TRABAJO MONOGRÁFICO

Modelos aleatorios en Genética de Poblaciones: Estimación de parámetros de mutación.

María Inés Fariello Rico

Orientador: Dr. Gustavo Guerberoff

Centro de Matemática

Licenciatura en Matemática
Facultad de Ciencias
Universidad de la República

3 de Setiembre de 2008

Uruguay

Prólogo

El tema de esta monografía es el producto de una serie de casualidades.

Empecé a estudiar matemáticas, sólo por saber un poco más, pero nunca pensé que esa iba a ser la carrera que iba a terminar estudiando, sólo quería hacer primero y depués dedicarme a otra cosa, porque pensaba que la matemática " era muy cuadrada " y que me iba a encerrar demasiado en algo, quería hacer algo más multidiciplinario.

El primer año estuvo bueno y decidí seguir con segundo. Resulta que a las clases de Probabilidad y Estadística iba un señor que yo nunca había visto. La última clase este señor, que en ese momento se convirtió en Enrique Lessa y nos enteramos que era profesor de Evolución, nos contó cómo se aplicaban las cadenas de Markov al coalescente. Ese día descubrí que había un punto de encuentro entre la biología y la matemática y que con la matemática podía aprender mucho sobre temas relacionados con la genética, un área de la biología que en el liceo me había gustado mucho, pero que pensaba que en Uruguay no había. Si Enrique no hubiera dado esa última clase, no sé si esta monografía existiría.

En realidad gracias a Enrique y al Pájaro, que tuvimos que ir a un congreso de biomatemática para conocernos y decidirnos a hacer una monografía en este tema es que estoy estudiando estos temas. Es más, creo que sino fuera por Enrique, no me hubiera involucrado con la Genética de Poblaciones.

Agradecimientos

Gracias a mami, papi y Pampi que me aguantaron toda la carrera y sobre todo antes de los exámenes, que supieron ser días complicados para mi, y bueno, para ellos también. Además por apoyarme aunque nunca tuvieron muy claro qué era lo que estaba haciendo y para qué servía, espero que la monografía los ayude un poquito.

A Ramirito por alivianarle la tarea durante estos casi dos últimos años a la flia., por acompañarme en lo que sea y porque siempre me da para adelante.

A Dalita, que sin ella la licenciatura no hubiera sido tan divertida y ni que hablar los veranos de estudio en Montevideo. A Ceci que también anduvo siempre por la vuelta. Al Gordo por prestarme sus cuadernos y su paciencia. Y a todos mis compañeros de la licenciatura que hicieron que ir a clase fuera un experiencia distinta todos los días.

A Laurita, que no dudó en escucharme dos veces la presentación, por sus aportes.

A Ana y Maryo que no pararon de darme empujoncitos para que terminara esta monografía.

Al Pájaro y a Enrique, el por qué ya lo dije en el prólogo.

Índice general

Capítulo 1

Inroducción

Este trabajo está basado principalmente en tres textos diferentes: unas notas de Simon Tavaré hechas para una escuela de verano de Probabilidad y Estadística del 2001 [12], un proyecto de libro de Rick Durret [1] y las notas sobre Genética de Poblaciones que escribió el profesor Enrique P. Lessa como apoyo para el curso de Evolución que se dicta en la Facultad de Ciencias [8].

El objetivo principal de esta monografía es presentar algunos estadísticos que permiten estimar el parámetro de mutación de poblaciones que cumplan ciertas características. Para hacerlo es necesario entender bien qué pasa con la población a lo largo del tiempo, es decir, cómo se transmiten los alelos de una generación a otra o cómo son las relaciones ancestrales entre los individuos de una misma población.

Se comenzó presentando el modelo genético poblacional desarrollado por Hardy y Weinberg, ya que es el modelo más simple dentro de los existentes referidos a la genética de poblaciones y a su vez el punto de partida para desarrollar otros modelos; ésto se debe a la gran cantidad de hipótesis que debe cumplir una población para que se cumpla el equilibrio de Hardy Weinberg.

El mecanismo empleado en los Capítulos II y III fue ir quitándole hipótesis al modelo del primer capítulo e ir viendo cómo cambiaba. Es así que en el segundo capítulo se presenta el modelo de Fisher y Wright; éste surge de quitárle la hipótesis de infinitud al tamaño poblacional de la muestra. Para analizar la evolución de la población se hizo de dos maneras diferentes: mirando al tiempo hacia adelante y de manera retrospectiva. En el primer caso surge un modelo que se basa en Cadenas de Markov y sirve para calcular cuánto tiempo es probable que demore un alelo en fijarse y en el segundo caso surge el modelo del coalescente.

En el tercer capítulo se introduce la mutación, lo que supone un gran cambio, ya que es el recurso que permite que la evolución se sostenga en el tiempo, ya que sin ella un aleló se fijará y a partir de ese momento no habrá más evolución de ese locus. Se verán diferentes modelos:entre ellos se destacan el modelo de Morán, que es un modelo de tiempo continuo y el de la Urna de Hoppe que es de tiempo discreto. Utilizando este modelo se demuestra la fórmula de muestreo de Ewens. Con esta fórmula se ve que el estadístico K_n que también se ve en este capítulo es suficiente. Éste fue uno de los primeros estadísticos del parámetro de mutación de una muestra; el único problema que presenta es que al ser asintótico para estimar el parámetro con error bajo, hay que tomar una muestra muy grande de la población.

En el último capítulo se presentan dos estimadores del parámetro de mutación: uno se debe a Watterson y el otro a Tajima. El uso de uno u otro estimador depende de las hipótsis que cumpla la muestra. También se pueden utilizar en conjunoto para elaborar hipótesis sobre una población. Por ejemplo para saber si se trabaja bajo un modelo neutral o si la población está en su estado de equilibrio. Estas dos aplicaciones de las diferencias entre las estimaciones fueron publicada por Tajima en 1989 y hoy en día se siguen utilizando.

Capítulo 2

Equilibrio de Hardy-Weinberg

G.H. Hardy y W. Weinberg crearon un modelo genético poblacional de manera independiente, en 1908. [4]

Si bien este modelo fue criticado por su falta de realismo, es un modelo se utiliza como hipótesis nula en pruebas de hipótesis que evalúan si una población cumple o no con una serie de propiedades, que se verán más adelante. Cualquier desviación significativa de los valores esperados del modelo indicaríía que alguna hipótesis del modelo está fallando.

Se desarrollará el modelo para un único gen representado por dos alelos. En general el pasaje de dos a más de ellos es directo.

2.1. El modelo de Hardy-Weinberg

Hipótesis del modelo: Las hipótesis se dividen en 2 grandes grupos. Si se modifican las hipótesis del primer grupo, la obtención de un modelo que se adapte a la nueva situación es directa, mientras que hacerlo con las segundas implicará la creación de otros modelos, algunos de los cuales serán objeto de estudio de este trabajo.

$1^{\underline{er}}$ **grupo**

- organismos diploides
- reproducción sexuada
- generaciones que no se solapan
- gen autosómico

- frecuencias que no difieren entre los sexos

$2^{\underline{do}}$ **grupo**

- población infinita
- apareamientos al azar
- ausencia de mutación
- ausencia de migración desde otras poblaciones
- ausencia selección natural sobre los genes considerados

Si bien la primera hipótesis de Hardy y Weinberg sobre los organismos es que sean diploides, esta es una hipótesis prescindible, porque, como se verá más adelante, lo que intresa en realidad son las frecuencias alélicas, en fase haploide (i.e. cociente entre la cantidad de alelos de un tipo y el total de alelos, 2N).

En la fase diploide u organísmica, cada locus está representado por dos alelos, mientras que en la fase aploide o gamética cada locus está representado por sólo uno.

Se considera un gen con 2 alelos (A y a) e individuos diploides.

