
DEPENDENCIA ESTOCÁSTICA EN GENÉTICA

TRABAJO FIN DE GRADO

Autor:

Ana Almansa Carricondo

Tutor:

Alicia María Juan González

GRADO EN MATEMÁTICAS



SEPTIEMBRE, 2016
Universidad de Almería

Índice general

1	Objetivos	1
2	Antecedentes	3
3	Fundamento matemático	7
3.1.	Procesos estocásticos	7
	Definición y descripción, 7.— Funciones de distribución de transición., 9.— Homogeneidad en el tiempo, 10.— Características de un proceso estocástico, 10.— Clases de procesos: procesos de Markov, 11.	
3.2.	Cadenas de Markov	12
	Ecuaciones de Chapman-Kolmogorov, 13.— Matrices de transición, 14.— Cadenas de Markov homogéneas, 15.— Clases de estados, 16.— Estados recurrentes y transitorios, 18.— Estados recurrentes positivos y recurrentes nulos, 19.— Estados periódicos, 20.— Estados absorbentes, 20.— Cadenas de Markov irreducibles y ergódicas: comportamiento límite, 22.	
4	Aplicación: el modelo de Wright-Fisher	25
4.1.	Principio de Hardy-Weinberg	26
4.2.	Apareamiento aleatorio sin mutación	30
	Heterocigosidad, 32.	
4.3.	Apareamiento aleatorio con mutación	36
5	Conclusiones	41
	Bibliografía	43

Abstract

A lot of studies have proved that there is a narrow relationship between genetics and mathematics. In 1908, the British mathematician G.H Hardy and the German doctor Wilhelm Weinberg proved that the composition of a population that mates at random remains constant at equilibrium in the absence of endogenous and exogenous factors, such as mutation, natural selection and migrations. This result is known as “the Hardy-Weinberg principle”. Subsequently, various mathematicians and biologists reduced their study of genetic evolution to finite populations and they set the mathematical basis for evolution. Among these illustrious figures, it is worth mentioning the British statistician and biologist Ronald Aylmer Fisher and the American geneticist Sewall Green Wright, who developed the “Wright-Fisher model”.

The purpose of our work is to develop the Wright-Fisher model, which is considered one of the simplest stochastic models for ADN evolution, in such a way that we will start introducing the necessary mathematical content, which is based on the theory of stochastic processes and, in particular, on Markov chains.

Afterwards, we will apply those contents to analyze two versions of the Wright-Fisher model: the first one, without mutation, and the second one, with mutation, including examples for each case.

Resumen

Numerosos estudios han demostrado que existe una estrecha relación entre la genética y las matemáticas. En el año 1908, el matemático inglés G.H Hardy y el médico alemán Wilhelm Weinberg demostraron que la composición de una población que se aparee al azar, permanece en equilibrio en ausencia de otros factores, endógenos y exógenos, tales como la mutación, la selección natural o las migraciones. Este resultado es conocido como “principio de Hardy - Weinberg”. Posteriormente, diversos matemáticos y biólogos redujeron su estudio de la evolución genética a poblaciones finitas y sentaron las bases matemáticas de la evolución. Entre estos ilustres personajes cabe destacar al estadístico y biólogo británico Ronald Aylmer Fisher y al genetista estadounidense Sewall Green Wright, los cuales desarrollaron el llamado “modelo de Wright - Fisher”.

El objetivo de nuestro trabajo es desarrollar dicho modelo, considerado como uno de los modelos estocásticos más sencillos de evolución del ADN, de modo que comenzaremos introduciendo el contenido matemático necesario, y que se basa en la teoría de los procesos estocásticos y, en concreto, las cadenas de Markov.

Posteriormente, aplicamos dichos contenidos para analizar dos variantes del modelo de Wright-Fisher: la primera, sin mutación, y la segunda con ella, incluyendo ejemplos de cada caso.

Objetivos

En el año 1953, los científicos Francis Crick y Jim Watson publicaron un modelo para la estructura de la molécula del ácido desoxirribonucleico (ADN), el responsable de la transmisión de la información genética de generación en generación. A partir de entonces los avances en genética han sido espectaculares, dando lugar a técnicas que están influyendo e influirán notablemente tanto en el mundo social mediante la terapia génica como en el económico con la producción de transgénicos.

Desde las bases estadísticas de las primeras leyes de Mendel en 1865, las matemáticas han ido acompañado este desarrollo en el campo de la biología y la genética como lo demuestran: los modelos probabilistas de transmisión de caracteres; las ecuaciones en derivadas parciales no lineales para describir la constitución de formas características de ciertas especies como, por ejemplo, los dibujos de las alas de las mariposas, las manchas de la piel de los leopardos; elementos de topología y geometría para entender cómo se forman las estructuras de las proteínas y cómo esta estructura determina su función, métodos de computación sofisticados para explotar e interpretar gran cantidad de datos,...

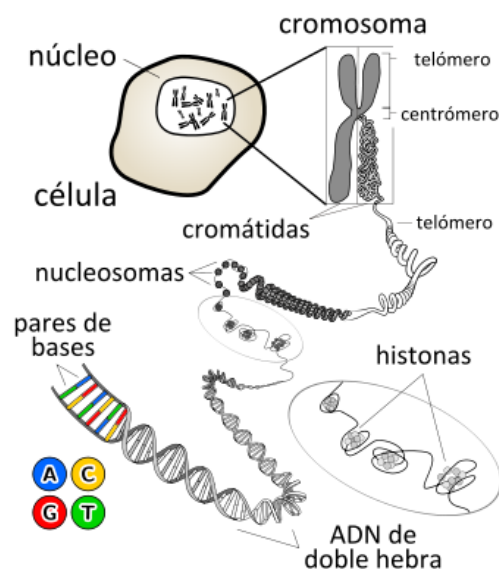
El objetivo de este trabajo es describir el modelo de Wright-Fisher, que es uno de los modelos aleatorios más simples de evolución del ADN en genética de poblaciones y hacerlo a través del tipo de proceso estocástico subyacente.

Antecedentes

El fenómeno mediante el cual los caracteres de los seres vivos se transmiten de padres a hijos recibe el nombre de herencia biológica y la ciencia que estudia este fenómeno es la Genética.

La Genética clásica o Mendeliana parte de los descubrimientos de Gregorio Mendel que, entre 1858-1866, realizó cruzamientos entre individuos de la misma especie pero de razas distintas, que se distinguían perfectamente en algún carácter. Estudió cómo se transmitían esos caracteres de generación en generación, llegando a deducir las leyes que regulan esta transmisión. El gran mérito de los trabajos de Mendel se debe a que se realizaron cuando aún no se conocía el ADN, ni se sabía cómo pasa la información genética de padres a hijos y, a pesar de este desconocimiento, las Leyes formuladas por Mendel son ciertas y se pueden explicar hoy en base a lo que sabemos del ADN.

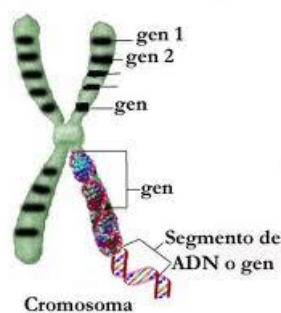
A partir del descubrimiento de la estructura del ADN y de su capacidad de replicación, se desarrolla la actual Genética Molecular, que ha permitido conocer cómo está dentro del ADN la información que determina los caracteres y cómo dicha información se transmite de padres a hijos. Así, cada célula de una especie tiene un número fijo de cromosomas. La especie humana, por ejemplo, tiene 46 cromosomas en el núcleo de cada célula pero, como somos organismos diploides, los cromosomas son semejantes dos a dos y así, tenemos 23 pares de cromosomas en cada célula.



Un cromosoma es un orgánulo en forma de filamento y está compuesto por moléculas de ADN. El ADN es el material hereditario. Cada molécula de ADN tiene forma de doble hélice, es como una escalera de cuerda floja suspendida en el espacio, cuyos peldaños están formados por bases nitrogenadas o nucleótidos, mientras que la baranda son azúcares y fosfatos. Las bases nitrogenadas son cuatro: la Guanina (G), la Citosina (C), la Adenina (A) y la Timina (T). Cada peldaño de la escalera está formado o bien por una pareja de G y C, o bien de A y T.

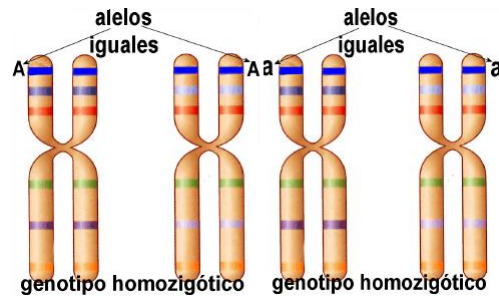


Los genes están en los cromosomas, se disponen alineados a lo largo del cromosoma siguiendo un orden fijo, y el lugar que ocupa cada gen es su locus. Cada gen, por tanto, es un segmento de ADN dispuesto en un lugar específico del cromosoma, y constituye la unidad de información genética para determinar una característica particular en un individuo.



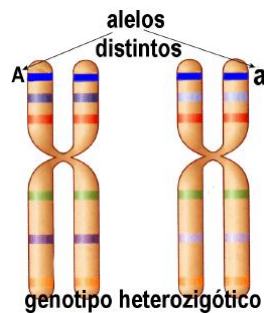
Los genes que informan sobre el mismo carácter se llaman alelos. Así, el gen que determina el color de los ojos, por ejemplo, puede tener diversos alelos: ojos verdes, azules, marrones, negros,... En los organismos diploides, cada carácter está regido por dos alelos situados en locus equivalentes de cromosomas homólogos. Por tanto, de los 2 alelos que rigen cada carácter, uno es de procedencia paterna y el otro de procedencia materna. Pueden darse dos situaciones:

1. Los 2 alelos que rigen un carácter son idénticos; se dice entonces que el individuo es homocigótico o puro para ese carácter. En este caso, los dos alelos se representan con la misma letra (por ejemplo, AA , aa). Por ejemplo, en el clásico experimento de los guisantes de Mendel, en el guisante existen 2 alelos para el color de la semilla: A , determina color amarillo; a determina color verde. Hay guisantes en los que los alelos son AA (color amarillo) y guisantes en los que los alelos son aa (color verde)
2. Los 2 alelos son distintos; se dice que el individuo es híbrido o heterocigótico para ese carácter. En el caso de los guisantes, si un alelo es A y el otro es a pueden presentarse, a su vez, dos situaciones:
 - a) Que de los 2 alelos que rigen el carácter, sólo uno de ellos se manifieste exteriormente, mientras que el otro no se manifiesta. El gen que se manifiesta se llama dominante y se representa con una letra mayúscula; el gen



que queda oculto se llama recesivo y se representa con la misma letra pero minúscula. Cuando esto ocurre se dice que la herencia es dominante. En el ejemplo anterior, el gen que determina color amarillo es dominante sobre el que determina color verde; por ello el alelo para el amarillo se representa con letra mayúscula (por ejemplo, A) y el alelo para el verde con la misma letra pero minúscula (a). Los guisantes heterocigóticos para el color, presentan la combinación de alelos Aa y su color es amarillo. Los genes recesivos sólo se manifiestan exteriormente cuando no van acompañados de ningún dominante, es decir, cuando están en homocigosis (aa).

- b) Que los 2 alelos tengan la misma “fuerza” para manifestarse; en este caso, el carácter que rigen se manifiesta exteriormente como una mezcla de ambos; se dice entonces que la herencia es intermedia. Los dos alelos pueden representarse también con letras mayúsculas y minúsculas, sin que eso signifique que el alelo representado con mayúscula sea el alelo dominante. Por ejemplo, cuando se cruzan dondiegos rojos con dondiegos blancos, los dondiegos resultantes tienen flores rosas; si llamamos R al alelo que determina color rojo y r al alelo para el blanco, los dondiegos rojos serían RR , los blancos rr y los rosas Rr .



El conjunto de todos los genes de un individuo constituye su material genético o *genotipo*. La interacción del genotipo con el ambiente del individuo, dan lugar a su *fenotipo*:

$$\text{genotipo} + \text{ambiente} \rightarrow \text{fenotipo}$$

Así, el fenotipo constituye definitiva los rasgos de un individuo, la manifestación externa de su genotipo como consecuencia del medio en el que se encuentra.

Fundamento matemático

Si el término estocástico es sinónimo de aleatorio, azar o incertidumbre, proceso estocástico significa entonces “fenómeno aleatorio en el transcurso del tiempo” y, en consecuencia, siempre que se observa un fenómeno empírico que se desarrolla en el tiempo de manera controlada por leyes de probabilidad, se está observando un proceso estocástico.

3.1 Procesos estocásticos

Definición y descripción

Formalmente, un proceso estocástico (p.e.) es un conjunto arbitrario de variables aleatorias (v.a.) $\{X_t; t \in T\}$ definidas sobre un mismo espacio de probabilidad (Ω, \mathcal{A}, P) , donde el índice t representa, generalmente, el tiempo, de modo que X_t es la variable aleatoria observada en el instante t . No obstante, t puede representar también espacio, longitud, superficie, volumen, etc. El conjunto de índices T , o rango del parámetro t , es el *espacio paramétrico* y puede ser infinito numerable

$$\{T = 0, \pm 1, \pm 2, \dots\} \quad T = \{0, 1, 2, \dots\}$$

o infinito no numerable

$$T = \{t; -\infty < t < \infty\} \quad T = \{t; t \geq 0\} \quad T = \{t; 0 \leq t \leq 1\}$$

En el primer caso, $\{X_t; t \in T\}$ es un p.e. *de parámetro discreto*, y notaremos $X_t \equiv X_n$ para cada $n \in T$; en el segundo caso, es un p.e. *de parámetro continuo* y notaremos $X_t \equiv X(t)$.

