un RNA no-codificante de expresión diferencial sobre el X inactivo

y es el responsable primario de la iniciación y desarrollo del proceso de inactivación del cromosoma X.

Las causas (2) y (3) se deben a mecanismos genéticos y en general se asocian a un patrón de inactivación con sesgo extremo, es decir, un XIP superior al 90%. Si alguna de las causas genéticas interviene, es decir, el patrón de inactivación no fuera al azar, se observaría que todos los tejidos presentan sesgo extremo [24]. De lo contrario, es decir, cuando no intervienen las causas (2) y (3), la causa (1) es una importante fuente de sesgo [31].

4.3. Procesos de deriva aplicados al X

Como ya hemos mencionado existen mujeres que presentan un sesgo en el patrón de inactivación. Sin embargo, si las causas (2) y (3) no intervienen lo más probable es que una mujer tenga aproximadamente la mitad de sus células con inactivación del cromosoma X materno y la otra mitad con inactivación del cromosoma X paterno. Esto motivó nuestro estudio de posibles cambios por puro azar en la frecuencia de células con inactivación del cromosoma X materno durante la expansión clonal de las células precursoras de un tejido que ocurre con posterioridad a la inactivación inicial, suponiendo que dicha expansión es no sincrónica (proceso que será modelado en la Subsección 4.3.2), y durante los episodios de renovación celular (proceso que modelaremos en la Subsección 4.3.1). Veremos en primer lugar el caso de renovación celular, sobre el cual existe modelado estocástico en la bibliografía ([36] y [37]) pero nosotros abordamos introduciendo la renovación parcial.

4.3.1. Deriva cromosómica en la renovación celular

Las células de un organismo no viven indefinidamente y su vida media depende del tipo de célula que se considere. Hay células cuyo período de vida es largo, como las musculares o las neuronas, en comparación con la vida de por ejemplo, las células sanguíneas y epiteliales, que tienen una mayor tasa de renovación a partir de sus células progenitoras. El número de células que componen un tejido en un organismo adulto permanece aproximadamente constante; las células que mueren se sustituyen por otras, proceso que permite el mantenimiento de un balance adecuado entre la pérdida y la renovación celular. El recambio celular en los tejidos de un

4.3. PROCESOS DE DERIVA APLICADOS AL X

organismo se fundamenta en el mantenimiento de un equilibrio (homeostasis) entre proliferación y muerte celular a fin de garantizar la población adecuada en cada momento [16]. Entonces cada tejido posee sus propias características fisiológicas de recambio.

Con el fin de estudiar las posibilidades de un cambio en el patrón de inactivación a lo largo de la vida de una mujer por puro azar, desarrollamos un modelo probabilístico de deriva. Nos basamos en el proceso descrito en la Sección 2.1, con el que representamos los eventos de renovación celular en los cuales se producen nacimientos y muertes de determinadas células. Para estudiar la evolución de las frecuencias de células con inactivación de uno u otro cromosoma X en un tejido (ya no frecuencias alélicas en este caso), se describe el recambio celular como un proceso estocástico de muestreo y expansión clonal con reemplazo parcial de la población en cada generación. Esta manera de describir el recambio celular se basa en lo que sucede en la mitosis, proceso por el cual se obtienen células hijas con información genética idéntica a las de las células progenitoras, que tiene lugar en las células somáticas y permite la renovación de tejidos.

Por falta de datos suficientes, este modelo no puede ajustarse a lo observado *in vivo* hasta el presente, pero permite estudiar mediante acotaciones el rol que juega el azar en el cambio de las frecuencias de células con inactivación de uno u otro cromosoma X en los episodios de renovación celular.

Se considera una población con un número constante, N , de individuos, donde cada individuo representa una célula con el cromosoma X materno o paterno inactivo. Por lo tanto, N representará el número total de células de un tejido. Las sucesivas generaciones representan los episodios de renovación celular. En estos episodios hay células que se reproducen, células que mueren y células que simplemente permanecen. Es posible interpretar a la generación $n + 1$ como el resultado de cuatro etapas. En las primeras dos etapas, del total de N células de la generación n se eligen al azar r células que son las que mueren y de las $N - r$ que restan, se eligen s ($s < N - r$) células que son aquellas que se reproducen clonalmente hasta que sus descendientes alcanzan la cantidad r , reponiendo la cantidad que había originalmente. Dicha expansión clonal, que tiene lugar en la tercera etapa, mantiene la proporción de células con inactivación del cromosoma X materno presente en la muestra aleatoria de tamaño s . Finalmente la generación $n + 1$ se forma con las $N - r$ células que permanecen y las nuevas r células producto de la tercera etapa.

CAPÍTULO 4. APLICACIONES

Si definimos X_n como el número de células con inactivación del cromosoma X materno luego del n -ésimo episodio de recambio, el rango de X_n es $\{0, 1, \dots, N\}$ para cada $n \geq 0$. Vimos en la Sección 2.1, Capítulo 2 que

$$E(X_{n+1} | X_0 = i) = i$$

para cada $n \in \mathbb{N}$ lo cual se deducía de (2.1.17). Este resultado indica que el número esperado de células con inactivación del cromosoma X materno se mantiene igual al inicial, en términos probabilísticos. Sin embargo, podría ocurrir que el número de células que tienen inactivado el cromosoma materno, X_n , pasa en algún momento a ser N o 0 indicando la pérdida de uno de los tipos de células si se dispusiera del tiempo suficiente. Cuando uno de los dos tipos de células se pierde de la población, nunca retorna, es por eso que 0 y N son estados absorbentes. Como $\{X_n\}_{n \geq 0}$ es una cadena de Markov absorbente, la probabilidad de que algún tipo de célula se fije en la población es 1. Eso significa que, independientemente del estado inicial de la población y pese a no haber factores selectivos que favorezcan uno u otro tipo, con el transcurrir del tiempo la población termina siendo de un tipo homogéneo. Cabe destacar que esto no implica que una mujer que cuando nace tiene un XIP del 50% llegue a tener un XIP del 100% conforme avance su edad.

El tiempo de fijación es:

$$\tau = \min \{n : X_n = 0 \text{ o } X_n = N\}.$$

Es decir, corresponde al primer instante en que la población consiste en todas células que tienen inactivado el cromosoma materno o todas que tienen inactivado el cromosoma paterno. Como se vió en (1.1.17) se tiene que

$$P(X_\tau = N | X_0 = i) = i/N.$$

Esto es, la probabilidad de fijación para células con el X materno inactivo es igual a la proporción de células con el X materno inactivo en la población inicial (y análogamente para células con el X paterno inactivo).

Para estudiar los efectos de la deriva que tienen lugar en la etapa de renovación de la población, es relevante la aleatoriedad en la elección de las s células que se reproducirán. Luego de este muestreo aleatorio la expansión es determinística, es decir, las r células hijas tendrán el mismo patrón de inactivación que las s células

4.3. PROCESOS DE DERIVA APLICADOS AL X

progenitoras. Consideramos entonces los cálculos de la varianza y heterocigosidad para el modelo de deriva con expansión determinística de la población realizados en la Sección 2.1. En este caso el término heterocigosidad alude a la condición donde coexisten ambos clones de inactivación del cromosoma X y no a una coexistencia de tipos de alelos.

Para estudiar cómo cambia el sesgo hacia uno de los dos tipos de células luego de n recambios celulares por deriva, estudiemos la fórmula de la heterocigosidad. La heterocigosidad se define como la probabilidad de tomar, sin reemplazo, dos células de distinto tipo en la generación n , es decir, una de ellas tiene inactivado el cromosoma materno, y la otra, el paterno. Una baja probabilidad de este evento indica un alto sesgo en el patrón de inactivación y viene dada por:

$$h(n) = h(0) \left(1 - \frac{r(r+s)}{N(N-1)s} \right)^n,$$

según (2.1.25) (ver Capítulo 2, Subsección 2.1.3).

Con el objetivo de estudiar para qué valores de los parámetros r y s se puede producir un cambio en el patrón de inactivación inicial luego de n generaciones, hallamos una función del número de células que mueren, r , en función del número de células que se reproducen, s , en cada episodio de renovación celular. Para ello despejamos r en función de s de la expresión hallada para la heterocigosidad y obtenemos la siguiente función:

$$r : [0, N] \rightarrow [0, N] \text{ tal que } r(s) = \frac{\sqrt{s^2 + 4 \left(1 - \sqrt[n]{\frac{h(n)}{h(0)}} \right) N(N-1)s} - s}{2}.$$

En efecto, si partimos de la expresión que hallamos para la heterocigosidad

$$h(n) = h(0) \left(1 - \frac{r(r+s)}{N(N-1)s} \right)^n,$$

luego de algunas operaciones algebraicas obtenemos que esta última expresión se puede reescribir de la siguiente manera

$$0 = r^2 + rs - \left(1 - \sqrt[n]{\frac{h(n)}{h(0)}} \right) N(N-1)s. \quad (4.3.1)$$

Para cada posible valor de s (4.3.1) es una ecuación de segundo grado cuya solución

CAPÍTULO 4. APLICACIONES

es

$$r(s) = \frac{\sqrt{s^2 + 4 \left(1 - \sqrt[n]{\frac{h(n)}{h(0)}}\right) N (N-1) s - s}}{2}. \quad (4.3.2)$$

Por otro lado, tenemos que la varianza de la variable aleatoria X_n , según (2.1.23), es

$$\text{Var}(X_n) = X_0 (N - X_0) \left(1 - \left(1 - \frac{r(r+s)}{s(N-1)N}\right)^n\right).$$

La varianza se puede expresar en función de la heterocigosidad de la siguiente manera

$$\text{Var}(X_n) = \frac{N^2}{2} (h(0) - h(n)). \quad (4.3.3)$$

En efecto, si reemplazamos $\left(1 - \frac{r(r+s)}{s(N-1)N}\right)^n$ por $\frac{h(n)}{h(0)}$ en (2.1.23) obtenemos que

$$\begin{aligned} \text{Var}(X_n) &= X_0 (N - X_0) \left(1 - \frac{h(n)}{h(0)}\right) \\ &= \frac{N^2}{2} (h(0) - h(n)). \end{aligned}$$

Observemos que cuanto menor es la heterocigosidad en la generación n , mayor es la dispersión de la distribución de la variable aleatoria X_n .