Gen diploide $\rightarrow$ 2 alelos $\rightarrow$ 3 genotipos: AA, aa, Aa.

Observación 2.1.1. Si hubieran m alelos, habría $\mathrm{C}_2^m = \frac{m(m-1)}{2}$ genotipos.

Construcción del modelo: Los cálculos siguientes se aplican tanto en el caso en que existe codominancia, como en el caso que existe dominancia completa.

Los gametos se forman en la meiosis, obteniéndose células con un juego de cromosomas, por lo tanto cada una tendrá solamente un sólo represante del gen. Los individuos del tipo AA, producirán solamente alelos del tipo A; los aa del tipo a y los Aa, la mitad A y la otra mitad a.

Los individuos de la nueva generación se formarán a parir de dos gametos parentales que portarán cada uno un alelo; en el caso de la reproducción sexuada, uno de la madre y otro del padre.

Se intoducirá la siguiente notación:

d**:** Frecuencia de genotipos dominantes en la generación parental.

h**:** Frecuencia de genotipos heterocigotos en la generación parental.

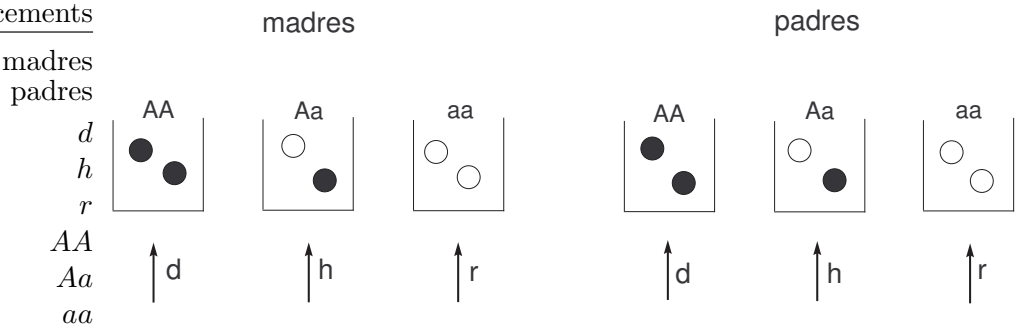
r**:** Frecuencia de genotipos recesivos en la generación parental.

p**:** Frecuencia de alelos dominantes.

q**:** Frecuencia de alelos recesivos.

A demás se utilizará el subíndice 1 para los hijos de la generación parental, el 2 para los hijos de ésta y así sucesivamente para las generaciones siguientes.

PSfrag replacements



- $p = d + \frac{h}{2}$
- $q = r + \frac{h}{2}$

Para calcular las probablilidades que tiene un individuo de la generación siguiente de portar determinado genotipo, se utilizará la siguiente notación:

- AA, aa, Aa para los genotipos de los individuos de la nueva generación.
- $M =$ para indicar que el genotipo es el de la madre.
- $P =$ para indicar que el genotipo es el del padre.

Además se tendrán en cuenta los siquientes hechos e hipótesis:

- Como los gametos se forman en la meiosis, la probabildad de heredar cada uno de los alelos tanto del padre como de la madre es la misma, o sea $\frac{1}{2}$.
- Como la reproducción es al azar, heredar un determinado alelo de la madre no influye en qué alelo se heredará del padre (o sea, los sucesos son independientes)[1].

[1]Las poblaciones de este tipo se dicen *panmíticas*.

- Como las frecuencias genotípicas no dependen del sexo, la probabilidad de heredar un determinado alelo de la madre y otro del padre es la misma que la de heredar el primero del padre y el segundo de la madre. Por ejemplo: $\mathbb{P}(Aa|M = AA, P = aa) = \mathbb{P}(Aa|M = aa,\ P = AA)$ Observar que si las frecuencias genotípicas fueran diferentes en cada sexo, el resultado de la cuenta que sigue es directo.

$$\begin{aligned}
\mathbb{P}(AA) \quad &= \underbrace{\mathbb{P}(AA|M = AA,\ P = AA)\mathbb{P}(M = AA,\ P = AA)}_{1\cdot 1\cdot d\cdot d} + \\
&+ \underbrace{\mathbb{P}(AA|M = AA,\ P = Aa)\mathbb{P}(M = AA,\ P = Aa)}_{1\cdot \frac{1}{2}\cdot d\cdot h} + \\
&+ \underbrace{\mathbb{P}(AA|M = AA,\ P = aa)\mathbb{P}(M = AA,\ P = aa)}_{0} + \\
&+ \underbrace{\mathbb{P}(AA|M = Aa,\ P = AA)\mathbb{P}(M = Aa,\ P = AA)}_{\frac{1}{2}\cdot 1\cdot h\cdot d} + \\
&+ \underbrace{\mathbb{P}(AA|M = Aa,\ P = Aa)\mathbb{P}(M = Aa,\ P = Aa)}_{\frac{1}{2}\cdot \frac{1}{2}\cdot h\cdot h} + \\
&+ \underbrace{\mathbb{P}(AA|M = Aa,\ P = aa)\mathbb{P}(M = Aa,\ P = aa)}_{0} + \\
&+ \underbrace{\mathbb{P}(AA|M = aa,\ P = AA)\mathbb{P}(M = aa,\ P = AA)}_{0} + \\
&+ \underbrace{\mathbb{P}(AA|M = aa,\ P = Aa)\mathbb{P}(M = aa,\ P = Aa)}_{0} + \\
&+ \underbrace{\mathbb{P}(AA|M = aa,\ P = aa)\mathbb{P}(M = aa,\ P = aa)}_{0} = \\
&= d^2 + 2\cdot \tfrac{1}{2} d\cdot h + \tfrac{1}{4} h^2 \\
&= (d + \tfrac{h}{2})^2 = p^2
\end{aligned}$$

De manera análoga se obtiene:

$$\mathbb{P}(Aa) = 2pq$$
$$\mathbb{P}(aa) = q^2$$

Este cálculo implica que si se consideran poblaciones diferentes, no importa en realidad qué valores que tomen d, h y r, sino que si se mantienen p y q constantes, los genotipos posibles de la generación siquiente tendrán las mismas probabilidades de ocurrir, ya que dependen solamente de estos últimos parámetros. De manera que en la generación siguiente se tendrá con probabilidad 1:

$$d_1 = p^2, \quad h_1 = 2pq, \quad r_1 = q^2$$

Al calcular los nuevos p_1 y q_1 se obtiene:

$$p_1 = d_1 + \frac{h_1}{2} = p^2 + \frac{2pq}{2} = p(p+q) = p$$

$$q_1 = r_1 + \frac{h_1}{2} = q^2 + \frac{2pq}{2} = q(p+q) = q$$

Por lo tanto las frecuencias alélicas se manterndrán constantes con probabilidad 1. A las distirbuciones genotípicas que cumplen esta propiedad se les llama *estacionarias* o distribuciones equilibradas.

En poblaciones muy grandes[2] las frecuencias observadas de los 3 genotipos estarán cerca de las probabilidades teóricas dadas por las ecuaciones anteriores. Por lo tanto si las frecuencias observadas están cerca[3] de las frecuencias esperadas por el modelo, se puede decir que la descendencia tendrá una disribución genotípica estacionaria, que continuará prácticamente sin cambios en las generaciones siguientes.

En la práctica se obseravarán desviaciones, pero para poblaciones con tamaños muy grandes se puede decir que: *Cualquiera sea la composición de la población parental, la reproducción al azar producirá una distribución genotípica aproximadamente estacionaria con frecuencias genotípicas constantes.*

En el caso que haya una población que cumpla las hipótesis pero que no se encuentre en equilibrio, éste se reestablecerá luego de una generación. Esto se debe a que las frecuencias alélicas dependen solamente de las genotípicas y que las frecuencias genotípicas de la siguiente generación, dependen de las frecuencias alélicas de la generación anterior. O sea, $d_1 = p^2,\ d_2 = p^2, \ldots$, sin importar cuál haya sido el d inicial.