Los valores que toma el proceso $\{X_t\}$, esto es, las v.a. que lo constituyen, son sus *estados*. Un número real x es un valor posible o *estado* de un p.e. $\{X_t; t \in T\}$ si, y sólo si

$$\exists t \in T/P[x - h < X_t \leq x + h] > 0 \quad \forall h > 0$$

Se llama *espacio de estados* del p.e. y lo notaremos por E , al conjunto de sus posibles valores. Dicho espacio es discreto si consta de un número finito o infinito numerable de estados. En caso contrario, el espacio de estados es continuo.

Es útil clasificar los procesos estocásticos según la naturaleza de sus espacios paramétrico y de estados. En este sentido, todo p.e. pertenece a una y sólo una de las 4 clases siguientes:

- *de parámetro discreto y espacio de estados discreto*: Por ejemplo, cuando se observan las preferencias de un consumidor con respecto a 3 marcas de cierto producto, y X_n representa la marca consumida por dicho individuo en la n -ésima etapa. Si asociamos valores numéricos a las marcas, esto es,

$$A \longrightarrow 1, \quad B \longrightarrow 2, \quad C \longrightarrow 3$$

entonces el espacio de estados del p.e. es $E = \{1, 2, 3\}$.

- *de parámetro continuo y espacio de estados discreto*: Por ejemplo, la cantidad de inventario disponible, supuesto que los artículos inventariados son numerables, durante cierto intervalo de tiempo; o también el número de estudiantes que esperan el autobús durante el día.
- *de parámetro discreto y espacio de estados discreto*: La cantidad disponible de un producto no numerable (tela, carburante,...), cuando la observación se realiza en instantes discretos, por ejemplo, cuando tiene lugar el cierre diario del establecimiento que lo expende; o el volumen de papel en el n -ésimo contendedor de cierta localidad.
- *de parámetro continuo y espacio de estados continuo*: Por ejemplo, el nivel de líquido de un embalse durante cierto periodo de tiempo; o también, el tiempo que un trabajo o tarea tiene que esperar hasta que entra en servicio, es decir, hasta que empieza a desarrollarse. En este caso, el instante en el que el trabajo entra en servicio es el parámetro.

Cuando el espacio de estados de un p.e. es discreto, éste se conoce como *cadena* y, por tanto, tendremos cadenas en tiempo discreto $\{X_n\}$, y cadenas en tiempo continuo $\{X(t)\}$.

Del mismo modo que el objetivo del Cálculo de Probabilidades es el conocimiento de las leyes de probabilidad de las v.a. que describen las características de un fenómeno aleatorio,

$$X : (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B}, P_X) \rightarrow \mathcal{L}(X) \equiv P_X$$

el propósito fundamental de la Teoría de los Procesos Estocásticos es describir la ley de probabilidad de cualquier proceso,

$$\{X_t; t \in T\} : (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^T, \mathcal{B}^T, P_T) \rightarrow \mathcal{L}(\{X_t; t \in T\}) \equiv P_T$$

y que viene dada, en virtud del *Teorema de Consistencia de Kolmogorov*, por la familia de distribuciones de probabilidad finitodimensionales,

$$\mathcal{L}(\{X_t; t \in T\}) \equiv P_T \iff \{P_{T_n}; \forall T_n \subset T, \forall n \geq 1 \text{ finito}\} \text{ consistentes}$$

donde $P_{T_n} = \{P_{t_1, \dots, t_n} : \forall t_i \in T, \forall i = 1, \dots, n\}$ es, a su vez, la familia de todas las distribuciones de probabilidad n -dimensionales.

No obstante, puesto que las probabilidades son funciones de conjunto y, por tanto, más difíciles de manejar con los métodos del análisis clásico, podemos, en virtud del *Teorema de Correspondencia*, expresar la consistencia en términos de sus correspondientes funciones de distribución de probabilidad (fdp) y así,

$$\mathcal{L}(\{X_t; t \in T\}) \equiv P_T \iff \{F_{T_n}; \forall T_n \subset T, \forall n \geq 1 \text{ finito}\} \text{ consistentes}$$

es la familia de fdp finitodimensionales, donde

$$F_{T_n} = \{F_{t_1, \dots, t_n}; \forall t_i \in T, \forall i = 1, \dots, n\}$$

es, a su vez, la familia de todas las fdp n -dimensionales. La consistencia significa:

- a) Todas las permutaciones que se efectúen con los subíndices de las variables se corresponden con las que se realizan en los subíndices de T_n , es decir, hay simetría en las permutaciones de los subíndices:

$$F_{t_{i_1}, t_{i_2}, \dots, t_{i_n}}(x_{i_1}, x_{i_2}, \dots, x_{i_n}) = F_{t_1, t_2, \dots, t_n}(x_1, x_2, \dots, x_n)$$

para cada permutación $\{i_1, i_2, \dots, i_n\}$ de $\{1, 2, \dots, n\}$.

- b) La fdp de orden $n + 1$ es convergente a la fdp de orden n , esto es,

$$\begin{aligned} F_{t_1, \dots, t_n, t_{n+1}}(x_1, \dots, x_n, +\infty) &= \lim_{x_{n+1} \rightarrow +\infty} F_{t_1, \dots, t_n, t_{n+1}}(x_1, \dots, x_n, +\infty) \\ &= F_{t_1, \dots, t_n}(x_1, \dots, x_n) \end{aligned}$$

Funciones de distribución de transición.

La Teoría de los Procesos Estocásticos estudia familias de v.a. desde el punto de vista de su interdependencia. Las *funciones de distribución de transición* son las que determinan la relación de interdependencia entre las v.a. que constituyen el proceso, y quedan determinadas, a su vez, por las distribuciones finitodimensionales.

$$\forall l, \tau \in T \quad \text{con} \quad \tau < l \quad F(x, t|y, \tau) = P[X_t \leq x | X_\tau = y] \quad \forall x, y \in E$$

- Si $\{X_t; t \in T\}$ es una cadena (E es discreto), las funciones de distribución de transición determinan las *funciones masa de probabilidad de transición* y viceversa, dadas por

$$\forall n, k \quad \text{con} \quad k < n \quad p_{i,j}(k, n) = P[X_n = j | X_k = i] \quad \forall i, j \in E$$

si T es discreto; o bien

$$\forall \tau, t \quad \text{con} \quad \tau < t \quad p_{i,j}(\tau, t) = P[X_t = t | X_\tau = i] \quad \forall i, j \in E$$

si T es continuo. Se tiene entonces que

$$\forall \tau, t \quad \text{con} \quad \tau < t \quad F(j, t|i, \tau) = \sum_{k \leq j} p_{i,k}(\tau, t) \quad \forall i, j \in E$$

- Si $\{X_t; t \in T\}$ es un proceso (E es continuo), las funciones de distribución de transición determinan las *densidades de probabilidad de transición* y viceversa, dadas por

$$\forall n, k \quad \text{con} \quad k < n \quad f(x, n|y, k) = \frac{\partial F(x, n|y, k)}{\partial x} \quad \forall x, y \in E$$

si T es discreto; o bien

$$\forall \tau, t \quad \text{con} \quad \tau < t \quad f(x, t|y, \tau) = \frac{\partial F(x, t|y, \tau)}{\partial x} \quad \forall x, y \in E$$

si T es continuo. En cualquier caso,

$$\forall \tau, t \quad \text{con} \quad \tau < t \quad F(x, t|y, \tau) = \int_{-\infty}^x f(u, t|y, \tau) du \quad \forall x, y \in E$$

Homogeneidad en el tiempo

Un proceso estocástico se dice que es *homogéneo en el tiempo* cuando la probabilidad de ocurrencia de cierto suceso en un instante determinado, no depende del instante en sí, sino del tiempo transcurrido desde el instante inicial, es decir, si notamos por t_0 al instante inicial, la probabilidad de transición

$$F(x, t|y, t_0) = P[X_t \leq x | X_{t_0} = y]$$

no depende directamente de t , sino del intervalo de tiempo $t - t_0$, que llamamos *lapso*. En consecuencia, $F(x, t|y, t_0)$ es realmente una función de tres argumentos:

$$F(x, t|y, t_0) \equiv g(x, y, t - t_0)$$

y así, fijados $x, y \in E$ tendremos que $\forall t, \tau \in T$ con $t > \tau$, tales que $t - \tau = h > 0$, las funciones de distribución de probabilidad de transición que verifican

$$F(x, t|y, \tau) = g(x, y, t - \tau) = g(x, y, h) = F(x, h|y, 0)$$

se dice que son estacionarias, pudiendo considerarse la notación

$$\forall t > 0, \quad F(x, t|y) = P[X_t \leq x | X_0 = y] \quad \forall x, y \in E$$

Ejemplo 3.1. *Una oficina tiene 5 líneas telefónicas. Supongamos que en un instante determinado puede haber un número cualquiera de líneas ocupadas. Durante cierto tiempo se observan las líneas telefónicas a intervalos regulares de 2 minutos. Si X_n es la variable aleatoria que describe el número de líneas telefónicas que están ocupadas en la n -ésima etapa, ¿el proceso aleatorio que describe este fenómeno aleatorio es homogéneo en el tiempo?*

El proceso $\{X_n; n \geq 1\}$ que describe el fenómeno aleatorio considerado es una cadena en tiempo discreto con espacio de estados $E = \{0, 1, 2, 3, 4, 5\}$. Dicho proceso podrá considerarse homogéneo en el tiempo si la proporción entre llamadas telefónicas entrantes y salientes y el promedio de duración de las mismas no cambia durante el periodo de observación del proceso. Esto significa que dicho periodo no puede incluir intervalos de tiempo densos en los que se esperan más llamadas o intervalos de tiempo tranquilos en los que se esperan pocas llamadas. Por ejemplo, si en una etapa cualquiera, sólo hay 1 línea ocupada, entonces la probabilidad de que en la etapa siguiente hayan k líneas en uso ha de ser constante respecto a n , es decir,

$$P[X_{n+1} = k | X_n = 1] = p_{1,k}(n, n+1) = p_{1,k}$$

Características de un proceso estocástico

De la misma forma que, desde el punto de vista de las aplicaciones, es necesario sustituir la ley de probabilidad de una v.a. por un conjunto reducido de cantidades numéricas: su esperanza matemática, su varianza, ..., que expresan las características de tal v.a., vamos a establecer conceptos paralelos para un p.e. No

obstante y a diferencia de lo que sucede en el caso de una v.a., tales conceptos no se reducirán a constantes, sino a funciones numéricas.

Así, por ejemplo, dado el p.e. $\{X_t; t \in T\}$, introducimos los conceptos siguientes:

- Su función de medias, cuando existe, es la función numérica $\mu_X : T \rightarrow \mathbb{R}$ definida por

$$\mu_X(t) = E[X_t]$$

- Su función de varianzas, cuando existe, es la función real σ_X^2 sobre T dada por

$$\sigma_X^2(t) = Var[X_t] = E[(X_t - \mu_X(t))^2]$$

- Su función de autocovarianzas, es la función $K_X(t_1, t_2) = cov(X_{t_1}, X_{t_2})$ que mide el grado de dependencia entre las v.a. del proceso, para los diferentes valores del parámetro.

Directamente vinculada a la función de autocovarianzas, podemos también considerar la función de autocorrelaciones,

$$\rho_X(t_1, t_2) = \frac{K_X(t_1, t_2)}{\sigma_X(t_1)\sigma_X(t_2)} \quad \forall t_1, t_2 \in T$$

que proporciona el coeficiente de correlación de las v.a. X_{t_1} y X_{t_2} .

Clases de procesos: procesos de Markov

Los procesos estocásticos describen muchos fenómenos naturales en los que, para cualquier conjunto finito de índices $t_1 < t_2 < \dots < t_n$, sus correspondientes variables aleatorias $X(t_1), X(t_2), \dots, X(t_n)$ manifiestan algún tipo de interdependencia.

La clasificación de los procesos estocásticos viene dada por dicha estructura, según la cual existen diversas clases de procesos: los de Markov, los de renovación, los estacionarios, los gaussianos, ... La complejidad del análisis de cada tipo de proceso depende directamente de la complejidad de la estructura de interdependencia entre las variables aleatorias que lo constituyen.

Un p.e. $\{X_t; t \in T\}$ se dice que es un *proceso de Markov* si describe un fenómeno que obedece al *principio* o *condición de Markov*, según el cual “se puede deducir la probabilidad de que un fenómeno aleatorio se encuentre en cierto estado en un instante determinado t , conocido su estado en un instante anterior τ , y ello no depende de la evolución del fenómeno antes del instante τ ” y que, intuitivamente, significa que, *conocido el presente, el futuro no depende del pasado*.