Resultados

Consideremos la siguiente función

$$r' : [0, 1] \rightarrow [0, 1] \text{ tal que } r'(s') = \frac{\sqrt{s'^2 + 4 \left(1 - \sqrt[n]{\frac{h(n)}{h(0)}}\right) (N-1) s' - s'}}{2}, \quad (4.3.4)$$

que se obtiene estandarizando el número de células que mueren y el número de células elegidas para reproducirse mediante $r' = r/N$ y $s' = s/N$ respectivamente, en (4.3.1).

A partir de (4.3.4), el número de células en un tejido, N , y el número promedio de episodios de renovación celular a lo largo de la vida de una mujer, n , queremos estudiar para qué valores de los parámetros r' y s' , según el modelo propuesto, una mujer que por ejemplo tiene un patrón de inactivación inicial (XIP_0) 50 : 50, podría pasar a tener un sesgo (XIP_{final}) 40 : 60 o 20 : 80 a lo largo de la vida.

Esto a su vez nos permitiría establecer si la deriva que se produce por la renovación de tejidos que ya alcanzaron el número de células presente en el nacimiento y la vida adulta, influye o no en la evolución del sesgo en el patrón de inactivación en una mujer.

Sesgo en el patrón de inactivación del cromosoma X

Vickers y colaboradores [37] estimaron la deriva en el sesgo del patrón de inactivación del cromosoma X en sangre periférica. Una de las hipótesis que consideran es que el sesgo que se observa en mujeres adultas se puede deber a la aleatoriedad en la división de células madre hematopoyéticas. A partir de los resultados que obtienen concluyen que un modelo estocástico de la división de células madre no es suficiente para explicar el sesgo en el patrón de inactivación del cromosoma X. En este trabajo consideran que el número de células madre es constante y es del orden de 10^6 . También estiman el número medio de divisiones por célula madre a lo largo de la vida. Este número fue estimado en 70 por Vickers *et al.* [36] en otro trabajo de investigación en el cual modelan la división de células madre hematopoyéticas.

Haciendo uso de las estimaciones de Vickers *et al.* estudiaremos para qué valores de los parámetros r' y s' podría producirse un sesgo en el patrón de inactivación por deriva a lo largo de la vida de una mujer. En particular, consideramos la estimación por parte de Vickers *et al.* [36] del número medio de divisiones por célula madre a lo largo de la vida para estimar el número de episodios de renovación celular. En nuestro modelo de renovación parcial sólo s células se renuevan en cada generación, entonces para que el promedio de divisiones por célula sea de 70, necesitaríamos aproximadamente $70 \cdot N/s = 70/s'$ generaciones de nuestro proceso. Luego si reemplazamos a N por 10^6 y n por $70/s'$ en (4.3.4) obtenemos

$$r' : [0, 1] \rightarrow [0, 1] \text{ tal que } r'(s') = \frac{\sqrt{s'^2 + 4 \left(1 - \frac{70}{s'} \sqrt{\frac{h_f}{h(0)}}\right) (10^6 - 1) s' - s'}}{2}, \quad (4.3.5)$$

donde h_f corresponde a la heterocigosidad final.

Correspondencia entre un valor de XIP y una heterocigosidad

La investigación médica utiliza el XIP para cuantificar el sesgo porque es un valor observable en una mujer. Las técnicas convencionales para medición del XIP

que usan la correlación X inactivo/marcas epigenéticas sólo indican el sentido del sesgo en un *loci* puntual del cromosoma. En la práctica ésto equivale a conocer la magnitud pero no el sentido del sesgo. Así, un XIP de 40 : 60 indica o bien 40 % de células con el X materno inactivo o bien 40 % de células con el X paterno inactivo. De manera que un valor de XIP corresponde a dos valores equidistantes de la media $N/2$ (suponiendo un XIP inicial no aleatorio de 50 : 50). Éstos son sólo dos valores puntuales que puede tomar la variable aleatoria X_{final} cuya distribución tiene una heterocigosidad h_f y una varianza $V(X_{final})$, parámetros que pueden calcularse conociendo s' , r' y el número de episodios de renovación.

Para tener una idea del XIP al que puede llegar una mujer en sangre periférica dados s' , r' y $70N/s$, hallamos el XIP que en algún sentido es el más representativo de la distribución de $X_{final} = X_{70N/s}$. Éste es el XIP tal que si la distribución estuviera concentrada simétricamente en los dos puntos que representa, tendría la misma varianza que X_{final} . De esta manera se verifica que la distancia esperada a la media para la distribución que asignamos al XIP es la misma que la de la distribución de X_{final} . Como la varianza determina la heterocigosidad y viceversa, establecemos así una correspondencia entre una heterocigosidad, que podemos calcular, y un XIP asociado a la misma.

Primero planteamos entonces para cada XIP_{final} la varianza de la distribución que está concentrada en los dos posibles valores que puede tomar el número de células con inactivación del cromosoma X materno según el XIP_{final} que se considere, llamémoslos x_{final_1} y x_{final_2} , con media $\frac{10^6}{2}$:

$$Var_{XIP_{final}} = \left(x_{final_1} - \frac{10^6}{2}\right)^2 \frac{1}{2} + \left(x_{final_2} - \frac{10^6}{2}\right)^2 \frac{1}{2}. \quad (4.3.6)$$

El objetivo es, para un XIP_{final} dado, hallar los valores de s' y r' que determinan una distribución final con una varianza igual a la varianza $Var_{XIP_{final}}$ definida en (4.3.6). Como en particular nos interesa la correspondencia con la heterocigosidad final, consideramos (4.3.3) y obtenemos

$$\frac{(10^6)^2}{2} (h(0) - h_f) = \left(x_{final_1} - \frac{10^6}{2}\right)^2 \frac{1}{2} + \left(x_{final_2} - \frac{10^6}{2}\right)^2 \frac{1}{2}. \quad (4.3.7)$$

4.3. PROCESOS DE DERIVA APLICADOS AL X

Despejando h_f de esta última expresión tenemos

$$h_f = h(0) - \left(\left(x_{final_1} - \frac{10^6}{2} \right)^2 \frac{1}{(10^6)^2} + \left(x_{final_2} - \frac{10^6}{2} \right)^2 \frac{1}{(10^6)^2} \right). \quad (4.3.8)$$

Finalmente a partir de esta expresión calcularemos la heterocigosidad final, h_f , de las distribuciones más representativas para cada posible XIP_{final} (XIP alcanzado a lo largo de la vida). Más aún, mediante (4.3.5) calcularemos los s' y r' que determinan dichas distribuciones.

Ahora estamos en condiciones de graficar a partir de (4.3.5) el conjunto de puntos (s', r') para los cuales, según el modelo propuesto, una mujer que tiene un patrón de inactivación inicial (XIP_0) 50 : 50 pasa a tener (en el sentido de la correspondencia que establecimos con la heterocigosidad) un XIP de 49 : 51, 48 : 52, 47 : 53, 46 : 54, 45 : 55, 40 : 60, 30 : 70 y 10 : 90. También incluimos el caso en el que el patrón inicial se mantiene en 50 : 50. Entonces para calcular la heterocigosidad final en cada caso consideramos (4.3.8). Estos valores de parámetros están resumidos en el Cuadro 4.2.

$h(0)$	h_f
0.5 (XIP_0 50 : 50)	0.5 (XIP_{final} 50 : 50)
0.5 (XIP_0 50 : 50)	0.4998 (XIP_{final} 49 : 51)
0.5 (XIP_0 50 : 50)	0.4992 (XIP_{final} 48 : 52)
0.5 (XIP_0 50 : 50)	0.4982 (XIP_{final} 47 : 53)
0.5 (XIP_0 50 : 50)	0.4968 (XIP_{final} 46 : 54)
0.5 (XIP_0 50 : 50)	0.495 (XIP_{final} 45 : 55)
0.5 (XIP_0 50 : 50)	0.48 (XIP_{final} 40 : 60)
0.5 (XIP_0 50 : 50)	0.42 (XIP_{final} 30 : 70)
0.5 (XIP_0 50 : 50)	0.18 (XIP_{final} 10 : 90)

Cuadro 4.2: En este cuadro se exponen los parámetros que se introducen en la fórmula (4.3.5) para hacer los gráficos de la Figura 4.6.