Es importante destacar que si las frecuencias genotípicas entre los sexos son diferentes, luego de una generación éstas se emparejarán, ya que si el gen no está ligado al sexo la probabilidad de obtener un genotipo determinado es independiente del sexo y la probabilidad de tener una hija o un hijo es la misma. Luego de emparejarse se necesitará otra generación para que se reestablezca el equilibrio. Por lo tanto, el equilibrio Hardy-Weinberg se establecerá luego de dos generaciones.

[2]Sin esta condición el modelo probabilístico no tiene significado operacional, ya que este enunciado se debe a la Ley de los Grandes Números y al Teorema Central del Límite, que permite estimar el efecto de las fluctuaciones

[3]Para evaluar si las frecuencias están cerca, se puede utilizar el test de chi cuadrado.

Hardy al desarrollar su modelo destacó que las frecuncias observadas y las esperadas estarán cerca, ya que, a pesar de que la distribución es estacionaria, en las poblaciones naturales se espera que ocurran fluctuaciones de una generación a otra, que irán cambiando a p y a q en pequeñas cantidades a lo largo de la historia de la población. Como no existe una fuerza restauradora de las frecuencias, en una población finita, las frecuencias de los alelos irá fluctuando hasta que uno de los alelos se fije. La consecuencia de esta será que los individuos serán de un tipo solamente. En el caso multialélico, si no hay mutación se irán perdiendo alelos hasta que quede sólo uno, en es caso se dice que dicho alelo se ha fijado.

En el capítulo que sigue se verá que la velocidad de fijación depende del tamaño de la población, de manera que cuanto más grande es esta, más va a tardar en fijarse algún alelo.

Muchas veces se interpreta que el equilibrio de Hardy-Weinberg implica que las frecuencias se mantendrán estrictamente estables a lo largo del tiempo, como si la ley de los grandes números fuera una fuerza recuperadora de las frecuencias iniciales, lo que no es correcto. Pero a pesar de esta afirmación en la naturaleza se encuentran muchas poblaciones que, si bien no tienen tamanño infinito (es imposible), se encuentran en este equilibrio. Este hecho hace que el uso del modelo sea el de hipótesis nula: si la población no está en equilibrio es porque además de tener tamaño finito no cumple alguna otra de las hipótesis.

Capítulo 3

Deriva genética y coalescente.

En este capítulo se verá la genética de poblaciones desde dos puntos de vista: el primero estará basado en la descendencia de los individuos, o sea, se estudiará el proceso hacia el futuro mientras que el otro se basará en las relaciones ancestrales entre los individuos, o sea, se estudiará el proceso hacia el pasado.

El modelo de Fisher-Wright [6], desrrollado en la década del 60, surge de quitarle la hipótesis de infinitud de la población al modelo de Hardy-Weinberg, lo que provoca la pérdida de equilibrio. Este pérdida se traduce en fluctuaciones alélicas, lo que causará evolución, en el sentido que la población cambiará sus características iniciales a lo largo del tiempo.

En esta misma década comenzaron a observarse las secuencias de proteínas; más tarde se secuanciaron alozimas y cadenas de ADN. Todo esto hace posible que el modelo se pueda aplicar a secuencias que codifican para genes o simplemente a pares de bases, entre otros, y no solamente sobre alelos cuya expresión sea visible.

Dados N individuos, se tienen $2N$ alelos en fase diploide. Los gametos son más numerosos, ya que un mismo individuo produce varios de éstos y puede reproducirse más de una vez, por lo que cada vez que ocurre un evento de reproducción dentro de una generación determinada las opciones posibles para formar un nuevo individuo son siempre las mismas, sin depender de si se trata del primer nacimiento de la generación o del último. Por esta razón es posible considerar que la formación de un individuo ocurre por muestreo con reposición de la generación anterior, o sea, el hecho que un individuo haya heredado un alelo de la generación anterior, no impone ninguna restricción

sobre las posibilidades de herencia de otro individuo; dicho de otra manera, los genotipos de los individuos de una misma generación son eventos independientes entre sí.

3.1. Modelo de Fisher-Wright

En esta sección se trabajará con una población diploide de tamaño constante N con dos alelos A y a y al final del capítulo se verá qué pasa con este mismo modelo cuando ocurren cambios del tamaño poblacionacional.

Uno de los objetivos de esta sección es ver cómo varía la proporción entre los alelos de un tipo y la cantidad total de ellos, $2N$.

Se utilizará individuos o alelos indistintamente.

Se denota:

$X_n =$ número de alelos A en la generación n

La generación $r+1$ es hija de los individuos de la anterior, la generación r. Si hay i individuos portadores del alelo A, la probabilidad de que un individuo de la generación $r+1$ porte también un alelo A es:

$\pi_i := \frac{i}{2N}$.

Por lo tanto la distribución de la variable aleatoria X_{r+1} dada X_r será una Binomial con parámetros $2N$ y $\frac{X_r}{2N}$. Entonces,

$$p_{ij} := P(X_{r+1} = j | X_r = i) = \binom{2N}{j} \pi_i^j (1-\pi_i)^{2N-j} \quad \forall\, 0 \leq i, j \leq 2N \quad (3.1)$$

Observar que el proceso $\{X_r : r = 0, 1, \ldots\}$ es una cadena de Markov homogénea en el tiempo cuya matriz de transición es $P = (p_{ij})$ y el espacio de estados $S = \{0, 1, \ldots, 2N\}$, de los cuales 0 y $2N$ son absorbentes: El estado 0, significa que se fijó el alelo a y el estado $2N$ significa que lo hizo el A.

Se calculará la esperanza, aplicando que la distribución es binomial:

$$\mathbb{E}(X_{r+1} | X_r) = 2N \frac{X_r}{2N} = X_r$$

Observando que el proceso es una martingala, se obtiene que $\mathbb{E}(X_r) = \mathbb{E}(X_0),\ \forall r \in \mathbb{N}$, lo que significa que en promedio la distribución de alelos de todas las generaciones se mantendrá constante. Esto es lo mismo que nos dice el equilibrio Hardy-Weinberg.

Pero el promedio no dice nada respecto a lo que va a pasar en realidad, ya que como se verá más adelante a medida que pasa el tiempo la variabilidad se irá perdiendo y uno de los alelos se fijará.

A continuación se calculará:

$a_i :=$ la probabilidad de que se fije el alelo A, dado que $X_0 := i$.

Se sabe que $a_0 = 0$, $a_{2N} = 1$, y

$$a_i = p_{i0} \cdot 0 + p_{i2N} \cdot 1 + \sum_{j=1}^{2N-1} p_{ij} a_j \tag{3.2}$$

Se tenía que:

$$\sum_{i=0}^{2N} p_{ij} j = \mathbb{E}(X_1 | X_0 = i) = i \tag{3.3}$$

Por lo tanto reescribiendo (3.2) se obtiene:

$$\sum_{j=i}^{2N-1} p_{ij} a_j - a_i = -p_{i2N} \quad \forall i$$

y reescribiendo (3.3) se obtiene:

$$\sum_{j=i}^{2N-1} p_{ij} j - i = -2N p_{i2N} = \sum_{j=i}^{2N-1} p_{ij} \frac{j}{2N} - \frac{i}{2N} = -p_{i2N} \quad \forall i$$

Al reordenar la matriz P, poniendo primero los estados transitivos y luego los absorbentes, adquiere la forma siguiente:

$$\tilde{P} = \left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array} \right)$$

Al considerar todas las ecuaciones con i variando desde 1 hasta $2N-1$, al haber reescrito (3.3) y (3.2) , se observa que éstos son de la forma $(Q-I)x = (p_{1\ 2N}, p_{2\ 2N}, \ldots p_{2N-1\ 2N})$, por lo tanto, (a_i) y $\left(\pi_i = \frac{i}{2N}\right)$ son solución del mismo sistema.