Formalmente, la dependencia markoviana de *primer orden* significa que para cualquier conjunto finito de índices $t_1 < t_2 < \dots < t_{n-1} < t_n$,

$$\mathcal{L}(X_{t_n} | X_{t_{n-1}}, X_{t_{n-2}}, \dots, X_{t_1}) \equiv \mathcal{L}(X_{t_n} | X_{t_{n-1}})$$

o equivalentemente, para $x_1, x_2, \dots, x_n \in \mathbb{R}$,

$$P[X_{t_n} \leq x_n | X_{t_{n-1}} = x_{n-1}, X_{t_{n-2}} = x_{n-2}, \dots, X_{t_1} = x_1] = P[X_{t_n} \leq x_n | X_{t_{n-1}} = x_{n-1}] \quad (3.1)$$

siempre que la probabilidad condicionada del miembro izquierdo esté definida. Si se verificase que

$$\mathcal{L}(X_{t_n}|X_{t_{n-1}}, X_{t_{n-2}}, \dots, X_{t_1}) \equiv \mathcal{L}(X_{t_n}|X_{t_{n-1}}, X_{t_{n-2}})$$

se trataría de un proceso de Markov de segundo orden. Análogamente, podría definirse un proceso de Markov de orden r .

Ejemplo 3.2. *El proceso estocástico del Ejemplo 3.1 será un proceso de Markov, si la probabilidad de cada posible número de líneas ocupadas en cualquier etapa depende únicamente del número de líneas que estaban ocupadas cuando el fenómeno fue observado 2 minutos antes, y no depende de otros valores observados anteriormente. Por ejemplo, si en la n -ésima etapa había 3 líneas ocupadas, entonces la probabilidad del número de líneas ocupadas en la siguiente etapa, debe ser la misma con independencia del número de líneas ocupadas en la $(n-1)$ -ésima etapa, es decir,*

$$\forall k \in E \quad P[X_{n+1} = k|X_n = 3] = P[X_{n+1} = k|X_n = 3, X_{n-1} = i] \quad \forall i \in E$$

Teorema 3.1. *La ley de probabilidad de un proceso de Markov queda descrita por las distribuciones unidimensionales y las distribuciones de probabilidad de transición.*

Demostración:

Probamos este teorema únicamente para cadenas de Markov, pues en el caso de procesos de Markov con espacio de estados E continuo la demostración es equivalente pero utilizando densidades de probabilidad.

En una cadena de Markov, las distribuciones finitodimensionales pueden expresarse por sus funciones de masa de probabilidad. Así, para $x_1, \dots, x_n \in E$ y un conjunto finito de índices $t_1 < t_2 < \dots < t_{n-1} < t_n$, aplicando la definición de probabilidad condicionada, se tiene que

$$\begin{aligned} P[X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_n} = x_n] &= \\ &= P[X_{t_2} = x_2, \dots, X_{t_{n-1}} = x_{n-1}, X_{t_n} = x_n | X_{t_1} = x_1] P[X_{t_1} = x_1] \\ &= P[X_{t_3} = x_3, \dots, X_{t_n} = x_n | X_{t_2} = x_2, X_{t_1} = x_1] P[X_{t_2} = x_2 | X_{t_1} = x_1] P[X_{t_1} = x_1] \\ &\vdots \\ &= P[X_{t_n} = x_n | X_{t_{n-1}} = x_{n-1}, \dots, X_{t_1} = x_1] \dots P[X_{t_2} = x_2 | X_{t_1} = x_1] P[X_{t_1} = x_1] \end{aligned}$$

y en virtud de la condición de Markov 3.1

$$\begin{aligned} P[X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_n} = x_n] &= P[X_{t_n} = x_n | X_{t_{n-1}} = x_{n-1}] \dots P[X_{t_2} = x_2 | X_{t_1} = x_1] P[X_{t_1} = x_1] \\ &= p_{x_{n-1}, x_n}(t_{n-1}, t_n) \dots p_{x_1, x_2}(t_1, t_2) p_{x_1}(t_1) \end{aligned}$$

■

3.2 Cadenas de Markov

Una cadena de Markov de parámetro discreto puede, por tanto, representarse mediante una sucesión de v.a. discretas $\{X_n; n \geq 0\}$ que satisface la condición de

Markov. Supondremos de aquí en adelante que la cadena es de primer orden. Si el espacio de estados lo notamos por $E = \{0, 1, 2, \dots\}$, entonces su ley de probabilidad queda determinada, en virtud del Teorema 3.1, por la familia de distribuciones unidimensionales, que notamos por

$$\{\mathbf{p}(n); n \geq 0\}$$

donde $\mathbf{p}(n)$ es la f.m.p. de la v.a. X_n , que puede representarse mediante el vector fila

$$\mathbf{p}(n) = (p_0(n), p_1(n), p_2(n), \dots, p_i(n), \dots) \quad p_i(n) = P[X_n = i]$$

y la familia de distribuciones de probabilidad de transición, que podemos representar mediante las matrices $\{\mathbf{P}(m, n); n > m \geq 0\}$ donde

$$\mathbf{P}(m, n) = \begin{pmatrix} P_{00}(m, n) & P_{01}(m, n) & \dots & P_{0j}(m, n) & \dots \\ P_{10}(m, n) & P_{11}(m, n) & \dots & P_{1j}(m, n) & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ P_{i0}(m, n) & P_{i1}(m, n) & \dots & P_{ij}(m, n) & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \end{pmatrix}$$

siendo $P_{ij}(m, n) = P[X_n = j | X_m = i]$ las funciones de probabilidad de transición. Estas matrices se dice que son *estocásticas* porque todos sus términos son no negativos y la suma de los elementos de cada fila es 1,

$$P_{ij}(m, n) \geq 0 \quad \forall i, j \in E \quad \text{y} \quad \sum_{j \in E} P_{ij}(m, n) = 1 \quad \forall i \in E$$

Ecuaciones de Chapman-Kolmogorov

Si un proceso es markoviano, entonces sus probabilidades de transición satisfacen las ecuaciones de Chapman-Kolmogorov

$$P_{ij}(m, n) = \sum_{k=0}^{\infty} P_{ik}(m, \tau) P_{kj}(\tau, n) \quad \forall n > \tau > m \geq 0, \quad \forall i, j \in E \quad (3.2)$$

En efecto,

$$P_{ij}(m, n) = P[X_n = j | X_m = i] = \frac{P[X_m = i, X_n = j]}{P[X_m = i]} = \frac{\sum_{k=0}^{\infty} P[X_m = i, X_\tau = k, X_n = j]}{P[X_m = i]}$$

para $n > \tau > m \geq 0$ y en virtud del Teorema de la Probabilidad Total, con lo cual

$$\begin{aligned}
 P_{ij}(m, n) &= \sum_{k=0}^{\infty} \frac{P[X_m = i, X_\tau = k, X_n = j]}{P[X_m = i]} \\
 &= \sum_{k=0}^{\infty} \frac{P[X_m = i]P[X_\tau = k, X_n = j|X_m = i]}{P[X_m = i]} \\
 &= \sum_{k=0}^{\infty} P[X_\tau = k, X_n = j|X_m = i] \\
 &= \sum_{k=0}^{\infty} P[X_\tau = k|X_m = i]P[X_n = j|X_\tau = k, X_m = i] \\
 &= \sum_{k=0}^{\infty} P[X_\tau = k|X_m = i]P[X_n = j|X_\tau = k] \\
 &= \sum_{k=0}^{\infty} P_{ik}(m, \tau)P_{kj}(\tau, n)
 \end{aligned}$$

Análogamente,

$$p_j(n) = P[X_n = j] = \sum_{k=0}^{\infty} p_k(0)P_{kj}(0, n)$$

Matricialmente, dichas ecuaciones pueden expresarse, respectivamente, por

$$\mathbf{P}(m, n) = \mathbf{P}(m, \tau)\mathbf{P}(\tau, n) \quad n > \tau > m \geq 0 \quad (3.3)$$

y

$$\mathbf{p}(n) = \mathbf{p}(0)\mathbf{P}(0, n) \quad n > 0 \quad (3.4)$$

En consecuencia, la ley de probabilidad de la cadena de Markov $\{X_n; n \geq 0\}$ queda determinada, en virtud del Teorema 3.1, por:

- la distribución inicial: $\mathbf{p}(0)$ y
- la familia de distribuciones de transición: $\{\mathbf{P}(m, n); n > m \geq 0\}$

Matrices de transición

Aplicando sucesivamente las ecuaciones de Chapman-Kolmogorov a las probabilidades de transición en (3.3), se deducen las siguientes relaciones recurrentes:

$$\mathbf{P}(m, n) = \mathbf{P}(m, n-1)\mathbf{P}(n-1, n) = \mathbf{P}(m, n-2)\mathbf{P}(n-2, n-1)\mathbf{P}(n-1, n)$$

esto es,

$$\mathbf{P}(m, n) = \mathbf{P}(m, m+1)\mathbf{P}(m+1, m+2)\dots\mathbf{P}(n-1, n) \quad m < n$$

con lo cual, la familia de matrices de transición $\{\mathbf{P}(m, n); n > m \geq 0\}$ queda determinada, a su vez, por la familia de matrices de transición en 1 etapa:

$$\{\mathbf{P}(n, n+1); n \geq 0\}$$

y puede enunciarse así el siguiente

Teorema 3.2. Si $\{X_n; n \geq 0\}$ es una cadena de Markov, su ley de probabilidad queda completamente determinada por

- La distribución inicial $\mathbf{p}(0)$, y
- La familia de distribuciones de probabilidad de transición en 1 etapa:

$$\{\mathbf{P}(n, n+1); n \geq 0\}$$

Cadenas de Markov homogéneas

Si la cadena de Markov es homogénea en el tiempo, esto es, sus funciones de probabilidad de transición, $P_{ij}(m, n)$, son estacionarias (dependen del lapso $n - m$),

$$P_{ij}(m, n) \equiv P_{ij}(n - m) \quad \forall n > m (\Leftrightarrow n - m > 0)$$

para cada $i, j \in E$. En consecuencia,

$$\mathbf{P}(m, n) \equiv \mathbf{P}(n - m) \quad n > m \geq 0$$

Dichas probabilidades las notamos por

$$P_{ij}(n) \quad \forall i, j \in E \quad \text{y} \quad n > 0$$

y se conocen como *probabilidades de transición en n etapas*,

$$P_{ij}(n) = P[X_{m+n} = j | X_m = i] \quad \forall i, j \in E \quad \text{con} \quad n, m \geq 0$$

y $\mathbf{P}(n)$ es la *matriz de transición en n etapas*. En particular, las probabilidades de transición en 1 etapa las notaremos simplemente por

$$P_{ij}(1) = P_{ij} \quad \forall i, j \in E \quad \Rightarrow \quad \mathbf{P}(1) \equiv \mathbf{P}$$

De (3.3) y (3.4) tenemos,

$$\mathbf{P}(m, n) = \mathbf{P}^{n-m} \quad \Rightarrow \quad \mathbf{P}(n) = \mathbf{P}^n$$

con lo cual

$$\mathbf{p}(n) = \mathbf{p}(0)\mathbf{P}^n$$

y puede enunciarse el siguiente

Teorema 3.3. Si $\{X_n; n \geq 0\}$ es una cadena de Markov homogénea en el tiempo, su ley de probabilidad queda completamente determinada por

- La distribución inicial: $\mathbf{p}(0)$, y
- La matriz de transición en 1 etapa: \mathbf{P} , conocida simplemente como *matriz de transición de la cadena*.

Si la cadena es finita, esto es, tiene un número finito k de estados, entonces su matriz de transición \mathbf{P} es cuadrada y de orden k .

Ejemplo 3.3. (Prediciendo el tiempo atmosférico) Supongamos que la probabilidad de lluvia mañana depende exclusivamente de si está lloviendo o no hoy, y no depende de las condiciones climatológicas pasadas. Supongamos también que si llueve hoy, entonces lloverá mañana con probabilidad α ; y si no está lloviendo hoy, entonces lloverá mañana con probabilidad β .

Si decimos que el proceso está en el estado 0 cuando llueve y en el estado 1 cuando no llueve, entonces estamos ante una cadena de Markov con dos estados cuyas probabilidades de transición están dadas por

$$P = \begin{pmatrix} \alpha & 1 - \alpha \\ \beta & 1 - \beta \end{pmatrix}$$

Supongamos ahora que $\alpha = 0,7$ y $\beta = 0,4$. Vamos a calcular la probabilidad de que llueva durante los próximos cuatro días, teniendo en cuenta que hoy está lloviendo.

La matriz de transición en 1 etapa en este caso es

$$P = \begin{pmatrix} 0,7 & 0,3 \\ 0,4 & 0,6 \end{pmatrix}$$

entonces,

$$P(2) = P^2 = \begin{pmatrix} 0,7 & 0,3 \\ 0,4 & 0,6 \end{pmatrix} \cdot \begin{pmatrix} 0,7 & 0,3 \\ 0,4 & 0,6 \end{pmatrix} = \begin{pmatrix} 0,61 & 0,39 \\ 0,52 & 0,48 \end{pmatrix},$$

$$P(4) = (P(2))^2 = \begin{pmatrix} 0,61 & 0,39 \\ 0,52 & 0,48 \end{pmatrix} \cdot \begin{pmatrix} 0,61 & 0,39 \\ 0,52 & 0,48 \end{pmatrix} = \begin{pmatrix} 0,5749 & 0,4251 \\ 0,5668 & 0,4332 \end{pmatrix}$$

y la probabilidad buscada es $P_{00} = 0,5749$.