En la Figura 4.6 se muestran los gráficos realizados a partir de las situaciones iniciales y finales del Cuadro 4.2.

CAPÍTULO 4. APLICACIONES

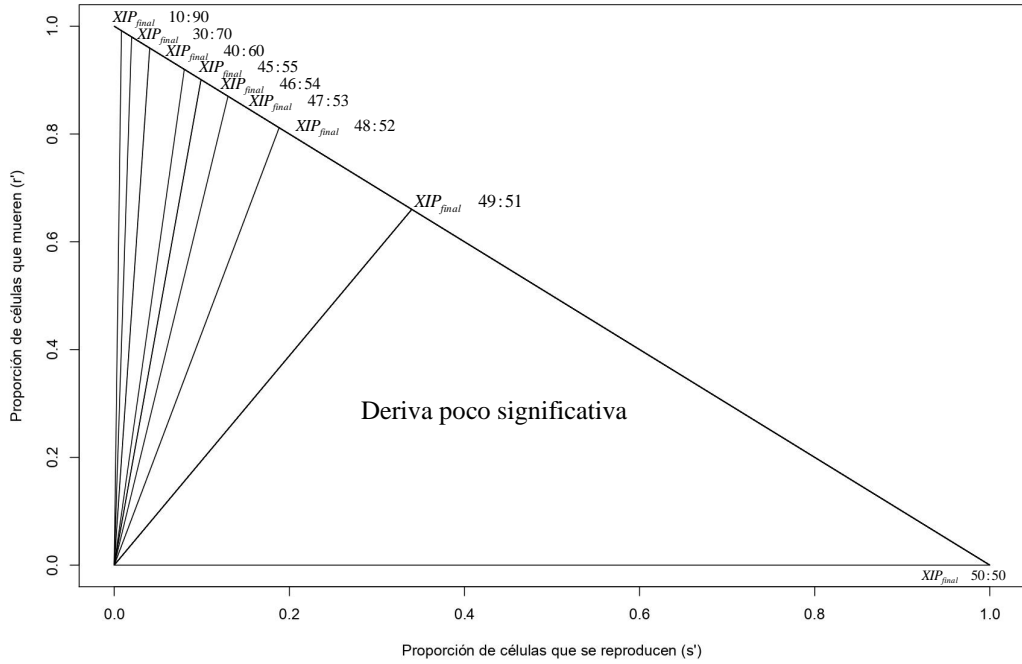


Figura 4.6: Proporción de células que mueren en función de la proporción de células que se reproducen para $h(0) = 0,5$ y los valores de h_f establecidos en el Cuadro 4.2. Además graficamos la recta $r' = 1 - s'$.

Como sugieren los gráficos de la Figura 4.6 la función (4.3.5) se puede aproximar por una función lineal. Ésta función es

$$r'(s') = s' \frac{\left(\sqrt{\left(1 + \frac{4(10^6-1)}{70} \ln\left(\frac{h(0)}{h_f}\right)\right)} - 1 \right)}{2}, \quad (4.3.9)$$

siempre que $\frac{s'}{70} \ln\left(\frac{h(0)}{h_f}\right)$ sea cercano a cero. Esto último se obtiene aproximando $\sqrt{\frac{h_f}{h(0)}} = e^{-\frac{s'}{70} \ln\left(\frac{h(0)}{h_f}\right)}$ por $1 - \frac{s'}{70} \ln\left(\frac{h(0)}{h_f}\right)$ para $\frac{s'}{70} \ln\left(\frac{h(0)}{h_f}\right)$ pequeño, usando que si x es pequeño $e^{-x} \approx 1 - x$.

Para cada uno de los valores de $h(0)$ y h_f considerados en el Cuadro 4.2 se tiene que $0 \leq \ln\left(\frac{h(0)}{h_f}\right)/70 < 0,015$, lo cual implica que $s' \ln\left(\frac{h(0)}{h_f}\right)/70$ es próximo a cero para todo s' ($0 \leq s' \leq 1$). Luego para los valores de parámetros considerados

4.3. PROCESOS DE DERIVA APLICADOS AL X

en el Cuadro 4.2, la función (4.3.5) se puede aproximar por (4.3.9) (ver el Cuadro 4.3).

$h(0)$	h_f	$r'(s') \approx$
0.5 (XIP_0 50 : 50)	0.4998 (XIP_{final} 49 : 51)	1,94s'
0.5 (XIP_0 50 : 50)	0.4992 (XIP_{final} 48 : 52)	4,31s'
0.5 (XIP_0 50 : 50)	0.4982 (XIP_{final} 47 : 53)	6,69s'
0.5 (XIP_0 50 : 50)	0.4968 (XIP_{final} 46 : 54)	9,09s'
0.5 (XIP_0 50 : 50)	0.495 (XIP_{final} 45 : 55)	11,49s'
0.5 (XIP_0 50 : 50)	0.48 (XIP_{final} 40 : 60)	23,65s'
0.5 (XIP_0 50 : 50)	0.42 (XIP_{final} 30 : 70)	49,41s'
0.5 (XIP_0 50 : 50)	0.18 (XIP_{final} 10 : 90)	120,31s'

Cuadro 4.3: En este cuadro se exponen las fórmulas de las rectas que aproximan a (4.3.5) para $h(0)$ y h_f considerados en el Cuadro 4.2

Deriva poco significativa

Primero consideremos aquellas combinaciones de los parámetros s y r que dan lugar a eventos de renovación celular que producen cambios poco significativos en el patrón de inactivación inicial a lo largo de la vida, según el modelo propuesto. En estos casos el valor establecido por el trabajo de Vickers para el número de células madre en el tejido sanguíneo lo podemos considerar como una cota inferior para nuestro parámetro N (el número de células en un tejido) porque la cantidad de células madre es menor que la de células somáticas ($N \geq 10^6$).

Veamos que con esta cota tenemos el mejor caso para la deriva. Observemos que si $N \geq 10^6$ resulta que

$$h(n) = h(0) \left(1 - \frac{r(r+s)}{N(N-1)s} \right)^n \geq h(0) \left(1 - \frac{r(r+s)}{10^6(10^6-1)s} \right)^n.$$

De manera que cuanto más pequeño sea el tamaño de la población más significativa será la deriva, porque se reduce la probabilidad de elegir dos células de distinto tipo.

En la Figura 4.6 representamos el conjunto de puntos s' y r' tales que el sesgo a lo largo de la vida partiendo de un patrón de inactivación inicial 50 : 50 pasa a ser 49 : 51. Tomando como referencia esta curva obtenemos el conjunto de puntos

para los cuales la deriva es poco significativa, que se corresponde con el área del gráfico que se encuentra por debajo de la misma.

Si el número de células que mueren y se reproducen en cada episodio de renovación celular pertenecen a dicha región del gráfico, entonces nuestro modelo predice que la heterocigosidad disminuye poco, aún en tejidos de alto recambio celular, como la sangre ([5], [6] y [7]). Luego, menores aún serán dichos efectos si $N \geq 10^6$. Vickers *et al.* [37] también concluyen que la probabilidad de sesgo por deriva es demasiado pequeña, pero lo hacen aplicando un modelo estocástico (basado en el modelo de Wright-Fisher) en el que la renovación a cada paso es total sin tener en cuenta la posibilidad de que sólo una parte de la población de células se expanda como lo hacemos en nuestro análisis.

Deriva significativa

A partir de la Figura 4.6 observamos ahora los puntos del gráfico por encima de la curva que representa un XIP_{final} de 49 : 51.

Considerando la aproximación de (4.3.5) por la recta (4.3.9) (ver Cuadro 4.3) se tiene que:

- Para cada evolución desde $h(0) = 0,5$ hasta un h_f final dado s' y r' resultan proporcionales.
- Las pendientes de las rectas toman valores cada vez más grandes conforme aumenta el sesgo final, XIP_{final} . En términos del modelo, ocurre que cuanto mayor es el sesgo final, mayor es el número de células que mueren por célula que se reproduce.
- Para cada XIP_{final} tenemos la acotación $s' \leq c_{h_f}$ donde

$$c_{h_f} = \frac{2}{\left(\sqrt{\left(1 + \frac{4(10^6-1)}{70} \ln\left(\frac{h(0)}{h_f}\right)\right)} + 1\right)}. \quad (4.3.10)$$

En efecto, si consideremos que $r' \leq 1 - s'$ resulta

$$s' \frac{\left(\sqrt{\left(1 + \frac{4(10^6-1)}{70} \ln\left(\frac{h(0)}{h_f}\right)\right)} - 1\right)}{2} \leq 1 - s',$$

de donde se obtiene el valor de la cota c_{h_f} , que en la Figura 4.6 es la abscisa de las intersecciones de las gráficas de $r'(s')$ y la recta $r' = 1 - s'$. Esta cota representa la máxima proporción de células que se pueden reproducir en cada evento de renovación celular, si inicialmente $h(0) = 0,5$, para cada h_f . Cuanto mayor es el sesgo del XIP_{final} (más deriva), menor es el valor de la cota.