La cadena es absorbente, por lo tanto la matriz $(Q - I)$ es invertible[1], de donde las soluciones de los sistemas son únicas, entonces $a_i = \pi_i,\ \forall i$.

Observación 3.1.1. Como la cadena es absorbente la probabilidad que sea absorbida por algún estado es 1. Pero también se puede calcularla sumando la probablilidad de que se fije cada uno de los alelos en juego: como las a_i son sus frecuencias iniciales, el total tiene que dar 1.

[1]Para ver la prueba que dicha matriz es invertible ver el apéndice

3.1.1. Heterocigosidad

Definición 3.1. *Si dos alelos son idénticos por descendencia, se dice que existe* endogamia.

Observación 3.1.2. Aunque se comience con una población con alelos todos diferentes, los hermanos tendrán los mismos alelos por descendencia, por lo tanto la deriva genética generará endogamia inevitablemente.

Si se eligen dos alelos al azar con reposición, éstos serán idénticos con probabilidad $\frac{1}{N}$ o diferentes con probabilidad $1 - \frac{1}{N}$.

Observación 3.1.3. Hablar de endogamia tiene sentido solamente cuando los individuos son homocigotos, ya que los heterocigotos son diferentes sin importar cómo se formó ese individuo.

Definición 3.2. *Se denominará F, al* coeficiente de endogamia*, definido como la probabilidad que dos individuos sean idénticos por descendencia.*

Bajo las hipótesis del modelo de Fisher-Wright, que es un modelo neutral de evolución, los alelos de un mismo tipo son generados por descendencia. Es importante destacar que dada una población, ésta tendrá endogamia inicial y además los alelos de los individuos homcigotos vendrán del ancestro común de sus padres. Por lo tanto el coeficiente de endogamia y el de homocigosis tendrán el mismo valor, en poblaciones cuyo modelo de evolución sea neutral.

Definición 3.3. *La* heterocigocidad $h(n)$ *se define como* $1 - F$.

Observación 3.1.4. La heterocigocidad es la probabilidad de obtener individuos heterocigotos.

La probabilidad que dos alelos sean sean diferentes por descendencia en la n-ésima generación, muestreando con reposición es $H_n^0 = 2\underbrace{\overbrace{X_n(2N - X_n)}^{casos\ favorables}}_{casos\ totales}{2N(2N-1)}$.

El factor 2 indica que hay dos maneras diferentes posibles de heredar cada uno de los alelos.

Definición 3.4. $h(n) = E(H_n^0)$

Teorema 3.1. $h(n) = \left(1 - \frac{1}{2N}\right)^n \cdot h(0)$

Demostración: Para la prueba se numerarán a las $2N$ "copias"del locus $1, 2, \ldots, 2N$ y se denominarán individuos.

Se elegirán dos individuos $x_1(0)$ y $x_2(0)$ en la generación n. Estos son descendientes de individuos $x_i(1)$, con $i = 1, 2$ de la generación $n-1$, quienes son a su vez descendientes de $x_i(2), i = 1, 2$ de la generación $n-2$, y así sucesivamente. De manera que $x_i(m)$ con $m = 1, 2, \ldots$ describe el linaje de $x_i(0)$, i.e., los ancestros yendo hacia atrás en el tiempo.

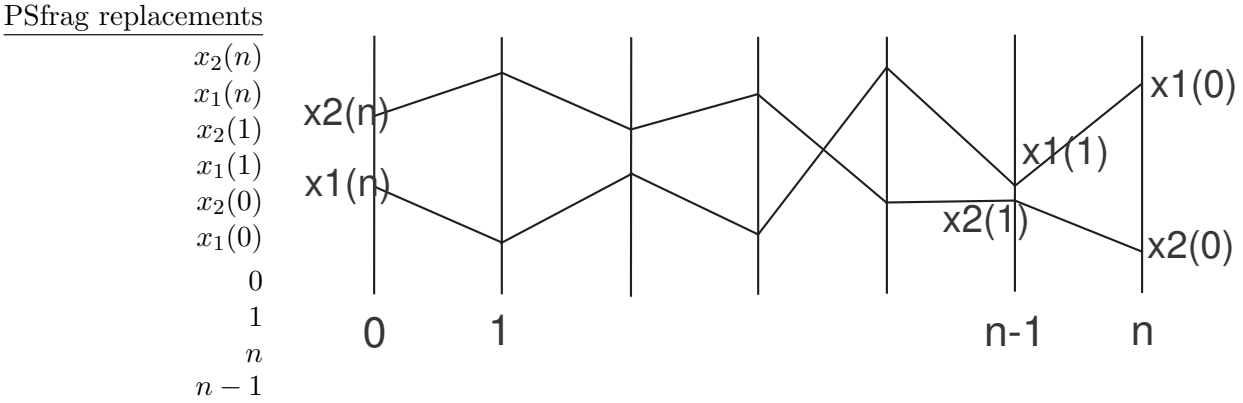


Figura 3.1: La figura ilustra las relaciones ancestrales de dos individuos hacia atrás en el tiempo.

Notar que:

Si $x_1(m) = x_2(m)$, se tendrá que $x_1(l) = x_2(l)\ \forall l/\ m < l \leq n$.

Si $x_1(m) \neq x_2(m)$, entonces como los individuos eligen a sus padres de manera indepen-diente, se cumple que $x_1(m+1) \neq x_2(m+1)$ con probabilidad $1 - \frac{1}{2N}$.

Para que $x_1(n) \neq x_2(n)$, debe haber ocurrido el segundo caso desde $m = 0$ hasta $m = n$, o sea: $x_1(k) \neq x_2(k)\ \forall 0 \leq k \leq n$.

La probabilidad de que $x_1(k) \neq x_2(k)$ dado que $x_1(k-1) \neq x_2(k-1)$ es $\left(1 - \frac{1}{2N}\right)^n$. Además como los individuos $x_1(n)$ y $x_2(n)$ se eligen al azar en la generación 0, entonces serán diferentes con probablidad H_0^0, que es por definición $h(0)$. ♠

Observación 3.1.5. Si N es grande, entonces $h(n) \approx e^{-\frac{n}{2N}} h(0)$. Por lo tanto la heterocigosidad tiende a 0 exponencialmente cuando $\frac{n}{2N} \to +\infty$. Por lo tanto, cuanto más grande es la población habrá menor variabilidad,

de dónde se deduce que le probabilidad que un alelo se fije disminuye a medida que el tamaño poblacional aumenta.

3.2. Coalescente

En esta sección se estudiará el modelo retrospectivamente, ya que naturalmente el proceso genealógico proporciona una manera directa de estudiar las frecuencias genéticas.

En ausencia de recombinación y mutación la secuencia de ADN que representa al gen de interés es una copia exacta de una secuencia de la generación anterior; sucesivamente cada una de las secuencias será una copia de una de la generación previa.

Para desarrollar el modelo del coalescente se asume:

- El número de nacimientos es independiente entre generaciones. La única restricción es que el total de individuos nacidos por generación sea N (es decir que se consideran $2N$ copias de ADN).
- Se numerarán arbitrariamente los individuos de 1 a $2N$ en cada generación, de manera que aunque dos secuencias provengan de la misma célula, no necesariamente sus números asignados guardarán alguna correlación.

Notar que si ν_i es el número de hijos del $i-ésimo$ individuo, entonces:

$$\mathbb{P}\left(\nu_1 = m_1, \ldots, \nu_{2N} = m_{2N}\right) = \frac{(2N)!}{m_1! \ldots m_{2N}!} \left(\frac{1}{2N}\right)^{2N} \tag{3.4}$$

con la restricción $\quad m_1 + \ldots + m_{2N} = 2N$

Notar entonces que, $(\nu_1, \ldots, \nu_{2N})$ tiene disribución multinomial simétrica.