Clases de estados

Definición 3.1. Se dice que el estado j es accesible desde el estado i si, empezando en i , es posible que el proceso llegue al estado j en un número finito de pasos. Si un estado j es accesible desde otro estado i , y a su vez i es accesible desde j , se dice que los estados i y j se comunican. En términos probabilísticos

$$\begin{aligned} i \rightarrow j \text{ (} j \text{ accesible desde } i \text{)} & \quad \text{si para algún } n \geq 0, \quad P_{ij}(n) > 0 \\ i \leftrightarrow j \text{ (} i \text{ y } j \text{ se comunican)} & \quad \text{si para algún } n \geq 0, \quad P_{ij}(n) > 0 \\ & \quad \text{y para algún } m \geq 0, \quad P_{ji}(m) > 0 \end{aligned}$$

Al contrario,

$$\begin{aligned} i \nrightarrow j \text{ (} j \text{ no accesible desde } i \text{)} & \quad \text{si para todo } n \geq 0, \quad P_{ij}(n) = 0 \\ i \nleftrightarrow j \text{ (} i \text{ y } j \text{ no se comunican)} & \quad \text{si para todo } n \geq 0, \quad P_{ij}(n) = 0 \\ & \quad \text{o para todo } n \geq 0, \quad P_{ji}(n) = 0 \\ & \quad \text{o para todo } n \geq 0, \quad P_{ij}(n) = 0 \text{ y } P_{ji}(n) = 0 \end{aligned}$$

Nótese que cada estado se comunica consigo mismo, ya que, por definición,

$$P_{ii}(0) = P[X_0 = i | X_0 = i] = 1 \quad \forall i \in E$$

La relación de comunicación satisface las siguientes tres propiedades:

- (i) Reflexiva: $i \leftrightarrow i$
- (ii) Simétrica: si $i \leftrightarrow j$, entonces $j \leftrightarrow i$
- (iii) Transitiva: si $i \leftrightarrow j$ y $j \leftrightarrow k$, entonces $i \leftrightarrow k$

Las propiedades (i) y (ii) se siguen inmediatamente de la definición de comunicación. Para probar (iii), supongamos que i se comunica con j y que j se comunica con k . Luego, existen enteros n y m tales que $P_{ij}(n) > 0, P_{jk}(m) > 0$. Entonces, por las ecuaciones de Chapman-Kolmogorov, tenemos que

$$P_{ik}(n+m) = \sum_{r=0}^{\infty} P_{ir}(n)P_{rk}(m) \geq P_{ij}(n)P_{jk}(m) > 0$$

Por consiguiente, el estado k es accesible desde el estado i . De igual modo se prueba que el estado i es accesible desde el estado k . Luego, los estados i y k se comunican.

Como una relación que satisfaga las propiedades reflexiva, simétrica y transitiva es una *relación de equivalencia*, podemos decir que la relación de comunicación lo es. El conjunto de todos los estados de una cadena de Markov que se comunican puede entonces agruparse en una *clase de equivalencia*. Una cadena de Markov puede tener más de una clase de equivalencia. Si hay más de una, entonces no es posible que estados que se comunican estén en diferentes clases de equivalencia, pero sí es posible que haya estados en una clase que sean accesibles desde otros de otra clase.

Definición 3.2. Si todos los estados de una cadena de Markov pertenecen a una misma clase de equivalencia, se dice que esta cadena es irreducible.

Ejemplo 3.4. Consideremos la cadena de Markov que consta de los tres estados 0, 1, 2 y que tiene la matriz de probabilidades de transición

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

Es fácil comprobar que esta cadena de Markov es irreducible. Por ejemplo, es posible ir del estado 0 al estado 2, ya que

$$0 \rightarrow 1 \rightarrow 2$$

Esto es, un camino para llegar desde el estado 0 al estado 2 es ir del estado 0 al estado 1 (con probabilidad $\frac{1}{2}$) y luego ir del estado 1 al estado 2 (con probabilidad $\frac{1}{4}$).

Estados recurrentes y transitorios

Veamos ahora otra clasificación de los estados, muy ligada a la clasificación dada por la relación de comunicación, que se refiere a la naturaleza “interna” de cada estado y se basa en la noción de *probabilidad del tiempo de primer paso*. Así, por ejemplo, la probabilidad de que, partiendo de i , la cadena pase por primera vez por el estado k en la n -ésima etapa, la notamos por $f_{ik}(n)$ y viene dada por

$$f_{ik}(n) = P[X_n = k, X_m \neq k, m = 1, 2, \dots, n-1 | X_0 = i]$$

con la convención de que $f_{ik}(0) = 0$. En cambio, es evidente que $f_{ik}(1) = P_{ik}$.

La probabilidad de pasar alguna vez por k , partiendo de i viene dada, entonces, por

$$f_{ik} = \sum_{n=1}^{\infty} f_{ik}(n)$$

Cuando $k = i$, entonces f_{ii} es la *probabilidad de recurrencia de i* .

Definición 3.3. (*Estados recurrente y transitorio*). Un estado i es recurrente cuando $f_{ii} = 1$. En caso contrario ($f_{ii} < 1$), es un estado transitorio.

El resultado siguiente entre las probabilidades de transición y las del tiempo de primer paso que, en cierto sentido, es intuitivo, es fundamental para caracterizar la recurrencia.

Teorema 3.4.

$$P_{ik}(n) = \sum_{m=0}^n f_{ik}(m) P_{kk}(n-m) \quad n \geq 1 \quad (3.5)$$

esto es, la probabilidad de que, partiendo del estado i , la cadena pase por el estado k en la n -ésima etapa es equivalente a la probabilidad de que en alguna etapa intermedia m , pase por primera vez por dicho estado k y vuelva a pasar por él al cabo de $(n-m)$ etapas.

$$P_{ii}(n) = \sum_{m=0}^n f_{ii}(m) P_{kk}(n-m) \quad n \geq 1 \quad (3.6)$$

Demostración:

En efecto, sean $i, k \in E$ y consideremos todas las posibles trayectorias de la cadena en las que, partiendo de i , la cadena pasa por k en la n -ésima etapa ($[X_n = k]$), aunque el primer paso por k ha ocurrido en alguna etapa intermedia m . Formalmente, si

$$A_n = [X_n = k] \quad \text{y} \quad B_m = [X_m = k, X_v \neq k, v = 1, \dots, m-1]$$

entonces $B_m = A_m - (A_{m-1} \cup A_{m-2} \cup \dots \cup A_1)$ para $m = 1, \dots, n$, los sucesos $\{B_m; m = 1, \dots, n\}$ son mutuamente excluyentes y,

$$\sum_{m=1}^n B_m = \bigcup_{m=1}^n A_m$$

Por tanto,

$$A_n = A_n \cap (\cup_{m=1}^n A_m) = A_n \cap \left(\sum_{m=1}^n B_m \right) = \sum_{m=1}^n (A_n \cap B_m)$$

y

$$P[A_n | X_0 = i] = \sum_{m=1}^n P[A_n \cap B_m | X_0 = i] \quad (3.7)$$

donde

$$\begin{aligned} P[A_n \cap B_m | X_0 = i] &= P[A_n | B_m, X_0 = i] P[B_m | X_0 = i] \\ &= P[X_n = k | X_m = k] P[B_m | X_0 = i] \\ &= P_{kk}(n-m) f_{ik}(m) \end{aligned}$$

con lo cual

$$P_{ik}(n) \sum_{m=1}^n f_{ik}(m) P_{kk}(n-m) = \sum_{m=0}^n f_{ik}(m) P_{kk}(n-m)$$

dado que $f_{ik}(0) = 0$. Cuando $k = i$ y por un razonamiento análogo al anterior, obtenemos (3.6). ■

Corolario 3.4.1. (Criterio de recurrencia).

$$f_{ii} \begin{cases} < 1 & \text{si } \sum_{n=0}^{\infty} P_{ii}(n) < \infty \\ = 1 & \text{si } \sum_{n=0}^{\infty} P_{ii}(n) = \infty \end{cases}$$

Corolario 3.4.2. (Recurrencia como propiedad de clase). Si dos estados se comunican ($i \leftrightarrow k$), entonces o son ambos recurrentes o son ambos transitorios,

$$f_{ii} = 1 \quad \text{y} \quad i \leftrightarrow k \quad \Rightarrow \quad f_{kk} = 1$$

En consecuencia, los estados de una misma clase comunicante o son todos recurrentes o todos transitorios.

Estados recurrentes positivos y recurrentes nulos

Si i es recurrente y T_i es la v.a. que describe el número de etapas necesarias para que la cadena retorne, vuelva a pasar por dicho estado, entonces su f.m.p. es $\{f_{ii}(n); n \geq 0\}$ y notaremos por μ_i al tiempo medio de recurrencia,

$$\mu_i = \sum_{n=1}^{\infty} n f_{ii}(n)$$

Definición 3.4. Un estado recurrente se dice que es recurrente positivo, cuando $\mu_i < \infty$; y recurrente nulo, cuando $\mu_i = \infty$.

Pueden probarse, además, los siguientes resultados:

- La recurrencia nula sólo es posible en cadenas de Markov infinitas.
- La recurrencia positiva es también una propiedad de clase, es decir, si dos estados recurrentes i, k de una cadena de Markov se comunican, o son ambos recurrentes positivos o son ambos recurrentes nulos.

Estados periódicos

Definición 3.5. Un estado i es periódico y su periodo es $n(> 1)$, cuando n es el mayor número natural tal que $P_{ii}(n) > 0$, es decir, la cadena únicamente puede pasar por él en las etapas $n, 2n, 3n, \dots$,

$$P_{ii}(m) \begin{cases} > 0 & \text{si } m = n \\ = 0 & \text{si } m \neq n \end{cases}$$

Un estado aperiódico es un estado cuyo periodo es 1.

Al igual que en otros casos, la periodicidad es una propiedad de clase. Puesto que el periodo de una clase comunicante es el periodo común a todos los estados de la misma, si su periodo es 1 la clase es aperiódica.

Definición 3.6. Un estado ergódico es un estado recurrente positivo y aperiódico.

Teorema 3.5. Si el estado i tiene periodo $d_i (\geq 1)$, entonces existe un entero N tal que para todo $n \geq N$, $P_{ii}(nd_i) > 0$

Demostración:

La demostración de este teorema depende de un resultado de teoría de números, el cual afirma que si n_1, n_2, \dots, n_k son enteros positivos con máximo común divisor d , entonces existe un entero positivo N tal que para todo $n \geq N$ podemos encontrar c_i (enteros no negativos) satisfaciendo la relación

$$nd = \sum_{i=1}^k c_i n_i \tag{3.8}$$

Para el estado i , sean n_1, n_2, \dots, n_k los enteros para los que $P_{ii}(n_r) > 0$ ($r = 1, 2, \dots, k$) y sea d_i su máximo común divisor. Sin pérdida de generalidad, asumimos c_i ($i = 1, 2, \dots, k$) como enteros positivos. Por (3.8) tenemos que

$$\begin{aligned} P_{ii}(nd_i) &= P_{ii} \left(\sum_1^k c_r n_r \right) \\ &\geq [P_{ii}(n_1)]^{c_1} [P_{ii}(n_2)]^{c_2} \dots [P_{ii}(n_k)]^{c_k} > 0 \end{aligned}$$

■

Corolario 3.5.1. Si $P_{ij}(m) > 0$, entonces existe un entero positivo N tal que, para todo $n \geq N$, $P_{ij}(m + nd_i) > 0$

Estados absorbentes

Existen también estados que actúan como agujeros negros: si se llega a ellos ya no se sale nunca jamás. Se denominan *estados absorbentes*.

Definición 3.7. Un estado i se dice que es un estado absorbente si y solo si $P_{ii} = 1$.

Si i es un estado absorbente, $f_{ii}(1) = P_{ii} = 1$, de modo que $f_{ii} = 1$ y $\mu_i = 1$, por lo que i es recurrente positivo. Por tanto, puede afirmarse que *todo estado absorbente es recurrente positivo*.

Teorema 3.6. *Si una cadena de Markov homogénea tiene un número finito de estados, de los cuales, al menos uno es absorbente y accesible desde algún estado no absorbente, entonces la absorción es segura.*

Demostración:

Sea E el espacio de estados y $A = \{a_k\}$ el subespacio de estados absorbentes, donde $k \geq 1$. Supuesto que al menos un estado absorbente es accesible desde el subespacio $E - A = B = \{b_j\}$, probaremos que $\lim_n P[X_n \in A] = 1$.