Si los valores que toman r y s en nuestro escenario biológico particular estuviesen incluidos en la región del gráfico donde la deriva es significativa, sería de esperar que una mujer adulta presente sesgo en el patrón de inactivación como consecuencia de la deriva en los episodios de renovación celular.

4.3.2. Deriva en el crecimiento por expansión clonal

En esta subsección analizaremos si el patrón inicial de inactivación del cromosoma X podría cambiar durante las primeras divisiones celulares de la expansión clonal de las células precursoras de cada tejido, suponiendo que las causas genéticas (2) y (3) expuestas en la Subsección 4.2.2 no intervienen. Se trata de un análisis que no encontramos en la bibliografía y que llevamos a cabo para comparar nuestros resultados con los existentes.

Nos proponemos estudiar si a partir de la inactivación, en el proceso de crecimiento del tejido, es significativa la deriva. En particular, si puede llevar a un porcentaje de 90 % o más con una probabilidad mayor que la que da la inactivación inicial. Por ejemplo, podría ocurrir que una mujer con 80 % de inactivación inicial tenga una alta probabilidad de pasar a 90 % o más por deriva durante la expansión clonal que ocurre inmediatamente después de la inactivación.

4.3.3. Modelo de deriva de Polya-Eggenberger aplicado a la expansión clonal

Con el fin de estudiar la posibilidad de deriva en el patrón de inactivación, no en los episodios de renovación celular sino durante el desarrollo embrionario, analizamos un modelo apropiado para una población pequeña en crecimiento que consiste en una variante del modelo de urna de Polya-Eggenberger, que desarrollamos en el Capítulo 3, Sección 3.2.

CAPÍTULO 4. APLICACIONES

En nuestro caso no partimos de una urna con bolillas de dos colores diferentes sino de una situación análoga que consiste en una población inicial de tamaño N_0 , que representa el número de células de un tejido al momento de la inactivación. La población se compone de dos tipos de células, las que tienen inactivación del cromosoma X materno o paterno, en la que a cada paso, una célula al azar se reproduce por clonación. La proporción inicial de células con el X materno inactivo es aleatoria. Esta aleatoriedad se debe a que el número de células con el cromosoma X materno inactivo luego del proceso de inactivación viene dado por una variable aleatoria, X_0 , que como vimos en la Subsección 4.2.1 tiene distribución binomial con parámetros N_0 y $1/2$, lo cual corresponde a la situación biológica que está en estudio.

El modelo de deriva aplicado a la expansión clonal que se propone, consiste en elegir aleatoriamente una célula e ir agregando exactamente una célula con inactivación en el mismo cromosoma que la que fue elegida a cada paso. Este procedimiento se repite hasta que se hayan agregado tantas células como las que conforman un tejido adulto.

Entonces, la generación n tendrá $N_0 + n$ células, donde n representa el número de células que se agregaron al tejido hasta el paso n , luego de la inactivación. La variable aleatoria X_n representa el número de células con el X materno inactivo en la generación n , es decir, luego de la incorporación de n nuevas células a las N_0 iniciales.

El objetivo es calcular para cada $n \in \mathbb{N}$, la probabilidad $P(X_n \geq 0,9 \cdot (N_0 + n))$ y de esta manera estudiar si durante la expansión clonal se puede producir un sesgo extremo en uno de los dos sentidos (90% o más de las células con la misma inactivación) por deriva con una probabilidad mayor que $P(X_0 \geq 0,9 \cdot N_0)$. Ésta última es la probabilidad que viene dada por la distribución binomial y representa la probabilidad de sesgo extremo al momento de la inactivación, la cual se mantendría si no hubiera deriva. Éste sería el caso si en cada generación todas las células experimentaran la clonación al mismo tiempo, mientras que nuestro modelo explora la posibilidad de que se verifique una cierta asincronicidad y la deriva que esto podría producir.

Cálculo de la probabilidad de que el 90 % o más de las células tengan el X materno inactivo

Para poder estudiar la evolución del sesgo a partir de la inactivación en el blastocisto según nuestro modelo, utilizaremos la fórmula exacta (3.2.11) (no asintótica) que obtuvimos en la Proposición 8 del Capítulo 3, Sección 3.2:

$$P(X_n \geq 0,9 \cdot (N_0 + n)) = \begin{cases} \left(\sum_{i=0,9 \cdot (N_0+n)}^{N_0+n-1} \left(\frac{1}{2}\right)^{N_0} \sum_{j=\max\{1, i-n\}}^{\min\{i, N_0-1\}} \binom{n}{i-j} \frac{\beta(i, N_0+n-i)}{\beta(j, N_0-j)} \binom{N_0}{j} \right) + \left(\frac{1}{2}\right)^{N_0} & \text{si } 0,9 \cdot (N_0 + n) \in \mathbb{N} \\ \left(\sum_{i=[0,9 \cdot (N_0+n)]+1}^{N_0+n-1} \left(\frac{1}{2}\right)^{N_0} \sum_{j=\max\{1, i-n\}}^{\min\{i, N_0-1\}} \binom{n}{i-j} \frac{\beta(i, N_0+n-i)}{\beta(j, N_0-j)} \binom{N_0}{j} \right) + \left(\frac{1}{2}\right)^{N_0} & \text{si } 0,9 \cdot (N_0 + n) \notin \mathbb{N} \end{cases}$$

donde $[0,9 \cdot (N_0 + n)]$ representa la parte entera de $0,9 \cdot (N_0 + n)$.

A partir del cálculo de la $P(X_n \geq 0,9 \cdot (N_0 + n))$ para cada $n \in \mathbb{N}$ podremos estudiar si dicha probabilidad de sesgo extremo aumenta, disminuye o se mantiene a medida que se incrementa el número de células, n , como veremos en la Subsección 4.3.4.

Fórmula asintótica de la probabilidad de que el 90 % o más de las células tengan el X materno inactivo

Además consideraremos la fórmula asintótica de la probabilidad de que el 90 % o más de las células tengan el X materno inactivo, que según (3.2.18) es:

$$\lim_{n \rightarrow \infty} P(X_n \geq 0,9 \cdot (n + N_0)) = \sum_{j=1}^{N_0-1} P(\beta_{j, N_0-j} \geq 0,9 | X_0 = j) \binom{N_0}{j} \left(\frac{1}{2}\right)^{N_0} + \left(\frac{1}{2}\right)^{N_0}$$

con $\beta_{j, N_0-j} \sim \beta(j, N_0 - j)$ (Capítulo 3, Sección 3.2). A partir de este resultado vamos a poder estimar la probabilidad de que el 90 % o más de las células tengan el X materno inactivo cuando el número de células que se agregaron es grande.

4.3.4. Evolución del sesgo a partir de la inactivación en el blastocisto

Para poder estudiar la evolución del sesgo a partir de la inactivación en el blastocisto mediante nuestro modelo, es necesario establecer el número de células de un tejido en el momento de la inactivación, el cual notamos por N_0 . La bibliografía consultada acerca del fenómeno de inactivación sugiere que hay entre 10 y 30 células precursoras de cada tejido cuando la inactivación se produce ([24] y [31]). Ilustramos nuestros resultados considerando que el proceso de expansión clonal se inicia con 12 células. Este número de células nos permitirá luego la comparación con trabajos previos, más precisamente con Shvestova *et al.* [31] como luego veremos.

En nuestro análisis vamos a suponer que no intervienen otros factores genéticos, entonces como ya se mencionó en la Subsección 4.2.2, la probabilidad de sesgo en el patrón de inactivación inicial viene dada por las colas de una distribución binomial con parámetros $N_0 = 12$ y $p = 1/2$.

De acuerdo a la notación que se introdujo anteriormente la probabilidad de sesgo extremo inicial es

$$P(X_0 \geq 0,9 \cdot 12) = P(X_0 \geq 10,8) \approx 0,002792,$$

que resulta de aproximar con la distribución normal. Recurrimos a esta aproximación porque X_0 , como ya hemos mencionado, tiene una distribución binomial, y por lo tanto para tener un valor más representativo de la probabilidad inicial de sesgo extremo es deseable reemplazar a $P(X_0 \geq 10,8)$ por un valor interpolado entre

$$P(X_0 \geq 10) = \sum_{k=10}^{12} \binom{12}{k} \left(\frac{1}{2}\right)^{12} = 79/4096 \approx 0,019287$$

y

$$P(X_0 \geq 11) = \sum_{k=11}^{12} \binom{12}{k} \left(\frac{1}{2}\right)^{12} = 13/4096 \approx 0,003174.$$

Por otro lado calculamos mediante la fórmula (3.2.11), que obtuvimos en la Subsección 4.3.3, la probabilidad de que el porcentaje de células con el X materno

inactivo sea mayor o igual al 90 % del total de células en cada generación. Esto último se realizó con el objetivo de estudiar si durante la expansión clonal, aplicando nuestro modelo no sincrónico, se puede producir un sesgo extremo por deriva con una probabilidad mayor que la probabilidad de sesgo extremo inicial.

Expansión clonal sincrónica vs asincrónica

Shvetsova y colaboradores publicaron que el sesgo en el patrón de inactivación de una mujer está determinado por el sesgo en el patrón de inactivación inicial, y se debe a la aleatoriedad del proceso de inactivación [31], que como vimos es la causa (1) que se mencionó en la Subsección 4.2.2.