Observación 3.2.1.

1. Comparando con el modelo de deriva genética de Fisher-Wright cada individuo porta un alelo A o a. Si los $2N$ individuos portaran el alelo A, como no hay mutaciones, toda la generación siguiente obtendrá alelos de ese tipo solamente.
2. Ordenando a los individuos según el alelo que porta, poniendo a los que llevan el A primero y los que llevan a después se obtiene que la distribución de los nacimientos tipo A es $\nu_1 + \ldots + \nu_i$, recuperando la distribución binomial con parámetro $2N$ y probabilidad de éxito $\frac{1}{2N}$, que se tenía anteriormente.

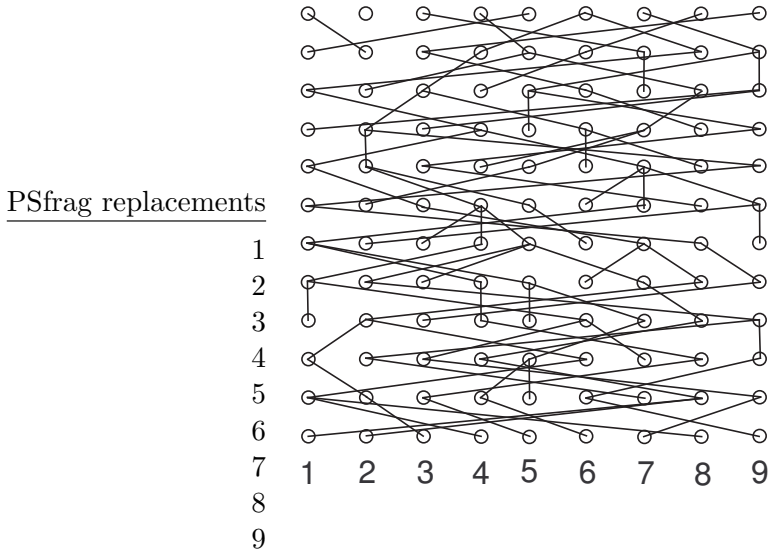


Figura 3.2: Simulación de las relaciones ancestrales durante 12 generaciones bajo las hipótesis del modelo Fisher-Wright para el caso en que la población tiene 9 individuos haploides. Las líneas indican las realciones padre-hijo

Notación: Como es habitual, en lo que sigue se denotará por MRCA al Ancestro Común Más Reciente y por T_{MRCA} al número de generaciones hasta llegar al MRCA de un grupo de individuos.

Las figuras 3.2, 3.3 y 3.4 muestran que los individuos 3 y 4 tienen su MRCA 2 generaciones atrás, mientras que el 2 y el 3 tienen su MRCA 11 generaciones atrás.

3.2.1. Tiempo al MRCA.

El objetivo de lo que sigue es estudiar la distribución de la variable aleatoria T_{MRCA} para un grupo de k individuos dentro de una población de tamaño $2N$.

Estudiando el proceso desde el presente hacia el pasado, los individuos eligen a sus padres al azar y de manera independiente; las elecciones sucesivas serán al azar de generación en generación.

En el capítulo anterior ya se vio que dados dos individuos la probabilidad que tengan padres diferentes es:

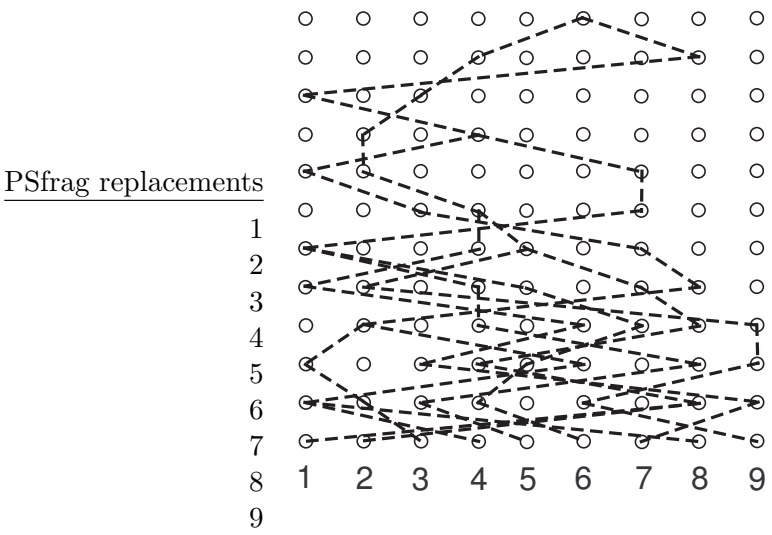


Figura 3.3: Modelo de la figura 3.2. Están marcadas solamente las relaciones que van desde el MRCA de la última generación hasta dicha generación.

$$\mathbb{P}(Dos\ individuos\ tengan\ padres\ diferentes) = 1 - \frac{1}{2N}$$

Si se elige un grupo de k individuos, la probabilidad que 2 de ellos elijan el mismo padre es $\underbrace{\frac{k(k-1)}{2}}_{C_2^k} \frac{1}{2N}$.

Observación 3.2.2. En todo lo que sigue se trabajará con $k \ll 2N$. Las probabilidades de que 3 o más individuos tengan el mismo padre se despreciarán, ya que el orden es menor o igual que $\frac{1}{(2N)^2}$.

Usando el mismo razonamiento que antes, se obtiene que la probabilidad de que no haya eventos de coalescencia en las primeras n generaciones dependiendo de la cantidad de individuos es:

Para dos individuos: $\left(1 - \frac{1}{2N}\right)^n$

Para k individuos:

$$\begin{aligned}[1 - \mathbb{P}(al\ menos\ 2\ tengan\ el\ mismo\ padre)]^n &\approx \left(1 - \frac{k(k-1)}{2}\frac{1}{2N}\right)^n \\ &\approx e^{-\frac{k(k-1)}{2}\frac{n}{2N}}\end{aligned}$$

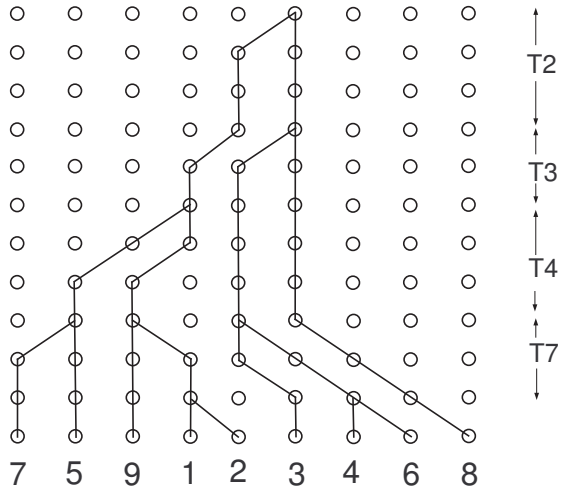


Figura 3.4: Modelo de la figura 3.2. Las relaciones de ancestralidad son las mismas que en la figura 3.3, pero cada generación está reordenada para que queden claras las relaciones entre los individuos de las mismas.

Notación: En lo que sigue se denotará:

$T_2 = \frac{T_{MRCA}}{2N}$; esto es, el tiempo al MRCA de dos individuos individuos medido en unidades de $2N$ generaciones.

$t = \frac{n}{2N}$.

T_j al tiempo durante el cual hay exactamente j individuos por generación de los k que habían inicialmante.

Con esto queda[2]:

$$\mathbb{P}(T_2 > t) = e^{-t}$$

$$\mathbb{P}(T_k > t) = e^{-\frac{k(k-1)}{2}t}$$

Observación 3.2.3. Si $N \to \infty$, entonces se puede considerar que t es continuo.