En efecto, si $X_0 \in A$, la afirmación es inmediata porque la absorción ya ha ocurrido. Supongamos, por tanto, que $X_0 \in B$. Puesto que, por hipótesis, existe al menos un estado en A que es accesible desde algún estado de B , esto es,

$$\exists a_k \in A \text{ y } \exists b_j \in B / b_j \longrightarrow a_k \quad \Rightarrow \quad \exists n_{j,k} / p_{j,k}(n_{j,k}) > 0$$

Sea entonces $n_{j,k}^0 = \min_j \{n_{j,k}\}$ la primera vez o etapa en la que la absorción tiene lugar en el estado a_k ($B \longrightarrow a_k$). Si existiesen más estados absorbentes accesibles desde B , sea

$$n^* = \max_j \min_k \{n_{j,k}\}$$

esto es, la última etapa en la que algún estado de B alcanza algún estado absorbente, se tiene entonces

$$P[X_{n^*} \in A] = p > 0 \quad \Leftrightarrow \quad P[X_{n^*} \in B] = (1 - p)$$

con $0 < (1 - p) < 1$. Puesto que, además, la cadena es markoviana y homogénea,

$$\begin{aligned} P[X_{2n^*} \in B] &= P[X_{n^*} \in B] \times P[X_{2n^*} \in B | X_{n^*} \in B] \\ &= P[X_{n^*} \in B] \times P[X_{n^*} \in B] \\ &= (1 - p)^2 \end{aligned}$$

y procediendo sucesivamente,

$$P[X_{kn^*} \in B] = (1 - p)^k \quad k \geq 1$$

Dado que, $\lim_k P[X_{kn^*} \in B] = \lim_k (1 - p)^k = 0$, obtenemos, en definitiva, que

$$\lim_n P[X_n \in B] = 0 \quad \Leftrightarrow \quad \lim_n P[X_n \in A] = 1$$

■

Ejemplo 3.5. *Consideremos la cadena de Markov que consta de los cuatro estados 0, 1, 2, 3 y que tiene la matriz de probabilidades de transición*

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Las clases de esta cadena de Markov son $\{0,1\}$, $\{2\}$ y $\{3\}$. Nótese que mientras que el estado 0 (o 1) es accesible desde el estado 2, el contrario no es cierto. Como el estado 3 es un estado absorbente, esto es, $P_{33} = 1$, ningún otro estado es accesible desde él.

Cadenas de Markov irreducibles y ergódicas: comportamiento límite

Si una cadena de Markov homogénea en el tiempo es finita e irreducible, entonces es recurrente positiva y suponiendo, además, que es aperiódica, es ergódica.

En el Ejemplo 3.3 calculamos $\mathbf{P}(4)$ para una cadena de Markov de 2 estados y obtuvimos como resultado

$$\mathbf{P}(4) = \begin{pmatrix} 0,5749 & 0,4251 \\ 0,5668 & 0,4332 \end{pmatrix}$$

de donde se sigue que $\mathbf{P}(8) = \mathbf{P}(4) \cdot \mathbf{P}(4) = \mathbf{P}^4 \cdot \mathbf{P}^4$ viene dada por

$$\mathbf{P}(8) = \begin{pmatrix} 0,572 & 0,428 \\ 0,570 & 0,430 \end{pmatrix}$$

Se percibe que la matriz $\mathbf{P}(8)$ es casi idéntica a la matriz $\mathbf{P}(4)$ y que cada una de las filas de $\mathbf{P}(8)$ tiene sus entradas casi idénticas. De hecho, parece que $P_{ij}(n)$ está convergiendo a algún valor (cuando $n \rightarrow \infty$) que es el mismo para todos los i . En otras palabras, parece existir una *probabilidad límite* de que el proceso esté en el estado j después de un gran número de transiciones, y este valor es independiente del estado inicial.

Las *matrices regulares* juegan un papel fundamental en el estudio del “comportamiento a largo plazo o estacionario” de las cadenas irreducibles y ergódicas.

Definición 3.8. Una matriz estocástica \mathbf{A} es regular cuando alguna potencia \mathbf{A}^n tiene todos sus términos positivos, esto es, $\mathbf{A}^n > 0$.

Caracterización 3.8.1. Una matriz estocástica \mathbf{A} es regular si y sólo si $\lambda = 1$ es un autovalor simple.

Como consecuencia, los demás autovalores son de módulo inferior a la unidad y el autovalor máximo no unidad, proporciona la tasa de aproximación de \mathbf{A}^n a su límite, $\lim_{n \rightarrow \infty} \mathbf{A}^n$.

Lema 3.1. En toda cadena de Markov finita, irreducible y aperiódica, con matriz de transición \mathbf{P} existe N , tal que $\forall n \geq N$ la matriz \mathbf{P}^n no contiene ceros, esto es,

$$\mathbf{P}^n > 0 \quad n \geq N$$

lo que significa que, después de N etapas, cada estado tiene una probabilidad no nula de ser ocupado, cualquiera que sea el estado inicial.

Demostración:

Si E es el espacio de estados de la cadena y, puesto que todos los estados se comunican entre sí, para cada par de estados $i, j \in E$ existen enteros no negativos $m(i, j)$ tales que

$$P_{ij}^{(m(i, j))} > 0$$

Por el Corolario 3.5.1 sabemos que para cada (i, j) existe un entero positivo $N(i, j)$ tal que para $n \geq N(i, j)$

$$P_{ij}^{(m(i, j) + n)} > 0$$

Si tomamos

$$N = \max_{i, j \in E} \{m(i, j) + N(i, j)\},$$

entonces es claro que para todo $n \geq N$, $P_{ij}^{(n)} > 0$

■

Se tiene, por tanto, que la matriz de transición \mathbf{P} es regular, y esta es la razón por la cual estas cadenas de Markov se llaman también cadenas regulares y, como consecuencia directa de las propiedades de las matrices regulares tenemos el siguiente

Teorema 3.7. Si $\{X_n; n \geq 0\}$ es una cadena de Markov homogénea, finita, irreducible y ergódica con matriz de transición \mathbf{P} , entonces

a)

$$\exists \lim_{n \rightarrow \infty} \mathbf{P}^n = \mathbf{\Pi}$$

y las filas de $\mathbf{\Pi}$ son una distribución de probabilidad $\pi > 0$.

b) La distribución de probabilidad π es la única distribución estacionaria de la cadena.

c) Cualquiera que sea la distribución inicial de la cadena, $\mathbf{p}(0)$,

$$\exists \lim_{n \rightarrow \infty} \mathbf{p}(n) = \lim_{n \rightarrow \infty} \mathbf{p}(0)\mathbf{P}^n = \pi$$

y dado que $\pi > 0$, existe y es única la distribución de equilibrio.

Así pues, en toda cadena de Markov homogénea, finita, irreducible y ergódica con matriz de transición \mathbf{P} , la distribución de equilibrio existe, es única y coincide con la distribución estacionaria, pudiendo determinarse como solución única del sistema

$$\pi = \pi\mathbf{P} \tag{3.9}$$

sujeto a las condiciones

$$\pi_j > 0 \quad \forall j \quad \text{y} \quad \sum_{j \in E} \pi_j = 1 \tag{3.10}$$

Puede demostrarse, además, que la distribución de equilibrio:

(i) proporciona los tiempos medios de recurrencia de cada uno de los estados de la cadena, pues

$$\mu_j = \frac{1}{\pi_j} \quad \forall j \in E \quad (3.11)$$

(ii) proporciona también la proporción media del tiempo de ocupación a largo plazo de cada estado de la cadena.

Ejemplo 3.6. Consideremos el Ejemplo 3.3, en el que asumimos que si llueve hoy, entonces lloverá mañana con probabilidad α ; y si no llueve hoy, entonces lloverá mañana con probabilidad β . Si decimos que el estado es 0 cuando llueve y 1 cuando no llueve, entonces por (3.9) y (3.10) las probabilidades límite π_0 y π_1 satisfacen

$$\pi_0 = \alpha\pi_0 + \beta\pi_1,$$

$$\pi_1 = (1 - \alpha)\pi_0 + (1 - \beta)\pi_1,$$

$$\pi_0 + \pi_1 = 1$$

de donde se obtiene que

$$\pi_0 = \frac{\beta}{1 + \beta - \alpha}, \quad \pi_1 = \frac{1 - \alpha}{1 + \beta - \alpha}$$

Por ejemplo, si $\alpha = 0,7$ y $\beta = 0,4$, entonces $\pi_0 = 0,571$ y $\pi_1 = 0,429$, con lo que puede decirse que a la larga es más probable que llueva.

Aplicación: el modelo de Wright-Fisher

La genética es la ciencia de la herencia, de los fenómenos hereditarios, que estudia la transmisión de los caracteres anatómicos, citológicos y funcionales de padres a hijos. Sus primeras leyes fueron establecidas por Mendel en 1865.

El modelo de Wright-Fisher, versión del modelo de Hardy-Weinberg para poblaciones pequeñas, es uno de los modelos más simples utilizados en genética de poblaciones. Al igual que el principio de Hardy-Weinberg, nos proporciona una idea de cómo cambian las frecuencias génicas en las condiciones siguientes:

- El apareamiento es aleatorio
- No hay mutación.
- No hay migración.
- La selección natural no está actuando sobre las características en cuestión.

Se trata de un modelo probabilista de transmisión de dos alelos A y a de un gen de una población de N organismos diploides. En cada locus hay, por tanto, un total de $2N$ alelos. Las combinaciones de los alelos A y a dan lugar a los genotipos AA , Aa y aa .

Supongamos, por ejemplo, que hay N_1 organismos con el genotipo AA ; N_2 organismos con el genotipo Aa y N_3 con el genotipo aa , entonces las frecuencias de tales genotipos son

$$f = \frac{N_1}{N} \quad g = \frac{N_2}{N} \quad h = \frac{N_3}{N} \quad (4.1)$$

siendo $f + g + h = 1$. El número de alelos A es $2N_1 + N_2$ y el de alelos a es $N_2 + 2N_3$ y sus frecuencias son, respectivamente,

$$fr(A) = \frac{2N_1 + N_2}{2N} = f + \frac{1}{2}g = x \quad fr(a) = \frac{N_2 + 2N_3}{2N} = \frac{1}{2}g + h = y \quad (4.2)$$

con $x + y = 1$. Suponiendo que el apareamiento es aleatorio y, además, las generaciones no se superponen, nos interesa analizar la evolución, a lo largo de las distintas generaciones, de las frecuencias de los alelos pero, dado que $x + y = 1$, será suficiente estudiar la de uno de ellos, por ejemplo, la de los alelos del tipo A .

Ejemplo 4.1. Para ilustrar la determinación de las frecuencias genotípicas de una población para un determinado locus autosómico, consideremos el sistema $M - N$ de grupos sanguíneos. Existen tres grupos sanguíneos, M , N y MN , que están determinados por dos alelos, L^M y L^N , de un mismo locus, de acuerdo con la siguiente relación

Grupo sanguíneo	Genotipo
M	$L^M L^M$
MN	$L^M L^N$
N	$L^N L^N$

Se analizó una muestra de 730 individuos de la población aborigen australiana y se encontró que 22 individuos eran del grupo M, 216 del grupo MN y 492 del grupo N. Luego, en virtud de (4.1), las frecuencias de los grupos sanguíneos M (genotipo $L^M L^M$), MN (genotipo $L^M L^N$) y N (genotipo $L^N L^N$) son, respectivamente,

$$f = \frac{22}{730} = 0,030 \quad g = \frac{216}{730} = 0,296 \quad h = \frac{492}{730} = 0,674$$

Como podemos comprobar, $f + g + h = 0,030 + 0,296 + 0,674 = 1$.

A partir de las frecuencias genotípicas calculadas podemos obtener, en virtud de (4.2), las frecuencias de los alelos L^M y L^N , siendo estas

$$fr(L^M) = f + \frac{1}{2}g = 0,030 + \frac{1}{2}0,296 = 0,178 = x$$

$$fr(L^N) = \frac{1}{2}g + h = \frac{1}{2}0,296 + 0,674 = 0,822 = y$$

Como podemos comprobar, $x + y = 0,178 + 0,822 = 1$.

Si asumimos que los 730 individuos constituyen una muestra aleatoria de aborígenes australianos, podemos considerar que tanto las frecuencias genotípicas observadas como las frecuencias alélicas se pueden interpretar como probabilidades y características de la población australiana.

4.1 Principio de Hardy-Weinberg

En poblaciones **infinitas** de individuos que se aparean al azar las frecuencias de los alelos se mantienen constantes, como se demuestra en el llamado *principio de Hardy-Weinberg*. En tales poblaciones, tanto las frecuencias de los genotipos como las de los alelos pueden aproximarse o interpretarse como probabilidades. Si observamos la evolución de dicha población y consideramos la n -ésima generación, sean f_n , g_n y h_n las frecuencias de los genotipos AA, Aa y aa, respectivamente, y sean x_n y y_n las de los alelos A y a.

Teorema 4.1. En una población diploide infinita que se aparee aleatoriamente,

(a) las frecuencias de los dos alelos situados en el mismo locus de un gen no cambian de generación en generación, esto es, $x_n = x_0$, $y_n = y_0$ para $n \geq 1$

Además, cualesquiera que sean las frecuencias de los genotipos,

(b) las frecuencias de las sucesivas generaciones permanecen fijas, esto es, $f_n = f_1$, $g_n = g_1$ y $h_n = h_1$ para $n \geq 2$ y

(c) están determinadas por las frecuencias iniciales de los alelos, pues f_1 , g_1 y h_1 dependen de x_0 y y_0 , dado que

$$f_1 = x_0^2 \quad g_1 = 2x_0y_0 \quad h_1 = y_0^2 \quad (4.3)$$

	Apareamiento		<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
	Masculino	Femenino	$P(M_i)$	$P(AA M_i)$	$P(Aa M_i)$	$P(aa M_i)$
M_1	AA	AA	f_0^2	1	0	0
M_2	AA	Aa	f_0g_0	$\frac{1}{2}$	$\frac{1}{2}$	0
M_3	AA	aa	f_0h_0	0	1	0
M_4	Aa	AA	f_0g_0	$\frac{1}{2}$	$\frac{1}{2}$	0
M_5	Aa	Aa	g_0^2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
M_6	Aa	aa	g_0h_0	0	$\frac{1}{2}$	$\frac{1}{2}$
M_7	aa	AA	f_0h_0	0	1	0
M_8	aa	Aa	g_0h_0	0	$\frac{1}{2}$	$\frac{1}{2}$
M_9	aa	aa	h_0^2	0	0	1

Cuadro 4.1: Posibles apareamientos, sus probabilidades y las de los genotipos de la descendencia.