En ese trabajo se calcula la distribución teórica que se ajusta a la distribución observada del sesgo en el patrón de inactivación en la población de mujeres en estudio. Para dicho cálculo los autores realizaron simulaciones considerando diferentes números de células, 8, 16 y 32, basadas en el hecho de que la inactivación es un fenómeno estocástico que se puede modelar a partir de una distribución binomial con parámetros N_0 y $\frac{1}{2}$ (la probabilidad de que se inactive el cromosoma X materno o paterno). El número de células, N_0 , representa el número de células madre en el momento de la inactivación. Finalmente concluyen que el sesgo observado empíricamente es explicado por un modelo según el cual la inactivación se produce en el estado embrionario cuando hay 8 células de un tejido, que se expanden clonalmente hasta conformar el tejido adulto.

Nosotros en cambio proponemos dar lugar a la aleatoriedad a través de la asincronicidad utilizada para modelar la expansión clonal, en la cual a cada paso se va agregando una célula elegida al azar entre todas las presentes. El sesgo empírico medido por Shvetsova *et al.* se puede justificar a partir de nuestro modelo no sincrónico considerando un número inicial de células (N_0) igual a 8, 9, 10, 11, 12 y 13 (ver Cuadro 4.4). Establecemos que inicialmente hay 12 células porque es el mayor número que da en el límite una probabilidad de sesgo extremo que supera al valor observado empíricamente por Shvetsova *et al.* En la Figura 4.7 representamos la probabilidad de que el 90 % o más de las células sean células con el X materno inactivo para cada una de las primeras 2000 generaciones. Cabe aclarar que en este gráfico de la evolución del sesgo desde el blastocisto comenzando con 12 células se puede observar un comportamiento zigzagante que se debe a la naturaleza discreta de cada variable aleatoria X_n .

CAPÍTULO 4. APLICACIONES

N_0	$P(X_0 \geq 0,9 \cdot N_0)$	Exp. Clonal Asíncronica: $\lim_{n \rightarrow \infty} P(X_n \geq 0,9 \cdot (N_0 + n))$	Exp. Clonal Sincrónica: $\lim_{n \rightarrow \infty} P(X_n \geq 0,9 \cdot (N_0 + n))$
8	0,011825	0,042987	0,011825
9	0,008197	0,032701	0,008197
10	0,005706	0,025002	0,005706
11	0,003985	0,019194	0,003985
12	0,002792	0,014785	0,002792
13	0,001960	0,011420	0,001960

Cuadro 4.4: Probabilidad de sesgo extremo inicial y su valor límite comenzando con 8, 9, 10, 11, 12 y hasta 13 células.

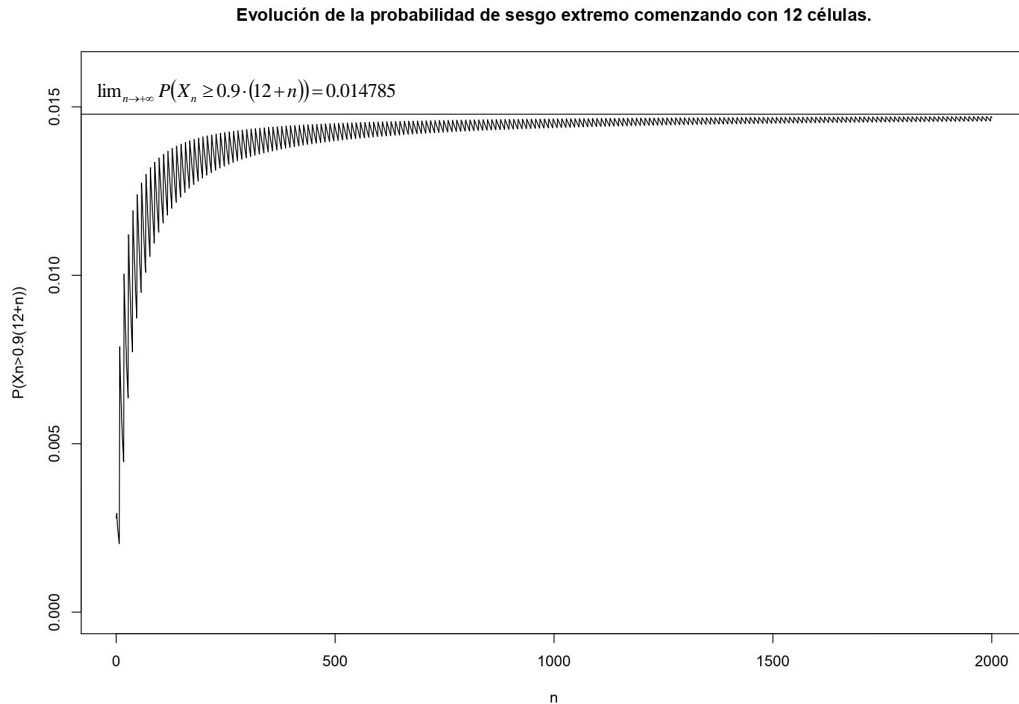


Figura 4.7: Probabilidad de sesgo extremo durante la expansión clonal partiendo de 12 células. La recta $y = 0,014785$ representa el $\lim_{n \rightarrow \infty} P(X_n \geq 0,9 \cdot (n + 12))$.

N_0	$P(X_0 \geq 0,9 \cdot N_0)$	$\lim_{n \rightarrow \infty} P(X_n \geq 0,9 \cdot (n + N_0))$
12	0,002792	0,014785

4.3. PROCESOS DE DERIVA APLICADOS AL X

En el gráfico de la Figura 4.7 se observa que en las primeras generaciones hay un aumento de la probabilidad de sesgo extremo inicial porque la deriva debida a la asincronicidad es significativa cuando el número de células que se agregaron es todavía pequeño, comparado con el número de células de un tejido adulto que suponemos mayor o igual a 10^6 [37] (ver Subsección 4.3.1). Además, en base a los cálculos se tiene que la probabilidad de que inicialmente el porcentaje de células con el X materno inactivo sea mayor o igual al 90 % del total de células, si bien tiende a aumentar a cada paso, luego de unas pocas generaciones (alrededor de 2000) se aproxima al valor asintótico de $P(X_n \geq 0,9 \cdot (n + 12))$, que se calculó a partir de la fórmula (3.2.18). Esto indica que durante la expansión clonal, aplicando nuestro modelo, se puede producir un sesgo extremo por deriva con una probabilidad mayor que la probabilidad de sesgo extremo inicial.

Concluimos que si se considerara una clonación no sincrónica, las células precursoras en el momento de la inactivación podrían no ser 8 sino 12, que según la bibliografía [24] es un número más acorde con el número de células que hay al momento del proceso de Lyonización. En efecto, después de unas pocas generaciones se alcanzaría e incluso superaría levemente la probabilidad de sesgo extremo observada empíricamente por Shvetsova *et al.* (como se representa en la Figura 4.8), explicando esta probabilidad a partir de 12 y no de 8 células iniciales.

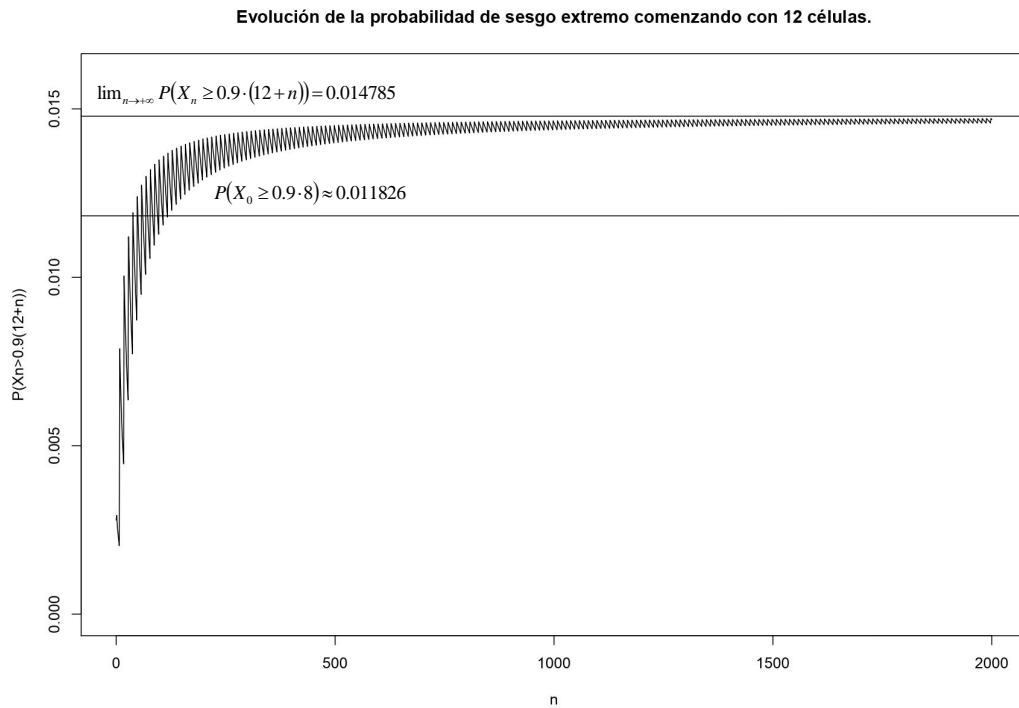


Figura 4.8: Probabilidad de sesgo extremo durante la expansión clonal partiendo de 12 células. La recta $y = 0,011826$ representa la probabilidad de sesgo extremo inicial con 8 células.