[2] T_k se puede ver como el mínimo de $\frac{k(k-1)}{2}$ variables exponenciales de parámetro 1

Al tomar el tiempo continuo se pasa de una variable geométrica a una exponencial. Por lo tanto para calcular la esperanza de las variables aleatorias basta observar que se cumple que si $\mathbb{P}(T_k > t) = e^{-\frac{k(k-1)}{2}t}$, su esperanza es el inverso del parámetro de la exponencial. Por lo tanto:

$E(T_2) = 1$

$E(T_k) = \frac{2}{k(k-1)}$.

Observación 3.2.4. Si se mide el tiempo en unidades de $2N$ generaciones, el tiempo durante el cual hay exactamente k linajes tiene una distribución exponencial de parámetro $\frac{2}{k(k-1)}$, cuya esperanza es $\frac{2}{k(k-1)}$.

Definición 3.5. *W_n será el tiempo que hay que esperar hasta llegar al MRCA medido de a $2N$ generaciones:* $W_n = \frac{T_{MRCA}}{2N}$

Observación 3.2.5. $W_n = T_n + \ldots + T_2$, de donde

$$W_n = \sum_{k=2}^{n} T_k \tag{3.5}$$

$$\mathbb{E}(W_n) = \mathbb{E}\left(\sum_{k=2}^{n} T_k\right) = \sum_{k=2}^{n} \mathbb{E}(T_k) = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right)$$

$$\overset{telescópica}{=} 2\left(-\frac{1}{n} + \frac{1}{2-1}\right) = 2\left(1 - \frac{1}{n}\right)$$

.

Por lo tanto, el tiempo que habrá que esperar en promedio para encontrar al MRCA:

$$\mathbb{E}(W_n) \overset{n\to\infty}{\to} 2$$

De ese tiempo, la mitad aproximadamente es el que se debe esperar para que ocurra la última colición ($\mathbb{E}(W_n) = 1$).

Recordar que el tiempo está medido de a $2N$ generaciones, por lo tanto, en promedio, habrá que esperar aproximadamente $4N$ generaciones para encontrar al MRCA.

Observación 3.2.6. $1 = \mathbb{E}(T_2) = \mathbb{E}(\sum_{k=3}^{n} T_k) \leq \mathbb{E}(W_n) < 2$ y W_n está cerca de 2, incluso para tamaños poblacionales moderados.

Observación 3.2.7. $\mathbb{E}(W_n - \widetilde{W_n}) = 2\left[1 - \frac{1}{2N} - \left(1 - \frac{1}{n}\right)\right] = 2\left(\frac{1}{n} - \frac{1}{2N}\right) < \frac{2}{n}$. Donde la población inicial para calcular W_n tiene $2N$ individuos y para calcular $\widetilde{W_n}$ tiene n.

La diferencia entre la esperanza del tiempo que puede llevar encontrar el MRCA de un subgrupo de la población entera y ella misma es pequeña: $\frac{2}{n}$, donde n es el tamaño poblacional del subgrupo. Entonces para saber aproximadamente hace cuánto tiempo vivió el MRCA, no es necesario tener en cuenta a la población entera, sino a un subconjunto de la misma. El tamaño de éste va a depender de la cota de error que se desee tener $\left(\frac{2}{n}\right)$.

Este resultado es de gran utilidad para quienes, por ejemplo se ocupan de estudiar las relaciones filogenéticas entre especies, ya que sin tener a todos los individuos de una población, les es posible estimar cuánto tiempo antes vivió su ancestro común, comapararlo con los datos obtenidos con otros métodos y darles consistencia, entre otras aplicaciones.

3.3. Tamaño de población variable

En esta sección se quitará otra hipótesis: el tamaño constante.

Como las generaciones siguen siendo discretas se tomará como 0 a la generación presente y se denotará $2N(j)$ al número de individuos j generaciones atrás.

Todos los tamaños, independientemente del número de generación, continuarán siendo relativamente grandes y se reescalará el tiempo en unidades de $2N \equiv 2N(0)$ generaciones.

Definición 3.6. *Tamaño relativo:*

$$f_{2N}(x) = \frac{2N(\lceil 2Nx \rceil)}{2N} = \frac{2N(j)}{2N}, \quad \frac{j-1}{2N} < x \leq \frac{j}{2N}, \quad j = 1, 2, \ldots \tag{3.6}$$

Se trabajará con tamaños relativos tales que:

$$\lim_{N \to \infty} f_{2N}(x) = f(x) \tag{3.7}$$

existe y es estrictamente positiva $\forall x \geq 0$.

En el modelo de reproducción de Fisher-Wright, se vio que la probabilidad de que dos individuos tuvieran su ancestro común más de s generaciones atrás, era la de elegir padres diferentes cada vez, siendo estos sucesos independientes. Por lo tanto, se tiene que:

$$\mathbb{P}\left(T_{MRCA} > s\right) = \prod_{j=1}^{s}\left(1 - \frac{1}{2N(j)}\right) = \prod_{j=1}^{s} e^{\log\left(1-\frac{1}{2N(j)}\right)} \approx e^{\sum_{j=1}^{s} -\frac{1}{2N(j)}}$$

$$\sum_{j=1}^{s} \frac{1}{2N(j)} = \sum_{j=1}^{s} \frac{2N}{2N(j)}\frac{1}{2N} = \int_0^{\frac{s}{2N}} \frac{dx}{f_{2N}(x)} = \int_0^t \frac{dx}{f_{2N}(x)} \overset{N\to\infty}{\to} \int_0^t \frac{dx}{f(x)}$$

$$\Rightarrow \mathbb{P}\left(T_{MRCA} > s\right) \to e^{-\int_0^{\frac{s}{2N}} \lambda(x)dx} \quad con\ \lambda(x) = \frac{1}{f_{2N}(x)}$$

Reescalando el tiempo t reescalado a $2N$ generaciones $\left(t = \frac{s}{2N}\right)$ se obtiene:

$$\mathbb{P}\left(\frac{T_{MRCA}}{2N} > \frac{s}{2N}\right) \to e^{-\int_0^{\frac{s}{2N}} \lambda(x)dx} \quad con\ \lambda(x) = \frac{1}{f_{2N}(x)}$$

$$\mathbb{P}\left(W_n > t\right) \to e^{-\int_0^{t} \lambda(x)dx} \quad con\ \lambda(x) = \frac{1}{f_{2N}(x)} \tag{3.8}$$

3.3.1. Crecimiento exponencial

Sea una población, tal que

$$N(x) = \begin{cases} N_0 e^{-\rho x} & x \le x_0 \\ 1 & x > x_0 \end{cases}$$

Donde $x_0 = \rho^{-1} \ln N_0$ para que $N(x_0) = 1$, ya que no puede haber menos de un invividuo en una generación. El valor 0 correspondería a la no existencia de la especie.

$$\Rightarrow f_{2N}(x) = \frac{2N(\lceil 2Nx \rceil)}{2N} = e^{-\rho(\lceil 2Nx \rceil)}$$

$$\int_0^t \lambda(x)dx = \int_0^t \frac{1}{e^{-\rho x}}dx = \int_0^t e^{\rho x}dx = \frac{e^{\rho t} - 1}{\rho}$$

$$\Rightarrow \mathbb{P}(W_n > t) = e^{-\frac{e^{\rho t}-1}{\rho}}$$

Ejemplo 3.1. Harding et al. [5] para estimar el tiempo hasta el MRCA de la β-globina humana utilizaron una población con $N_0 = 18,807$ y $\rho = 1{,}861 \times 10^{-5}$. El tiempo de coalescencia de la población se estimó gráficamente.

En la gráfica se expresó el tiempo negativamente, de manera que al ir hacia atrás en el tiempo en la gráfica se hace lo mismo, lo que facilita su interpretación.