Demostración:

En la población considerada, dado que tanto los individuos pertenecientes al sexo masculino como al femenino pueden tener cualquiera de los genotipos AA , Aa o aa , hay un total de $3^2 = 9$ apareamientos posibles, denotados por M_i , $i = 1, 2, \dots, 9$, los cuales se muestran en el Cuadro 4.1.

Las probabilidades de los diferentes apareamientos, recogidas en la columna *I* del Cuadro 4.1, se han calculado como sigue:

$$\begin{aligned}
 P(M_1) &= P(AA \cap AA) = P(AA)P(AA) = f_0f_0 = f_0^2 \\
 P(M_2) &= P(AA \cap Aa) = P(AA)P(Aa) = f_0g_0 \\
 &\vdots \\
 P(M_9) &= P(aa \cap aa) = P(aa)P(aa) = h_0h_0 = h_0^2
 \end{aligned}$$

Estos nueve apareamientos constituyen un sistema completo o partición, pues son sucesos mutuamente incompatibles cuyas probabilidades suman la unidad.

A su vez, las probabilidades de los genotipos de los descendientes, recogidas en las columnas *II*, *III* y *IV*, se calculan por

$$\begin{aligned}
 M_1(AA, AA) &\rightarrow AA \quad AA \quad AA \quad AA \quad \Rightarrow \quad P(AA|M_1) = 1 \\
 M_2(AA, Aa) &\rightarrow AA \quad Aa \quad AA \quad Aa \quad \Rightarrow \quad P(AA|M_2) = \frac{1}{2} \quad \text{y} \quad P(Aa|M_2) = \frac{1}{2} \\
 &\vdots \\
 M_8(aa, Aa) &\rightarrow aA \quad aa \quad aA \quad aa \quad \Rightarrow \quad P(Aa|M_8) = \frac{1}{2} \quad \text{y} \quad P(aa|M_8) = \frac{1}{2} \\
 M_9(aa, aa) &\rightarrow aa \quad aa \quad aa \quad aa \quad \Rightarrow \quad P(aa|M_9) = 1
 \end{aligned}$$

A partir de aquí, y en virtud del Teorema de la Probabilidad Total, la probabilidad de que un descendiente, esto es, un miembro de la primera generación, tenga el

genotipo AA es

$$\begin{aligned} f_1 = P(\text{descendiente sea } AA) &= \sum_{i=1}^9 P(AA|M_i)P(M_i) \\ &= f_0^2 + f_0g_0 + \frac{g_0^2}{4} \\ &= \left(f_0 + \frac{g_0}{2}\right)^2 \end{aligned}$$

y aplicando (4.2) obtenemos $f_1 = x_0^2$. Similarmente,

$$\begin{aligned} g_1 = P(\text{descendiente sea } Aa) &= \sum_{i=1}^9 P(Aa|M_i)P(M_i) \\ &= f_0g_0 + f_0h_0 + g_0h_0 + \frac{g_0^2}{2} \\ &= 2\left(f_0 + \frac{g_0}{2}\right)\left(\frac{g_0}{2} + h_0\right) \\ &= 2x_0y_0 \end{aligned}$$

$$\begin{aligned} h_1 = P(\text{descendiente sea } aa) &= \sum_{i=1}^9 P(aa|M_i)P(M_i) \\ &= \frac{g_0^2}{4} + g_0h_0 + h_0^2 \\ &= \left(\frac{g_0}{2} + h_0\right)^2 \\ &= y_0^2 \end{aligned}$$

obteniéndose (4.3).

Se demuestra, por tanto, que las frecuencias (o probabilidades) de los genotipos en cualquier generación están completamente determinadas por las frecuencias de los alelos en la generación anterior, cualesquiera que sean las frecuencias de los genotipos de ella (c).

En segundo lugar, si calculamos las frecuencias de los alelos A y a de la generación de los descendientes, obtenemos:

$$x_1 = f_1 + \frac{1}{2}g_1 = x_0^2 + x_0y_0 = x_0(x_0 + y_0) = x_0$$

con lo cual

$$y_1 = 1 - x_1 = 1 - x_0 = y_0$$

esto es, las frecuencias de los alelos permanecen constantes de generación en generación (a) y, finalmente, teniendo en cuenta (4.3), las frecuencias de los

genotipos de la segunda y sucesivas generaciones serían:

$$\begin{aligned}f_2 &= x_1^2 = x_0^2 = f_1 \\g_2 &= 2x_1y_1 = 2x_0y_0 = g_1 \\h_2 &= y_1^2 = y_0^2 = h_1\end{aligned}$$

lo que significa que dichas frecuencias permanecen fijas (b). ■

Dado que las frecuencias de los alelos permanecen fijas (b) y las de los genotipos son constantes de generación en generación (a), se dice que la población está en equilibrio (de Hardy-Weinberg).

Ejemplo 4.2. *Una de las aplicaciones prácticas del principio de Hardy-Weinberg es la estima de la frecuencia de los heterocigotos en una población. La frecuencia de un fenotipo recesivo se puede determinar normalmente contando tales individuos en una muestra de la población. Esta información y la ecuación de Hardy-Weinberg se pueden utilizar para calcular las frecuencias génicas y genotípicas. Veamos esta aplicación práctica en las poblaciones albinas.*

El albinismo es un carácter recesivo autónomo y en algunas poblaciones tiene una incidencia aproximada de 1/10000. Los albinos se distinguen fácilmente en la población por carecer de pigmento en la piel, en el pelo y en el iris. Debido a que este carácter es recesivo, un individuo tiene que ser homocigótico (aa). Su frecuencia en la población, suponiendo que estamos en la n-ésima generación de albinos, está representada por $h_n = y_0^2$, siempre y cuando se hayan cumplido en las generaciones anteriores todas las condiciones de dicho principio. Por consiguiente, la frecuencia del alelo recesivo a es:

$$y_0 = \sqrt{y_0^2} = \sqrt{0,0001} = 0,01$$

Dado que $x_0 + y_0 = 1$, entonces la frecuencia del alelo dominante A es

$$x_0 = 1 - 0,01 = 0,99$$

y en base al principio de Hardy-Weinberg, la frecuencia de los organismos heterocigóticos (Aa) es

$$g_n = 2x_0y_0 = 0,02$$

lo que significa que en el albinismo los organismos heterocigóticos son relativamente frecuentes en la población (2%), aún cuando la incidencia de los homocigóticos recesivos es de sólo 0,0001.

La relación entre las frecuencias genotípicas y alélicas se muestra en la Figura 4.1. Es importante advertir el rápido aumento de la frecuencia de los organismos heterocigóticos en una población cuando los valores de x_0 o de y_0 se alejan de cero. Esta observación confirma nuestra conclusión de que cuando un carácter recesivo, como el albinismo, es raro, la mayoría de los que llevan el alelo son heterocigóticos. En casos en los que las frecuencias de x_0 y de y_0 se encuentran entre 0,33 y 0,67, los heterocigóticos constituyen la clase más numerosa de la población.

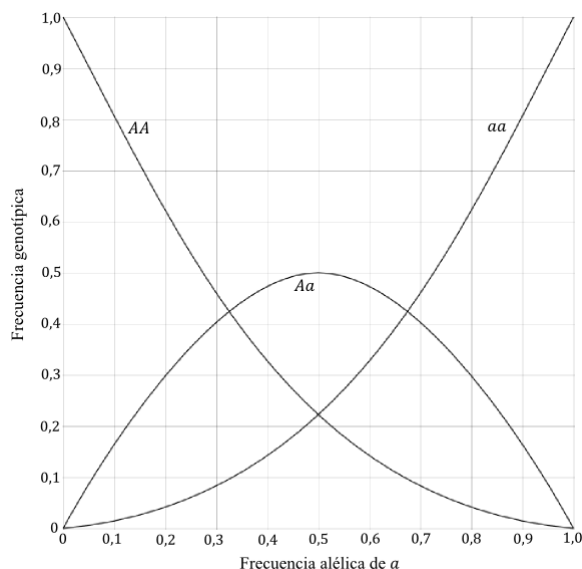


Figura 4.1: Relaciones entre las frecuencias genotípicas y alélicas de acuerdo con el principio de Hardy-Weinberg.

4.2 Apareamiento aleatorio sin mutación

En poblaciones **finitas**, sin embargo, existe la llamada *deriva aleatoria* que conduce finalmente a la eliminación de los genotipos heterocigóticos. El modelo matemático que describe este fenómeno es una cadena de Markov. En un primer caso, supongamos que no hay mutación, esto es, los genes no pueden cambiar de forma. No consideraremos tampoco otros factores que influyen en la evolución de las frecuencias génicas como son la selección natural o la migración.

Supongamos, por tanto, una población diploide finita que está formada por N organismos y nuevamente sean AA , Aa y aa los tres genotipos de los genes alelos. El número total de alelos se fija en $2N$ en todas las generaciones, suponiendo que el tamaño total de la población es constante en el tiempo.

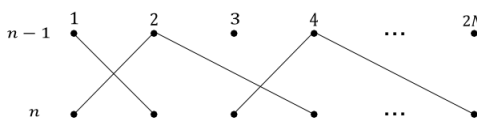


Figura 4.2: Ejemplo de reproducción aleatoria en el modelo de Wright-Fisher.

Notemos por X_n al número de alelos A de la n -ésima generación (en consecuencia, el número de alelos a será $2N - X_n$). Supuesto que los $2N$ alelos de cualquier generación se eligen al azar de entre los alelos de la generación anterior en $2N$ pruebas de Bernoulli, en las que la probabilidad de un alelo dado A o a es igual a su frecuencia en la generación anterior, entonces X_n es una variable aleatoria y la sucesión $\{X_n; n \geq 0\}$ es una cadena de Markov con espacio de estados $E =$

$\{0, 1, 2, \dots, 2N\}$, cuyas probabilidades de transición, $P[X_{n+1} = j | X_n = i]$ para $i, j \in E$, se determinan teniendo en cuenta que supuesto que $X_n = i$, entonces la v.a. X_{n+1} es una binomial de parámetros $2N$ y $p_i = \frac{i}{2N}$, esto es, $X_{n+1} | X_n = i \sim \mathcal{B}(2N, p_i)$. Luego,

$$P_{ij} = P[X_{n+1} = j | X_n = i] = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \quad i, j \in E \quad (4.4)$$

Se obtienen así las probabilidades de la matriz de transición \mathbf{P} de la cadena, que es una matriz cuadrada de orden $2N+1$. Dado que sus elementos no dependen de n , la cadena es homogénea en el tiempo y, en consecuencia, su comportamiento queda determinado por dicha matriz \mathbf{P} así como por la distribución inicial de la cadena, la de la v.a. X_0 .

De (4.4) se deduce que

$$P_{00} = P_{2N,2N} = 1$$

lo que significa que los estados 0 (genotipo aa) y $2N$ (genotipo AA) son absorbentes y, puesto que la cadena es finita, la absorción es segura (Teorema 3.6). Se tiene así que la población termina siendo homocigótica.

Por otra parte, puesto que $X_{n+1} | X_n = i \sim \mathcal{B}(2N, p_i)$, entonces

$$E[X_{n+1} | X_n = i] = 2N \left(\frac{i}{2N}\right) = i \quad (4.5)$$

para todo $n \geq 0$. En consecuencia,

$$E[X_{n+1} | X_n] = X_n \quad c.s \quad (4.6)$$

y la cadena es una *martingala*, lo que significa que el número esperado de alelos tipo A en cada generación se corresponde con el número de alelos tipo A observados en la generación inmediatamente anterior. Aplicando además las propiedades de la esperanza condicionada, se obtiene

$$E[E[X_{n+1} | X_n]] = E[X_n] \quad \Rightarrow \quad E[X_{n+1}] = E[X_n] \quad (4.7)$$

y puede afirmarse así que el número medio de alelos A permanece constante en el tiempo.

Dado que la cadena es finita, la absorción es segura (Teorema 3.6) y el siguiente resultado nos proporciona las probabilidades de absorción, llamadas *probabilidades de fijación*.

Teorema 4.2. (Probabilidades de fijación.) *En el modelo de Wright-Fisher con apareamiento aleatorio sin mutación, las probabilidades de fijación (absorción) en los alelos A (estado $2N$) y a (estado 0), son respectivamente*

$$P[X_\tau = 2N | X_0 = i] = \frac{i}{2N} \quad y \quad P[X_\tau = 0 | X_0 = i] = \frac{2N-i}{2N} \quad (4.8)$$

siendo

$$\tau = \min\{n : X_n = 0 \text{ ó } X_n = 2N\}$$

la generación en la que por primera vez todos los organismos son homocigóticos.