4.4. Enfermedades ligadas al cromosoma X

4.4.1. Hemofilia

En esta sección aplicamos los resultados obtenidos a genética de la hemofilia, una coagulopatía hereditaria ligada al cromosoma X que provoca hemorragias internas y externas y aparece por variantes patogénicas en individuos sanos afectando a sus descendientes.

La gravedad de las manifestaciones clínico-hemorrágicas de los pacientes está relacionada con los valores de actividad del factor de coagulación plasmático. Sin una adecuada presencia de este factor, la sangre no puede coagular adecuadamente para detener el sangrado. La actividad del factor está regulada por genes presentes

4.4. ENFERMEDADES LIGADAS AL CROMOSOMA X

en el cromosoma X. A partir de los experimentos pioneros con plasma de pacientes con hemofilia del Dr. Alfredo Pavlovsky a mitad del siglo XX, la hemofilia puede clasificarse en la hemofilia A y la hemofilia B [22]. La hemofilia A, identificada como HA está asociada a un defecto en el factor VIII de coagulación (FVIII) en plasma. La hemofilia B, identificada como HB se caracteriza por el déficit del factor IX de coagulación (FIX) en plasma. Ambos grupos son prácticamente indistinguibles clínicamente, es decir, en base a los síntomas del paciente.

En una enfermedad recesiva ligada al cromosoma X como la hemofilia, la condición se manifiesta en el varón si tiene afectado su único cromosoma X, mientras que la mujer heterocigota (con uno sólo de los dos cromosomas X afectados), que se denomina portadora, usualmente no padece síntomas de la enfermedad. Se suelen llamar simplemente portadoras a las portadoras heterocigotas. Cabe señalar que las portadoras homocigotas, o heterocigotas combinadas (doble heterocigota) afectando ambos cromosomas X con variantes patogénicas, son un porcentaje mínimo de la población dada la baja probabilidad de que se produzcan.

Expresión de hemofilia en mujeres

Si bien los portadores sintomáticos son usualmente los pacientes varones hemocigotas, también pueden ser las mujeres heterocigotas según su patrón de inactivación cromosómico en el hígado, el tejido de importancia en la producción del factor de coagulación.

De acuerdo a lo que vimos en la Subsección 4.2.2, como consecuencia de la aleatoriedad del proceso de inactivación del cromosoma X, se espera que una mujer portadora presente aproximadamente la mitad de actividad del FVIII o FIX respecto a un individuo de coagulación normal, lo cual es suficiente para que no presente síntomas. Por el contrario, cuando intervienen las causas (2) y (3) el proceso de inactivación de cualquiera de los dos cromosomas X no es al azar y puede producirse un sesgo extremo. Si este fenómeno de inactivación sesgada ocurre en una mujer portadora encontramos dos posibilidades: (a) que la inactivación se haya producido hacia el cromosoma X afectado, por lo tanto aumenta la producción del factor de coagulación; (b) que por el contrario esté inactivado preferentemente el X no afectado en la mayoría de las células, en consecuencia disminuya la actividad del factor de coagulación, y por lo tanto, se manifiesten síntomas de hemofilia [24].

Aunque no suele presentarse evidencia fenotípica de enfermedad en desórdenes recesivos ligados al cromosoma X, existen algunas mujeres que son sintomáticas.

La mayoría de estos casos son mujeres heterocigotas para mutaciones en el FVIII con síntomas de HA (o FIX con síntomas de HB) asociados al sesgo extremo hacia la inactivación del cromosoma X no afectado ([25], [26], [27] y [28]).

Correlación entre tejidos del patrón de inactivación

La determinación de la inactivación sesgada del cromosoma X es de interés para poder explicar por qué ciertas mujeres portadoras de hemofilia padecen síntomas propios de la enfermedad. El tejido de importancia en la producción del factor de coagulación es del hígado, por lo tanto, interesa conocer el patrón de inactivación del cromosoma X en dicho tejido en mujeres portadoras, y estudiar su eventual sesgo. Sin embargo, este tejido es de difícil acceso para obtener muestras celulares, para posteriormente realizar los análisis que determinan el XIP. Entonces el estudio del patrón de inactivación del cromosoma X en mujeres portadoras de hemofilia requiere una alta correlación del sesgo en el patrón de inactivación en diferentes tejidos, para poder estimar el sesgo en el hígado a partir del diagnóstico en tejidos más accesibles como la sangre periférica, mucosa bucal o células uretrales descamadas de la orina. Si se asume que existe esta correlación, un patrón de inactivación con sesgo extremo en sangre periférica, mucosa bucal o células uretrales descamadas de la orina podría utilizarse para determinar la causa de la expresión de síntomas en mujeres portadoras de la enfermedad [1].

4.4.2. Modelo de deriva y correlación entre tejidos

Radic *et al.* [28] afirman que es de esperar que cuando exista una causa sistémica (predisposición genética) para que se inactive preferencialmente un cromosoma X en todo el organismo, el XIP medido en sangre periférica será parecido a aquel presente en otros tejidos. Más precisamente, establecen que existe una correlación entre los niveles de FVIII:C y el patrón de inactivación del cromosoma X (XIP), en mujeres portadoras heterocigotas de mutaciones severas del factor VIII causales de HA en muestras de sangre periférica. Por lo tanto, tanto las células sanguíneas como los hepatocitos productores del FVIII presentarán patrones de inactivación similares. Luego, si se presenta sesgo extremo en los tejidos observados, sangre periférica y mucosa bucal, se puede inferir que el tejido productor del factor de coagulación también presenta sesgo extremo, y como vimos, esto podría explicar la sintomatología en una paciente portadora. Sin embargo, en aquellos casos donde

4.4. ENFERMEDADES LIGADAS AL CROMOSOMA X

los niveles de XIP en sangre periférica no permiten explicar los síntomas de sangrado en las mujeres portadoras heterocigotas de HA, o los altos niveles de FVIII:C, cercanos al valor medio esperado para no portadoras, sería de interés estimar el patrón de inactivación en el hígado, tejido involucrado en la expresión del FVIII (o FIX).

Considerando los aportes del artículo de Radic *et al.*, si el proceso de inactivación es aleatorio, es decir, si no existiera una predisposición genética a sesgar, no es posible suponer que existe una correlación entre tejidos para inferir el patrón de inactivación del cromosoma X en el hígado a partir de otros tejidos más accesibles.

Los resultados expuestos en esta tesis nos permiten estudiar si la divergencia entre los patrones de inactivación de distintos tejidos puede acentuarse por la deriva en el proceso de expansión clonal y en el proceso de renovación celular.

Primero analizamos si el patrón de inactivación a lo largo de la vida de una mujer, como consecuencia de los eventos de renovación celular, podría cambiar por deriva. Abordamos este problema a partir de una variante del modelo de Wright-Fisher de deriva genética, que desarrollamos en la Subsección 4.3.1, que permite describir el cambio en la frecuencia de células humanas con inactivación del cromosoma X materno o paterno. Obtuvimos una fórmula que permite calcular el número de células que mueren (r) en función del número de células que se replican (s) en cada episodio de renovación celular para que se produzca un sesgo en el patrón de inactivación respecto al patrón de inactivación inicial. Para utilizar esta fórmula fue necesario estimar el número de células en un tejido (N) y el número promedio de episodios de renovación celular a lo largo de la vida de una mujer (n). A partir del análisis realizado encontramos aquellas combinaciones de los parámetros s y r para las cuales la deriva podría ser significativa. Como también las combinaciones de s y r para las cuales nuestro modelo predice que un aumento de la divergencia del patrón de inactivación entre tejidos de una mujer es poco probable por deriva. Lo cual sería consistente con el trabajo de Vickers *et al.* [37], que establece que el sesgo en el patrón de inactivación en el tiempo promedio de vida de una persona no llegará a presentar cambios significativos por el puro azar involucrado en estos episodios de renovación.

En segundo lugar, abordamos la posibilidad de deriva significativa de una manera que no aparece en la bibliografía a la que hemos accedido. Estudiamos si el patrón inicial de inactivación del cromosoma X podría cambiar durante la expansión clonal de las células que tiene lugar en el proceso de crecimiento de un tejido

CAPÍTULO 4. APLICACIONES

mientras tiene un número reducido de células. Para ello modelamos la expansión clonal durante el desarrollo embrionario a partir de una variante del modelo de urna de Polya-Eggenberger desarrollado en la Subsección 4.3.4, bajo el supuesto de que en el momento de inactivación hay 12 células precursoras de un tejido. Calculamos la probabilidad de sesgo extremo en cada generación, es decir, la probabilidad de que el 90% o más de las células tengan la misma inactivación, que denotamos por $P(X_n \geq 0,9 \cdot (12 + n))$, como también su valor asintótico. A partir de este modelo encontramos que la probabilidad de sesgo extremo en el patrón de inactivación inicial, suponiendo que la expansión clonal no es sincrónica, tiende a aumentar por deriva durante la expansión clonal cuando todavía el número de células es pequeño, mientras que cuando se agregaron alrededor de 2000 células dicha probabilidad se aproxima al valor asintótico. Por lo tanto, según nuestro modelo podemos concluir que podría haber portadoras sintomáticas con una probabilidad mayor que la probabilidad de sesgo extremo en el patrón de inactivación inicial dada por las colas de una distribución binomial con parámetros $N_0 = 12$ y $p = 1/2$. Estas portadoras sintomáticas podrían presentar un patrón de inactivación no sesgado en los tejidos observables, porque la inactivación inicial y la posterior deriva que estamos estudiando son independientes en distintos tejidos. Concluimos que puede haber sesgo por puro azar, pero al producirse por procesos independientes en distintos tejidos, este tipo de sesgo no implica correlación.