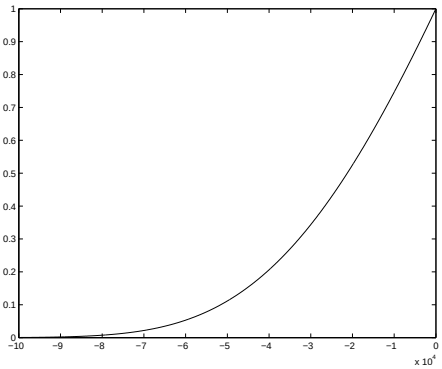


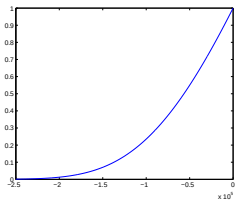
Figura 3.5: $\mathbb{P}(W_n > t) = exp(-\frac{e^{1,861\times10^{-5}t}-1}{1,861\times10^{-5}})$ con $t = \frac{s}{18807}$

En la gráfica se observa que la probabilidad que la coalescencia haya ocurrido más de 85.000 unidades de tiempo atrás es casi 0.

Es natural suponer que si una población creció exponencialemente se ramificó en forma de estrella, pero en este caso la gráfica contradice dicha suposición, ya que, en dicho caso, el cambio en las probabilidades debera ser brusco.

Al graficar utilizando diferentes parámetros se observa que la forma de la curva depende de $2N_0\rho$ sin importar los valores que toma N_0. De lo que se deduce que la forma del árbol genealógico de la población también depende de dicho coeficiente.

A continuación se encuentran tres gráficos más. El primero tiene $2N_0\rho = 0,7$, al igual que la gráfica anterior pero con $N_0 = 50000$.



Las dos gráficas siguientes se hicieron tomando $2\rho N_0 = 10$ y los tamaños poblacionales del orden de las anteriores.

3.3.2. Tamaño de población efectivo

Definición 3.7. *El* tamaño de población efectivo, N_e, *es el número de individuos que debería tener una población de tamaño constante para que el*

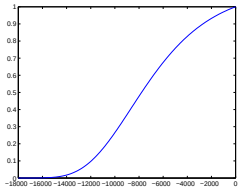


Figura 3.6: $N_0 = 18807$

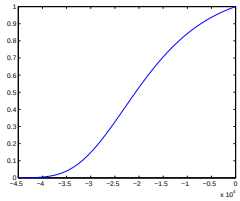


Figura 3.7: $N_0 = 50000$

estadístico con el que se trabaja dé el mismo resultado que daría para una población con cantidad variable de individuos.

Los estadísticos con los cuales se trabajará son:

- la probabilidad π_2 que dos genes elegidos al azar sean descendientes del mismo gen parental.
- El tiempo esperado de colaescencia ET_2.

Existen otros estadísticos que se pueden medir como el mayor valor propio menor que uno de la matriz de transición del proceso de Fisher-Wright o la varianza.

Para los estadísticos anteriores se definen los siguientes tamaños efectivos:

N_e^i = tamaño de población efectivo de endogamia = $(2\pi_2)^{-1}$

N_e^c = tamaño de población efectivo de coalescencia = $\frac{ET_2}{2}$

Variación periódica del tamaño poblacional

Dada una población cuya cantidad de individuos cumple un ciclo de k generaciones: $N_1,\ N_2,\ \ldots,\ N_k,\ N_1, \ldots, N_k, \ldots$. Aplicando la defininición 3.7 se puede generalizar la fórmula anterior de la siguiente manera:

$$\prod_{s=1}^{k}\left(1-\frac{1}{2N(s)}\right)=\left(1-\frac{1}{2N_e^i}\right)^k$$

Si los tamaños poblacionales son grandes, lo anterior implica:

$$\sum_{i=1}^{k}\frac{1}{2N(s)}\approx\frac{k}{2N_e^i}$$

Lo que implica que en este caso el tamaño poblacional efectivo es la media armónica:

$$N_e^i\approx k/\sum_{s=1}^{k}\frac{1}{N(s)} \tag{3.9}$$

Ejemplo 3.2. Dada una población cuyo tamaño poblacional aumenta de 10 a 100 de una generación a la siguiente, luego a 1000 y vuelve a 10, así sucesivamente.

De la ecuación 3.9 se tiene que el tamaño efectivo es aproximadamente:

$$N_e^i=3/\left(\frac{1}{10}+\frac{1}{100}+\frac{1}{1000}\right)=\frac{3000}{111}=27{,}0270\approx 27$$

Ahora se hará el cálculo directamente:

$$\left(1-\frac{1}{10}\right)\left(1-\frac{1}{100}\right)\left(1-\frac{1}{1000}\right)=\left(1-\frac{1}{2N}\right)=\left(1-\frac{1}{2N_e^i}\right)^3,$$

de donde $N_e^i\approx 26,656\approx 27$

Si bien los tamaños poblacionales no son muy grandes, se observa que los resultados obtenidos son similares.

Para calcular N_e^c basta observar que por ciclo la probabilidad de coalescencia es la misma, por lo tanto, en este caso $N_e^c=N_e^i$

A continuación se verán ejemplos de diferentes tipos de variación poblacional.

Modelo para dos sexos

Dada una población que está compuesta por N_1 machos diploides y N_2 hembras diploides, cada individuo nacido heredará un alelo materno, tomado al azar de los $2N_2$ disponibles y un alelo paterno tomado de los $2N_1$.

Para calcular el N_e^i se debe calcular la probabilidad que dos alelos tomados al azar de los $2N$ disponibles provengan de un mismo individuo.

Suponiendo que nacieron N individuos y tomando en cuenta los dos casos posibles de herencia del gen se tiene:

$$\begin{aligned}\pi_2 = \; & \mathbb{P}(\textit{los genes elegidos son ambos paternos y del mismo padre}) \\ & +\mathbb{P}(\textit{los genes elegidos son ambos maternos y de la misma madre}) \\ = \; & \tfrac{N(N-1)}{2N(2N-1)}\left\{\tfrac{1}{2N_1}+\tfrac{1}{2N_2}\right\}\end{aligned}$$

$$\Rightarrow \lim_{N\to\infty} \frac{N(N-1)}{2N(2N-1)} = \frac{1}{4}$$

Nota: La intersección de los eventos anteriores es despreciable.

De manera que, si la pobalción que nació es grande, no importa su tamaño exacto, sino el tamaño de la pobalción parental.

$$N_e^i = (2\pi)^{-1} \approx \frac{2}{(2N_1)^{-1}+(2N_2)^{-1}} = \frac{4N_1N_2}{N_1+N_2} \tag{3.10}$$

Ejemplo 3.3. Cuando se crían bovinos se tiene en general mucho más vacas que toros, ya que estos pueden fecundar a muchas vacas, lo que genera una gran endogamia. Para confirmar esta hipótesis se calculará el N_e^i suponiendo que se tienen 10 toros y 1000 vacas utilizando la ecuación 3.10.

$$\Rightarrow N_e^i = \frac{4\cdot 10\cdot 1000}{1010} = 39{,}6 \approx 40$$

Esto quiere decir que respecto a la endogamia, la población se comporta como si tuviera 40 individuos solamente. Esta es una de las razones por las cuales es difícil de eliminar las enfermedades de transmisión genética sin sacar a los individuos problema del ciclo reproductivo.

Observación 3.3.1. En general, si $N_1 \ll N_2 \Rightarrow N_e^i \approx 4N_1$

Capítulo 4

Mutación

En este capítulo se verá qué pasa con los modelos evolutivos cuando la mutación está presente en los mecanismos reproductivos. Los mecanismos posibles de mutación son diferentes, lo que da lugar a la existencia de diferentes modelos. De los casos que se presentarán en este trabajo, en algunos modelos los alelos mutarán a otro ya presente en la población, mientras que en otros, cada vez que ocurra una mutación, se introducirá un alelo diferente. No sólo se estudiará qué pasa en un modelo de tipo de Fisher-Wright, sino que además se desarrollará un modelo de tiempo continuo.