Demostración:

En efecto, puesto que

$$X_n = \begin{cases} X_n & \text{si } n < \tau \\ X_\tau & \text{si } n \geq \tau \end{cases}$$

se tiene que

$$E[X_n|X_0 = i] = E[X_\tau, \tau \leq n|X_0 = i] + E[X_n, \tau > n|X_0 = i] \quad (4.9)$$

con lo cual, en el límite para $n \rightarrow \infty$ y teniendo en cuenta que la v.a. X_n está acotada, pues $|X_n| \leq 2N$, se obtiene

$$\lim_{n \rightarrow \infty} E[X_n|X_0 = i] = E[X_\tau|X_0 = i] = i$$

en virtud de la propiedad de martingala y, puesto que además se trata de la esperanza matemática de una binomial,

$$E[X_\tau|X_0 = i] = 2NP[X_\tau = 2N|X_0 = i] = i$$

de donde

$$P[X_\tau = 2N|X_0 = i] = \frac{i}{2N}$$

con lo cual

$$P[X_\tau = 0|X_0 = i] = 1 - \frac{i}{2N} = \frac{2N - i}{2N}$$

■

Heterocigosidad

Dado que la población termina siendo homocigótica, los organismos heterocigóticos se extinguen, un aspecto interesante es analizar la rapidez con la que ocurre. Para ello, introducimos el concepto de *heterocigosidad*, que es una medida de la variación genética de una población respecto a un locus particular, y se define como la probabilidad de heterocigotos para ese locus. Así, puesto que X_n describe el número de alelos A en la n -ésima generación, entonces la probabilidad H_n de que en un mismo locus cualquier par de alelos seleccionados al azar y sin reemplazamiento sean diferentes entre sí viene dada por

$$H_n = 2 \frac{X_n}{2N} \frac{2N - X_n}{2N - 1} \quad (4.10)$$

ya que hay dos casos posibles y equiprobables según que la primera copia sea A y después a o al contrario.

Se tiene en definitiva que la heterocigosidad es una variable aleatoria y, en consecuencia, tiene sentido el siguiente

Teorema 4.3. *En el modelo de Wright-Fisher con apareamiento aleatorio sin mutación, el valor esperado de la heterocigosidad en la n -ésima generación es*

$$E[H_n] = \left(1 - \frac{1}{2N}\right)^n E[H_0] \quad (4.11)$$

Demostración:

Es conveniente numerar los $2N$ alelos y referirnos a ellos como individuos. Supongamos que, en el instante n , elegimos dos individuos numerados por $x_1(0)$ y $x_2(0)$. Cada individuo $x_i(0)$ es descendiente de otro individuo $x_i(1)$ en el tiempo $n-1$, que a su vez es descendiente de $x_i(2)$ en el tiempo $n-2$, etc. Luego, $x_i(m)$, $0 \leq m \leq n$ describe la ascendencia de $x_i(0)$ mirando hacia atrás en el tiempo.

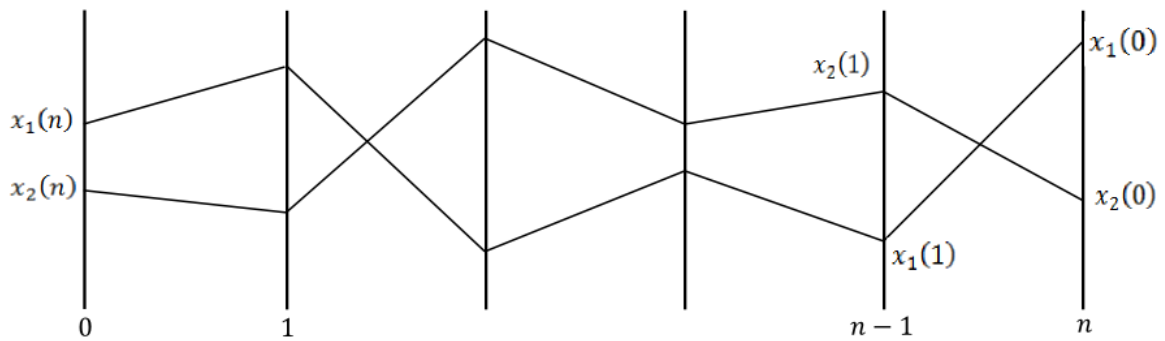


Figura 4.3: Un par de genealogías.

Si $x_1(m) = x_2(m)$, entonces se tiene que $x_1(l) = x_2(l)$ para $m < l \leq n$. Si $x_1(m) \neq x_2(m)$, entonces la probabilidad de que al elegir $x_1(m+1)$ y $x_2(m+1)$ se tenga $x_1(m+1) \neq x_2(m+1)$ es $1 - \frac{1}{2N}$. Para que $x_1(n) \neq x_2(n)$ debe cumplirse que $x_1(m)$ y $x_2(m)$ sean distintos para $1 \leq m \leq n$, lo que ocurre con probabilidad $\left(1 - \frac{1}{2N}\right)^n$. Cuando los dos linajes son independientes, $x_1(n)$ y $x_2(n)$ son dos individuos de la población elegidos al azar en el tiempo 0 , de modo que la probabilidad de que sean diferentes es H_0 . Luego,

$$E[H_n] = \left(1 - \frac{1}{2N}\right)^n E[H_0]$$

■

Si N es grande, utilizando la aproximación (dada por el desarrollo en serie de Taylor de la función exponencial) $1 - x \approx \exp(-x)$ para x pequeño, deducimos que

$$E[H_n] \approx \exp\left(-\frac{n}{2N}\right) E[H_0],$$

y así la heterocigosidad decrece exponencialmente a lo largo de las generaciones.

Ejemplo 4.3. Vamos a probar heurísticamente, mediante el siguiente ejemplo numérico, que una población finita de individuos que se aparean al azar y en la que no hay mutación termina siendo homocigótica. Consideremos $N = 2$. Entonces el espacio de estados es $E = \{0, 1, 2, 3, 4\}$. Calculando las probabilidades de transición según (4.4), obtenemos que la matriz de transición en 1 etapa es

$$P = \frac{1}{256} \begin{pmatrix} 256 & 0 & 0 & 0 & 0 \\ 81 & 108 & 54 & 12 & 1 \\ 16 & 64 & 96 & 64 & 16 \\ 1 & 12 & 54 & 108 & 81 \\ 0 & 0 & 0 & 0 & 256 \end{pmatrix}$$

Elegimos una distribución inicial para X_0 tal que la probabilidad de que la población sea homocigótica es 0:

$$\mathbf{p}(0) = \left(0, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0\right)$$

Calculamos ahora la distribución de X_1 , el número de alelos A de la primera generación, como $\mathbf{p}(1) = \mathbf{p}(0)P$, obteniéndose

$$\mathbf{p}(1) = (0,1113, 0,2422, 0,2930, 0,2422, 0,1113)$$

Similarmente, la distribución de X_2 se calcula por $\mathbf{p}(2) = \mathbf{p}(0)P^2$ y se obtiene

$$\mathbf{p}(2) = (0,2072, 0,1868, 0,2121, 0,1868, 0,2072)$$

Las distribuciones de probabilidad del número de alelos tipo A de las siguientes cuatro generaciones se han calculado de igual modo y son las siguientes:

$$\mathbf{p}(3) = (0,2803, 0,1406, 0,1583, 0,1406, 0,2803)$$

$$\mathbf{p}(4) = (0,3352, 0,1055, 0,1187, 0,1055, 0,3352)$$

$$\mathbf{p}(5) = (0,3764, 0,0791, 0,0890, 0,0791, 0,3764)$$

$$\mathbf{p}(6) = (0,4073, 0,0593, 0,0667, 0,0593, 0,4073)$$

La figura 4.4 muestra los diagramas de las distribuciones de X_0, X_1, \dots, X_6 . Puede verse que ya en la tercera generación ($n = 3$) hay una mayor masa de probabilidad concentrada en los estados homocigóticos que en los heterocigóticos. Esto contrasta con la situación de la población inicial en la que la probabilidad de estados homocigóticos era 0. Para la sexta generación la probabilidad de que la población sea homocigótica se ha elevado a 0,8146. Con el tiempo, la probabilidad de que la población sea heterocigótica es nula.

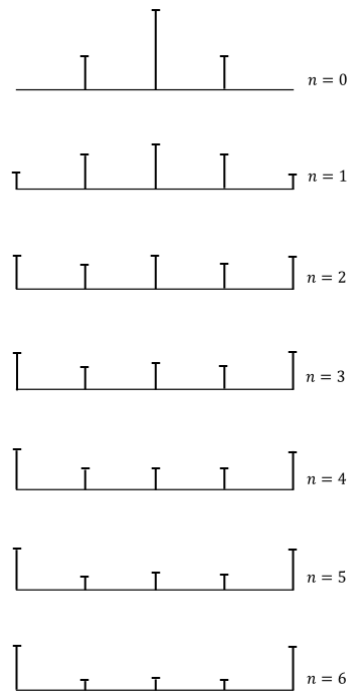


Figura 4.4: Evolución de la distribución de probabilidad del número de alelos tipo A cuando $N = 2$.

Vamos a calcular también la heterocigosidad a partir de la distribución inicial para X_0 que hemos considerado. Por (4.10) tenemos que para $N = 2$

$$H_0 = \frac{X_0(4 - X_0)}{8} \quad X_0 = 0, 1, 2, 3, 4$$

de donde se sigue que

$$\begin{aligned} E[H_0] &= \frac{0(4-0)}{8}P[X_0 = 0] + \frac{1(4-1)}{8}P[X_0 = 1] + \frac{2(4-2)}{8}P[X_0 = 2] + \\ &+ \frac{3(4-3)}{8}P[X_0 = 3] + \frac{4(4-4)}{8}P[X_0 = 4] = \frac{3}{8} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{2} + \frac{3}{8} \cdot \frac{1}{4} = \\ &= \frac{7}{16} \end{aligned}$$

de modo que el valor esperado para la heterocigosidad en la n -ésima generación viene dado, en virtud de (4.11), por

$$E[H_n] = \left(\frac{3}{4}\right)^n \cdot \frac{7}{16}$$

cuya representación gráfica nos muestra su decrecimiento exponencial a lo largo de las generaciones.

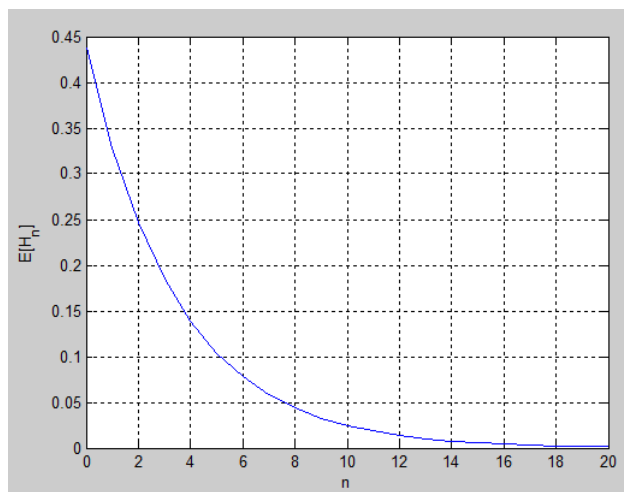


Figura 4.5: Representación gráfica de $E[H_n]$ respecto a n .

4.3 Apareamiento aleatorio con mutación

Aumentemos ahora la complejidad del modelo de Wright-Fisher anterior, suponiendo que en el paso de una generación a la siguiente puede haber mutación: un alelo A puede transformarse en a con probabilidad α_1 y un alelo a puede transformarse en A con probabilidad α_2 .

Suponiendo que partimos de una composición de i alelos tipo A en la n -ésima generación y en virtud del esquema siguiente,

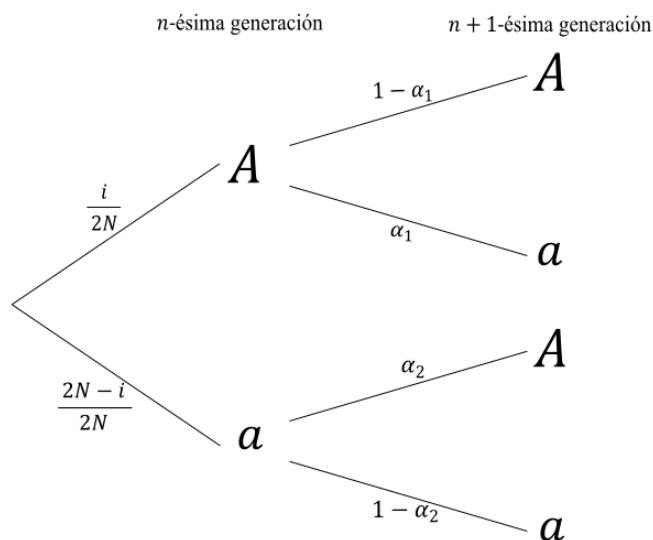


Figura 4.6: Mutaciones

la probabilidad de A en la $n + 1$ -ésima generación vendrá dada por

$$p_A(i) = (1 - \alpha_1) \frac{i}{2N} + \alpha_2 \frac{2N - i}{2N} \tag{4.12}$$

y la probabilidad de a será

$$p_a(i) = \alpha_1 \frac{i}{2N} + (1 - \alpha_2) \frac{2N - i}{2N} \quad (4.13)$$