Conclusión

Esta tesis se inscribe en la cercana relación existente entre la Teoría de Probabilidad y la Genética. Estudia los modelos de deriva genética de Wright-Fisher y el modelo de urna de Polya-Eggenberger proponiendo nuevas variantes, generalizaciones y parámetros de interés, y aportando demostraciones de sus propiedades. Los resultados se dirigen a generar herramientas para estimar la pérdida de diversidad en áreas en las que la bibliografía reciente señala la necesidad de cuantificar la deriva, en particular en la composición alélica en plagas y en la evolución del sesgo en el patrón de inactivación del cromosoma X en mujeres.

Conclusión

Apéndice

Determinante de la matriz de Vandermonde

Proposición 1. Para x_0, x_1, \dots, x_n todos distintos entre sí y

$$V = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix}$$

se tiene

$$\det V = \prod_{0 \leq i < j \leq n} (x_j - x_i).$$

Demostración. La prueba se realiza por inducción. En el caso $n = 1$ tanto $\det V$ como $\prod_{0 \leq i < j \leq n} (x_j - x_i)$ dan $x_1 - x_0$. Para llevar a cabo el paso inductivo, en primer lugar multiplicamos V por una matriz E de determinante igual a 1 cuyo efecto es que, para toda columna k salvo la primera, multiplica la columna anterior por x_0 y se la resta a la columna k . Esta matriz E tiene unos en la diagonal, $-x_0$ en cada posición de la subdiagonal superior y ceros en el resto de sus posiciones. Nos permite obtener

$$VE = \begin{pmatrix} 1 & x_0 - x_0 & \dots & x_0^n - x_0^{n-1}x_0 \\ 1 & x_1 - x_0 & \dots & x_1^n - x_1^{n-1}x_0 \\ \vdots & \vdots & x_i^k - x_i^{k-1}x_0 & \vdots \\ 1 & x_n - x_0 & \dots & x_n^n - x_n^{n-1}x_0 \end{pmatrix}$$

$$\begin{aligned}
 &= \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & x_1 - x_0 & \dots & (x_1 - x_0)x_1^{n-1} \\ \vdots & \vdots & (x_i - x_0)x_i^{k-1} & \vdots \\ 1 & x_n - x_0 & \dots & (x_n - x_0)x_n^{n-1} \end{pmatrix} \\
 &= \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & x_1 - x_0 & 0 & \vdots \\ \vdots & 0 & x_i - x_0 & 0 \\ 0 & \dots & 0 & x_n - x_0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \frac{1}{x_1 - x_0} & 1 & x_1 & \dots & x_1^{n-1} \\ \vdots & \vdots & \vdots & x_i^{k-1} & \vdots \\ \frac{1}{x_n - x_0} & 1 & x_n & \dots & x_n^{n-1} \end{pmatrix}.
 \end{aligned}$$

Luego $\det V = \det V \det E = \det VE$ donde VE resultó ser a su vez el producto de otras dos matrices, una con determinante igual a $(x_1 - x_0)\dots(x_n - x_0)$ y la otra con determinante igual a $1 \cdot \prod_{1 \leq i < j \leq n} (x_j - x_i)$ (por hipótesis inductiva). Multiplicando esos determinantes obtenemos $\det V = (x_1 - x_0)\dots(x_n - x_0) \prod_{1 \leq i < j \leq n} (x_j - x_i) = \prod_{0 \leq i < j \leq n} (x_j - x_i)$ como queríamos.

Corolario 1. *La matriz de Vandermonde es no singular.*

Cadenas de Markov con estados absorbentes

Definición 1. *Sea E un conjunto finito o numerable. Se dice que cada $x \in E$ es un estado y E es el espacio de estados. Una cadena de Markov de tiempo discreto con espacio de estados E es una sucesión de variables aleatorias $\{X_n : n \geq 0\}$ tales que para todo $x_0, \dots, x_n \in E$*

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}).$$

Esta igualdad se denomina propiedad de Markov.

La probabilidad de transición se define como la probabilidad de pasar del estado i al estado j en el paso n , $P(X_n = j | X_{n-1} = i)$, que se denota $p_{ij}(n)$.

Definición 2. *Una cadena de Markov de tiempo discreto se dice homogénea si*

$$P(X_n = x_n | X_{n-1} = x_{n-1}) = P(X_{n+m} = x_{n+m} | X_{n+m-1} = x_{n+m-1}),$$

es decir, $p_{ij}(n)$ no depende de n y simplemente se nota como p_{ij} .

Definición 3. En una cadena de Markov de tiempo discreto el estado j es accesible desde el estado i si existe un número de pasos n tal que $P(X_n = j | X_0 = i) > 0$.

Definición 4. Sea $\{X_n : n \geq 0\}$ una cadena de Markov homogénea de tiempo discreto con conjunto de estados finito E . Se dice que $E_0 \subset E$ es un conjunto absorbente si $P(X_1 \in E_0 | X_0 \in E_0) = 1$.

Definición 5. Sea $\{X_n : n \geq 0\}$ una cadena de Markov homogénea de tiempo discreto con conjunto de estados finito E . Decimos que $A \subset E$ es accesible desde $B \subset E$ si para todo $i \in A$ existe $j \in B$ tal que i es accesible desde j .

Definición 6. Sea $\{A_n : n \geq 0\}$ una sucesión de eventos, subconjuntos de un espacio Ω . Entonces el evento “ A_n ocurre infinitas veces” denotado mediante $\{A_n, i.v.\}$ se define como $\lim_{m \rightarrow \infty} \cup_{n \geq m} A_n$, lo cual significa $\cap_{m \geq 0} \cup_{n \geq m} A_n$ y coincide con $\{\omega \in \Omega : \forall m \exists n \geq m : \omega \in A_n\}$.

Proposición 1. Sea $\{X_n : n \geq 0\}$ una cadena de Markov homogénea de tiempo discreto con conjunto de estados finito E y $E_0 \subset E$ un conjunto absorbente. Sea τ el tiempo aleatorio en el que el proceso alcanza un elemento de E_0 por primera vez. Entonces si E_0 es accesible desde $E - E_0$ se tiene que para cualquier distribución inicial de X_0 , $P(\tau < \infty) = 1$, es decir, la probabilidad de alcanzar el conjunto absorbente en un número finito de pasos es 1.

Demostración. Sea $e = |E|$. Si algún estado de E_0 es accesible desde $i \in E - E_0$ entonces $P(X_e \in E_0 | X_0 = i) > 0$ porque si se llega con probabilidad positiva, existe un camino de probabilidad positiva que no repite estados. Sea $p = \min_{i \in E - E_0} P(X_e \in E_0 | X_0 = i)$ y para $k = 1, 2, \dots$ sea $A_k = \{X_{ke} \in E - E_0\}$. Luego para cualquier distribución inicial de X_0 tenemos $A_1 \supset A_2 \supset A_3 \supset \dots$ y $P(A_1) \leq 1 - p$.

Ahora calculemos $P(A_2)$:

$$\begin{aligned} P(A_2) &= P(X_{2e} \in E - E_0) \\ &= P(X_{2e} \in E - E_0 | X_e \in E - E_0) P(X_e \in E - E_0) \\ &\quad + P(X_{2e} \in E - E_0 | X_e \in E_0) P(X_e \in E_0) \\ &\leq (1 - p)(1 - p) + 0 \cdot P(X_e \in E_0) = (1 - p)^2. \end{aligned}$$

Apéndice

En general se tiene:

$$\begin{aligned} P(A_k) &= P(X_{ke} \in E - E_0) \\ &= P(X_{ke} \in E - E_0 | X_{(k-1)e} \in E - E_0) P(X_{(k-1)e} \in E - E_0) \\ &\quad + P(X_{ke} \in E - E_0 | X_{(k-1)e} \in E_0) P(X_{(k-1)e} \in E_0) \\ &\leq (1-p)(1-p)^{k-1} + 0 \cdot P(X_{(k-1)e} \in E_0) \\ &= (1-p)^k. \end{aligned}$$

Ahora, dado que $P(\tau < \infty) = P(\{A_k, i.v.\}^c)$, queremos ver que $P(A_k, i.v.) = 0$.

Tenemos que $P(A_k, i.v.) = P(\lim_{m \rightarrow \infty} \cup_{k \geq m} A_k)$ y como $A_1 \supset A_2 \supset A_3 \supset \dots$ resulta $\cup_{k \geq m} A_k = A_m$.