Así pues, la cadena de Markov $\{X_n, n \in \mathbb{N}\}$ que describe el modelo de Wright - Fisher se ha modificado al tenerse en cuenta el fenómeno de la mutación. En este caso, supuesto que $X_n = i$, la v.a. X_{n+1} sigue una binomial de parámetros $2N$ y $p_A(i)$, esto es, $X_{n+1}|X_n = i \sim \mathcal{B}(2N, p_A(i))$. Las probabilidades de transición, $P[X_{n+1} = j|X_n = i]$, vienen dadas entonces por

$$P_{ij} = \binom{2N}{j} p_A(i)^j p_a(i)^{2N-j} \quad (4.14)$$

de modo que todos los elementos de la matriz de transición \mathbf{P} de la cadena son estrictamente positivos y, por tanto, la cadena de Markov es irreducible. Como la cadena es finita, todos sus estados son recurrentes positivos, y además, dado que son aperiódicos, son ergódicos, y por tanto, en virtud del Teorema 3.7, existe la distribución de equilibrio

$$\pi_j = \lim_{n \rightarrow \infty} P_{ij}(n) \quad j = 0, \dots, 2N$$

que por coincidir con la estacionaria puede obtenerse como única solución de

$$\begin{aligned} \pi_j &= \sum_{i=0}^{2N} \pi_i P_{ij} \quad j = 0, \dots, 2N \\ \sum_{j=0}^{2N} \pi_j &= 1 \end{aligned}$$

Analicemos ahora el valor límite del número medio de alelos A . Definamos para ello

$$X_\infty = \lim_{n \rightarrow \infty} X_n$$

y teniendo en cuenta que $X_{n+1}|X_n = i \sim \mathcal{B}(2N, p_A(i))$,

$$\begin{aligned} E[X_{n+1}|X_n = i] &= 2N p_A(i) = 2N \left[(1 - \alpha_2) \frac{i}{2N} + \alpha_2 \frac{2N - i}{2N} \right] \\ &= (1 - \alpha_1)i + \alpha_2(2N - i) \end{aligned}$$

Generalizando,

$$E[X_{n+1}|X_n] = (1 - \alpha_1)X_n + \alpha_2(2N - X_n) \quad \text{c.s.}$$

Por consiguiente,

$$E[E[X_{n+1}|X_n]] = (1 - \alpha_1)E[X_n] + \alpha_2(2N - E[X_n])$$

y por las propiedades de la esperanza condicionada,

$$E[X_{n+1}] = (1 - \alpha_1)E[X_n] + \alpha_2(2N - E[X_n])$$

Tomando límites,

$$\lim E[X_{n+1}] = (1 - \alpha_1) \lim E[X_n] + \alpha_2(2N - \lim E[X_n])$$

esto es,

$$E[X_\infty] = (1 - \alpha_1)E[X_\infty] + \alpha_2(2N - E[X_\infty])$$

cuya solución es

$$E[X_\infty] = \frac{2N\alpha_2}{\alpha_1 + \alpha_2} \quad (4.15)$$

Ejemplo 4.4. Consideremos una población con un único organismo diploide, $N = 1$, y supongamos que las probabilidades de mutación de los alelos A y a son $\alpha_1 = \alpha_2 = \frac{1}{4}$. En este caso, $E = \{0, 1, 2\}$. Sustituyendo en (4.12) y (4.13) obtenemos que las probabilidades de A y a en la $n+1$ -ésima generación, partiendo de una composición de i alelos tipo A en la n -ésima generación, vienen dadas por

$$p_A(i) = \frac{i}{4} + \frac{1}{4} \quad p_a(i) = \frac{3}{4} - \frac{i}{4} \quad \text{para } i \in E$$

Los elementos de \mathbf{P} son entonces, por (4.14), para $j \in E$

$$P_{0j} = \binom{2}{j} \left(\frac{1}{4}\right)^j \left(\frac{3}{4}\right)^{2-j}$$

$$P_{1j} = \binom{2}{j} \left(\frac{1}{2}\right)^2$$

$$P_{2j} = \binom{2}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{2-j}$$

que dan lugar a la matriz de transición

$$\mathbf{P} = \frac{1}{16} \begin{pmatrix} 9 & 6 & 1 \\ 4 & 8 & 4 \\ 1 & 6 & 9 \end{pmatrix}$$

Se observa que no hay estados absorbentes, ya que no hay unos en la diagonal principal, y que cualquier transición de un estado a otro es posible, pues no hay ceros en \mathbf{P} .

Calculemos ahora la distribución de equilibrio

$$\pi_j = \lim_{n \rightarrow \infty} P_{ij}(n) \quad j = 0, 1, 2$$

Para ello hemos de resolver el sistema

$$\pi_j = \sum_{i=0}^2 \pi_i P_{ij} \quad j = 0, 1, 2$$

$$\sum_{j=0}^2 \pi_j = 1$$

esto es,

$$\begin{aligned}\pi_0 &= \frac{1}{16}(9\pi_0 + 4\pi_1 + \pi_2) \\ \pi_1 &= \frac{1}{16}(6\pi_0 + 8\pi_1 + 6\pi_2) \\ \pi_2 &= \frac{1}{16}(\pi_0 + 4\pi_1 + 9\pi_2) \\ \pi_0 + \pi_1 + \pi_2 &= 1\end{aligned}$$

obteniéndose

$$\pi_0 = \frac{2}{7} \quad \pi_1 = \frac{3}{7} \quad \pi_2 = \frac{2}{7}$$

El valor límite esperado del número de alelos tipo A será, en virtud de (4.15),

$$E[X_\infty] = \frac{2\frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

Ejemplo 4.5. Consideremos ahora $N = 2$ y repitamos los mismos pasos que en el Ejemplo 4.4, considerando también las probabilidades de mutación $\alpha_1 = \alpha_2 = \frac{1}{4}$. En este caso, $E = \{0, 1, 2, 3, 4\}$ y $p_A(i) = \frac{i}{8} + \frac{1}{4}$, $p_a(i) = \frac{3}{4} - \frac{i}{8}$ para $i \in E$. Las probabilidades de transición, para $j \in E$, son

$$\begin{aligned}P_{0j} &= \binom{4}{j} \left(\frac{1}{4}\right)^j \left(\frac{3}{4}\right)^{4-j} \\ P_{1j} &= \binom{4}{j} \left(\frac{3}{8}\right)^j \left(\frac{5}{8}\right)^{4-j} \\ P_{2j} &= \binom{4}{j} \frac{1}{16} \\ P_{3j} &= \binom{4}{j} \left(\frac{5}{8}\right)^j \left(\frac{3}{8}\right)^{4-j} \\ P_{4j} &= \binom{4}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{4-j}\end{aligned}$$

que dan lugar a la matriz de transición

$$P = \frac{1}{4096} \begin{pmatrix} 1296 & 1728 & 864 & 192 & 16 \\ 625 & 1500 & 1350 & 540 & 81 \\ 256 & 1024 & 1536 & 1024 & 256 \\ 81 & 540 & 1350 & 1500 & 625 \\ 16 & 192 & 864 & 1728 & 1296 \end{pmatrix}$$

Comparando esta matriz con la del Ejemplo 4.3, en el que considerábamos también una población con 2 organismos diploides pero obviando el fenómeno de la mutación,

podemos ver que en este caso no hay estados absorbentes al no haber unos en la diagonal principal, y tampoco hay estados que no se comuniquen entre sí al no haber ceros en la matriz.

Calculamos la distribución de equilibrio

$$\pi_j = \lim_{n \rightarrow \infty} P_{ij}(n) \quad j \in E$$

resolviendo el sistema

$$\pi_j = \sum_{i=0}^4 \pi_i P_{ij} \quad j = 0, 1, 2, 3, 4$$

$$\sum_{j=0}^4 \pi_j = 1$$

esto es,

$$\begin{aligned} \pi_0 &= \frac{1}{4096} (1296\pi_0 + 625\pi_1 + 256\pi_2 + 81\pi_3 + 16\pi_4) \\ \pi_1 &= \frac{1}{4096} (1728\pi_0 + 1500\pi_1 + 1024\pi_2 + 540\pi_3 + 192\pi_4) \\ \pi_2 &= \frac{1}{4096} (864\pi_0 + 1350\pi_1 + 1536\pi_2 + 1350\pi_3 + 864\pi_4) \\ \pi_3 &= \frac{1}{4096} (192\pi_0 + 540\pi_1 + 1024\pi_2 + 1500\pi_3 + 1728\pi_4) \\ \pi_4 &= \frac{1}{4096} (16\pi_0 + 81\pi_1 + 256\pi_2 + 625\pi_3 + 1296\pi_4) \\ \pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 &= 1 \end{aligned}$$

obteniéndose

$$\pi_0 = \frac{610}{6617} \quad \pi_1 = \frac{1632}{6617} \quad \pi_2 = \frac{2133}{6617} \quad \pi_3 = \frac{1632}{6617} \quad \pi_4 = \frac{610}{6617}$$

Por (4.15), el valor límite del número medio de alelos A es

$$E[X_\infty] = \frac{4 \frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = \frac{1}{\frac{1}{2}} = 2$$

Puede apreciarse que la distribución límite obtenida es simétrica, lo cual también ocurre en el Ejemplo 4.4. Esto se debe a que las probabilidades de transición obedecen a leyes de probabilidad binomiales. Además, podemos observar que tanto en la distribución límite de este ejemplo como en la del Ejemplo 4.4, el número de alelos tipo A de mayor masa de probabilidad coincide con el valor límite obtenido para el número medio de alelos A , lo cual es coherente pues en distribuciones simétricas la moda y la media coinciden.

Conclusiones

Como ya se mencionó anteriormente, el objetivo de nuestro trabajo ha sido describir el modelo de Wright-Fisher para la transmisión de caracteres genéticos en organismos diploides, que son los que tienen los cromosomas semejantes dos a dos. La especie humana es un ejemplo.

Nos hemos centrado en el estudio de la transmisión de dos alelos A y a de un gen de una población de organismos diploides y hacerlo desde el punto de visto probabilístico, dada la relación entre la genética y las matemáticas.

Para ello hemos recurrido a los conceptos y herramientas proporcionados por la teoría de los procesos estocásticos, considerados como la parte dinámica de la teoría de la probabilidad. En este contexto, hemos observado que el modelo de Wright-Fisher puede describirse mediante las llamadas cadenas de Markov, una clase importante y muy amplia de proceso estocástico que nos permite deducir la probabilidad de que el fenómeno aleatorio en estudio esté en cierto estado en un instante determinado t , conocido su estado en un instante anterior $\tau (< t)$ y ello no depende de la evolución o historia del fenómeno anterior a τ .

Nuestro modelo, además, resulta ser una cadena de Markov homogénea en el tiempo, lo que significa, en definitiva, que las transiciones entre los diversos estados del fenómeno son estacionarias, es decir, dependen del lapso o tiempo transcurrido, $t - \tau$. Este hecho simplifica el análisis del comportamiento o evolución, tanto transitorio como a largo plazo de la composición genética de la población.

Es un modelo que, aunque se basa en los mismo supuestos que el conocido principio de Hardy-Weinberg: los individuos se aparean al azar y no intervienen otros factores como las mutaciones, migraciones ni la selección natural, se refiere a poblaciones finitas.

Nuestro estudio, sin embargo, ha abarcado dos variantes del modelo. La primera de ellas, la básica, descarta la mutación entre los dos alelos A y a de un mismo gen. En este caso, al tratarse de una cadena de Markov homogénea finita, sus transiciones se explican mediante leyes de probabilidad binomiales y, dado que además, hay dos estados absorbentes que son los genotipos AA y aa , la población termina siendo homocigótica, esto es, los individuos heterocigóticos (Aa) se extinguen. Hemos analizado también la velocidad de dicha extinción, mediante el concepto de heterocigosidad, y el resultado es que el número medio de individuos u organismos heterocigóticos decrece exponencialmente a lo largo de las generaciones.

En la segunda variante hemos supuesto que los alelos A y a pueden mutar entre sí. Se trata nuevamente de una cadena de Markov homogénea finita cuyas probabilidades de transición también se describen por leyes de probabilidad binomiales pero, puesto que la cadena es ergódica, queda garantizada la existencia de la distribución de equilibrio o estacionaria para el número de alelos de la población. En este mismo caso, hemos analizado también el valor límite del número medio de alelos de la población.

El modelo de Wright-Fisher que hemos descrito es el más simple, pero tiene la virtud de ser el punto de partida de desarrollos más complejos, pudiendo gene-

5. CONCLUSIONES

realizarse a genes con k alelos A_1, \dots, A_k , con o sin mutaciones entre ellos. En tales situaciones, las probabilidades de transición de las correspondientes cadena markovianas estarán relacionadas con distribuciones de probabilidad multinomiales.

Bibliografía

- [1] U. Narayan Bhat, *Elements of applied stochastic processes*, John Wiley, 1984.
- [2] R. Durrett, *Probability models of DNA Sequence Evolution*. Probability and its Applications (2002).
- [3] S. M. Ross, *Introduction to probability models*, Elsevier, 2014.
- [4] M. Sanz Solé, *Paseos aleatorios en genética*, Las matemáticas y sus aplicaciones en el mundo social y económico (2006), 37–60.
- [5] H. C. Tuckwell, *Elementary applications of probability abrst*, Chapman Hall, 1995.
- [6] http://www.uam.es/personal_pdi/ciencias/joaquina/BOXES_CCAA/Constitucion_genetica_pob.htm.
- [7] <http://es.scribd.com/doc/48010509/Genetica-Poblaciones-03Jun07>
- [8] <http://www.elergonomista.com/biologiasselectividad/>