Luego $P(A_k, i.v.) = P(\lim_{m \rightarrow \infty} A_m)$, que por la propiedad de continuidad de la probabilidad es $\lim_{m \rightarrow \infty} P(A_m)$.

Finalmente se tiene que $P(A_k, i.v.) \leq \lim_{m \rightarrow \infty} (1-p)^m = 0$.

Bibliografía

- [1] Bittel, D. C., Theodoro, M. F., Kibiryeve, N., Fischer, W., Talebizadeh, Z. y Butler, M. G. (2008). Comparison of X-chromosome inactivation patterns in multiple tissues from human females. *Journal of medical genetics*, 45(5), 309–313. <https://doi.org/10.1136/jmg.2007.055244>
- [2] Caballero, A. (1994). Developments in the prediction of effective population size. *Heredity*, 73, 657–679. <https://doi.org/10.1038/hdy.1994.174>
- [3] Cannings, C. (1973). The Equivalence of Some Overlapping and Non-Overlapping Generation Models for the Study of Genetic Drift. *Journal of Applied Probability*, 10(2), 432-436. <https://doi:10.2307/3212359>
- [4] Cannings, C. (1974). The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models. *Advances in Applied Probability*, 6(2), 260-290. <https://doi:10.2307/1426293>
- [5] De Brasi, C., Maffei, A. L. y Tetzlaff, T. (Del 30 de julio al 2 de agosto de 2012). *Paseos al azar coalescentes en deriva cromosómica*. 5^a Escuela Argentina de Matemática y Biología BIOMAT. La Falda, Córdoba, Argentina.
- [6] De Brasi, C., Maffei, A. L. y Tetzlaff, T. (Del 13 al 17 de julio de 2015). *Dispersion of drift processes with application to X inactivation in different tissues*. Congreso latinoamericano de biomatemática, SOLABIMA 2015. Botucatu, Brasil.
- [7] De Brasi, C., Maffei, A. L. y Tetzlaff, T. (Del 1 al 5 de agosto de 2016). *Modelo de deriva y selección en el patrón de inactivación de cromosomas X con aplicación a hemofilia*. 7^{ta} Escuela Argentina de Matemática y Biología. Huerta Grande, Córdoba, Argentina.

BIBLIOGRAFÍA

- [8] De Brasi C., Maffei A. L. y Tetzlaff T. (Del 21 al 25 de septiembre 2020). *Esquema de Polya-Eggenberger aplicado a deriva cromosómica*. virtUMA 2020, LXIX Reunión Anual de Comunicaciones Científicas de la Unión Matemática Argentina.
- [9] Durrett, R. (2008). *Probability models for DNA sequence evolution*. Probability and its Applications. Springer.
- [10] Ewens, W. J. (1979). *Mathematical Population Genetics*. Springer.
- [11] Ewens, W. J. (2004). *Mathematical Population Genetics 1: Theoretical Introduction*. Springer.
- [12] Feller, W. (1951). *Diffusion processes in genetics*. Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, Los Angeles.
- [13] Fisher, R. (1930). *The genetical theory of natural selection*. Oxford, The Clarendon Press.
- [14] Greco, N.M., Tetzlaff, G.T. y Liljesthröm, G.G. (2004). Presence-absence sampling for *Tetranychus urticae* and its predator *Neoseiulus californicus* (Acari: Tetranychidae; Phytoseiidae) on strawberries. *International Journal of Pest Management*, 50(1), 23-27. <http://dx.doi.org/10.1080/09670870310001626338>
- [15] Hartl, D. L. y Clark, A. G. (1997). *Principles of population genetics*. Sinauer Associates, Inc. Publishers.
- [16] Lizarbe Iracheta, M. A. (2007). El suicidio y la muerte celular. *Real Academia de Ciencias Exactas, Físicas y Naturales*, 101(2),419-454.
- [17] Lyon, M. F. (1961). Gene action in the X-chromosome of the mouse. *Nature*, 190, 372–373. <https://doi.org/10.1038/190372a0>
- [18] Maffei, A. L. y Tetzlaff, T. (Del 7 al 11 de agosto de 2017). *Números efectivos para un modelo de deriva con renovación parcial de la población*. Congreso latinoamericano de biomatemática, SOLABIMA 2017. Cusco, Perú.
- [19] Mahmoud, H. M. (2008). *Pólya Urn Models*. CHAPMAN & HALL/CRC.
- [20] Moreno Letelier, A. (2007). *Tamaño efectivo de la población*. En: Eguiarte, L. E., Souza, V., Aguirre, X. (Comps.). *Ecología Molecular*. SEMARNAT. Ciudad de México. pp. 63-84.

-
- [21] Nagylaki, T. (1992). *Introduction to theoretical population genetics*. Springer.
- [22] Pavlovsky, A. (1947). Contribution to the pathogenesis of hemophilia. *Blood*, 2(2), 185-91.
- [23] Puck, J. y Willard, H. F. (1998). X inactivation in females with X-linked disease. *New England Journal of Medicine*, 338(5), 325-328. DOI:10.1056/NEJM199801293380611
- [24] Radic, Claudia Pamela. (2010). *Genética molecular de hemofilia: caracterización de mutaciones en hemofilia B, expresión de hemofilia en mujeres y desarrollo de nuevos métodos de análisis de inversiones*. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.
- [25] Radic, P., Tetzlaff, T., Bonduel, M., Candela, M., Rossetti, L. C. y De Brasi, C. (2010). X Inactivation Patterns and FVIII:C Activity Levels in Female Carriers and Non-Carriers of Severe Factor VIII Gene Mutations. *Haemophilia*, 16(4), 18-18.
- [26] Radic, P., Tetzlaff, T., Candela, M., Rossetti, L. C., Larripa, I. y De Brasi, C. D. (2010). Patrones de inactivación del cromosoma X y niveles de actividad del FVIII:C en mujeres portadoras y no portadoras de mutaciones causales de Hemofilia A severa. *Medicina*, 70(2), 491-491.
- [27] Radic, P., Rossetti, L., Tetzlaff, T., Abelleiro, M., Sciuccati, G., Candela, M., Bonduel, M. y De Brasi, C. (2011). *Análisis de los mecanismos moleculares que condicionan la expresión clínica de hemofilia en mujeres*. Premio Adolfo H. Aztiria de la Academia Nacional de Medicina de Buenos Aires, Argentina, al mejor trabajo sobre Investigación Clínica.
- [28] Radic, C. P., Rossetti, L. C., Abelleiro, M. M., Tetzlaff, T., Candela, M., Neme, D., Sciuccati, G., Bonduel, M., Medina-Acosta, E., Larripa, I. B., de Tezanos Pinto, M. y De Brasi, C. D. (2015). Phenotype-genotype correlations in hemophilia A carriers are consistent with the binary role of the phase between F8 and X-chromosome inactivation. *Journal of thrombosis and haemostasis: JTH*, 13(4), 530–539. <https://doi.org/10.1111/jth.12854>
- [29] REX Consortium. (2007). Structure of the scientific community modelling the evolution of resistance. *PLOS ONE*, 2(12): e1275. <https://doi.org/10.1371/journal.pone.0001275>

BIBLIOGRAFÍA

- [30] REX Consortium. (2010). The skill and style to model the evolution of resistance to pesticides and drugs. *Evolutionary applications*, 3(4), 375–390. <https://doi.org/10.1111/j.1752-4571.2010.00124.x>
- [31] Shvetsova, E., Sofronova, A., Monajemi, R. *et al.* (2019). Skewed X-inactivation is common in the general female population. *European Journal of Human Genetics*, 27, 455–465. <https://doi.org/10.1038/s41431-018-0291-3>
- [32] Stern, V., Smith, R., van den Bosch R. y Hagen, K. (1959). The integration of chemical and biological control of the spotted alfalfa aphid: The integrated control concept. *Hilgardia*, 29(2), 81-101. DOI:10.3733/hilg.v29n02p081
- [33] Tavaré, S. (2004). *Ancestral Inference in Population Genetics*. Springer.
- [34] Trumper, E. V. (2014). Resistencia de insectos a cultivos transgénicos con propiedades insecticidas: Teoría, estado del arte y desafíos para la República Argentina. *Agriscientia*, 31, 2, 109–126. <https://doi.org/10.31047/1668.298x.v31.n2.16538> G
- [35] Vélez, R. y Prieto, T. (2018). *Limit Behavior of Polya Urn Schemes*. In: Gil E., Gil E., Gil J., Gil M. (eds) *The Mathematics of the Uncertain. Studies in Systems, Decision and Control* (pp. 421–429). Springer, Cham.
- [36] Vickers, M. A., Brown, G. C., Cologne, J. B. y Kyoizumi, S. (2000). Modelling haemopoietic stem cell division by analysis of mutant red cells. *British Journal of Haematology*, 110, 54–62. <https://doi.org/10.1046/j.1365-2141.2000.02134.x>
- [37] Vickers, M. a., McLeod, E., Spector, T. D. y Wilson, I. J. (2001). Assessment of mechanism of acquired skewed X inactivation by analysis of twins. *Blood*, 97(5), 1274-1281. <https://doi.org/10.1182/blood.V97.5.1274>
- [38] Wright, S. (1930). Evolution in mendelian populations. *Genetics*, 16(2), 97–158.