La presencia de mutación es un cambio muy importante, porque permitirá evolución sostenida en el tiempo, ya que ésta es capaz de contrarrestar la deriva genética. En los modelos anteriores la deriva tendía a fijar un único alelo. Ahora la mutación es una fuente inagotable de variación, por lo que cada vez que se fije un alelo, hay posibilidades de que surja uno nuevo. Este hecho no quita que se pueda llegar a un equilibrio entre deriva y mutación.

4.1. Mutación bajo las hipótesis del modelo de Fisher-Wright

Se considera una población con 2 alelos, A y a, tales que $\mathbb{P}(A \text{ mute a } a) = v$ y $\mathbb{P}(a \text{ mute a } A) = u$. Debido a la presencia de mutación, la manera que un individuo de la nueva generación porte al alelo A ya no es únicamente por herencia, sino que puede haber heredado un a y éste haber mutado a A; por lo tanto las probabilidades de transición, p_{ij}, que se habían calculado en el capítulo anterior (3.1) deben modificarse.

$$p_{ij} = \binom{2N}{j} \pi_i^j (1-\pi_i)^{2N-j}$$

$$\text{con } \pi_i = \underbrace{\frac{i}{2N}(1-v)}_{(1)} + \underbrace{\frac{2N-i}{2N}u}_{(2)} \quad \forall\ 0 \leq i,j \leq 2N$$

(1) es la probabilidad de heredar un A y que no mute a un a.

(2) es la probabilidad de heredar un a y que mute a A.

Observación 4.1.1. $p_{ij} > 0\ \forall i,j$, de donde la cadena de Markov es irreducible y aperiódica, y como el espacio de estados es finito, tiene una única distribución estacionaria ν. Además:

$$\lim_{n\to\infty} \mathbb{P}(X_n = i) \to \nu(i)\ \forall i$$

Para calcular la distribución se debe resolver el sistema de ecuaciones $\sum_i \nu_i p_{ij} = \nu_j$ con $\nu_i > 0$ y $\sum_i \nu_i = 1$.

Se define

$$X_\infty := \lim_{n\to\infty} X_n$$

Teorema 4.1.

$$\mathbb{E}(X_\infty) = 2N\frac{u}{u+v} = 2N\rho \tag{4.1}$$

Demostración de 4.1:

$$\mathbb{E}(X_n|X_{n-1}) = 2N\pi_{n-1} = 2N\left(\frac{X_{n-1}}{2N}(1-v) + \frac{2N-X_{n-1}}{2N}u\right),$$

por lo tanto tomando esperanza:

$$\mathbb{E}(X_n) = \mathbb{E}(2N\pi_{n-1}) = (1-v)\mathbb{E}(X_{n-1}) + u\,(2N - \mathbb{E}(X_{n-1}))$$

En el límite: $\mathbb{E}(X_n) = \mathbb{E}(X_{n-1}) = x$

$$\Rightarrow x = (1-v)x + u(2N-x)$$

$$\Rightarrow x = 2N\frac{u}{u+v}$$

Para ver que $\mathbb{E}(X_n)$ converge a este límite sustituyendo x en la ecuación anterior y restándosela a la primera, se obtiene:

$$\mathbb{E}(X_n - 2N\rho) = (1-u-v)\mathbb{E}(X_{n-1}),$$

iterando esta igualdad se tiene que

$$\mathbb{E}(X_n - 2N\rho) \approx (1-u-v)^n \mathbb{E}(X_0),$$

como $0 < u+v < 1 \Rightarrow 0 < |1-u-v| < 1$

$$\Rightarrow \lim_{n\to\infty} \mathbb{E}(X_n - 2N\rho) = 0$$
$$\Rightarrow \lim_{n\to\infty} \mathbb{E}(X_n) = 2N\rho \qquad \spadesuit$$

Definición 4.1. *Se dice que dos individuos son* idénticos por descendencia *si hay un evento de coalscencia entre sus linajes antes que una mutación afecte a alguno de ellos.*

Notación: En lo que sigue μ denota la probabilidad de que un individuo mute en una generación.

Teorema 4.2. *Dada una población que cumple las hipótesis bajo las cuales se trabajó en este capítulo:*

$$\mathbb{P}(\textit{2 individuos sean idénticos por descendencia}) \approx \frac{\frac{1}{2N}}{2\mu + \frac{1}{2N}} = \frac{1}{1+4N\mu} \tag{4.2}$$

Demostración: En el paso de una generación a la siguiente los eventos posibles son: que uno de los linajes mute (esto tiene probabilidad $p_1 = 2\mu$), que los linajes coalescan (probabilidad $p_2 = \frac{1}{2N}$) o que no pase ni una cosa ni la otra. Está claro (ver figura) que la probabilidad γ de que haya un evento de mutación antes de un evento de coalescencia cumple la siguiente ecuación:

$$\gamma = \underbrace{p_1}_{mute} + \underbrace{(1-p_2)}_{no\ mute}\ \underbrace{(1-p_1)}_{no\ coalesca}\ \gamma$$

Como la probablidad de que coalescan y muten a la misma vez es despreciable por ser de orden $\frac{\mu}{2N}$, se obtiene:

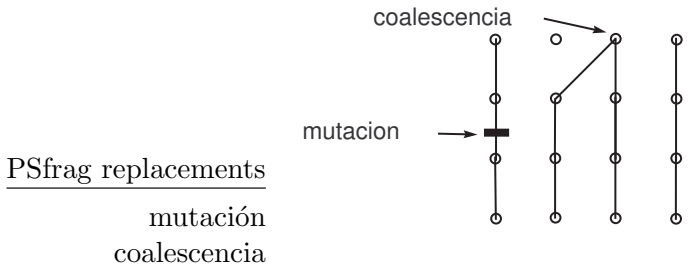


Figura 4.1: La figura ilustra un ejemplo de relaciones ancestrales entre 4 generaciones. La línea horizontal indica la presencia de una mutación.

$$\gamma = p_1 + (1 - p_2 - p_1)\gamma$$

de donde

$$\gamma = \frac{p_1}{p_1 + p_2}.$$

Por lo tanto la probabilidad de que dos individuos sean idénticos por descendencia es:

$$1 - \gamma = \frac{p_2}{p_1 + p_2} = \frac{1}{1 + 4N\mu}$$

♠

A continuación se calculará la varianza de X_∞.

Teorema 4.3.

$$var(X_\infty) = \left(2N + \frac{2N(2N-1)}{1 + 4N(u+v)}\right) \frac{uv}{(u+v)^2} \tag{4.3}$$

Demostración Para calcular EX_∞^2, se reescribirá X_∞ de la siguiente manera:

$X_\infty = \sum_{i=1}^{2N} \eta_i$, donde $\eta_i = 1$ si el i-ésimo individuo porta el alelo A y 0 si porta el alelo a. Por lo tanto:

$$X_\infty^2 = \sum_{i=1}^{2N} \sum_{j=1}^{2N} \eta_i \eta_j$$

Separando los $2N$ términos tales que $i = j$ de los $2N(2N-1)$ restantes se obtiene:

$$\mathbb{E}(X_\infty^2) = 2N\mathbb{P}(\eta_1 = 1) + 2N(2N-1)\mathbb{P}(\eta_1 = 1, \eta_2 = 1)$$

De (4.1) $\mathbb{P}(\eta_1 = 1) = \frac{u}{u+v}$. Considerando las posibilidades de coalescencia o no y usando (4.2) con $\mu = u + v$, se obtiene:

$$\mathbb{P}(\eta_1 = 1, \eta_2 = 1) = \underbrace{\frac{1}{1+4N\mu}}_{id.\ por\ desc.} \underbrace{\frac{u}{u+v}}_{\mathbb{P}(\eta_1=1)} + \left(1 - \frac{1}{1+4N\mu}\right)\left(\frac{u}{u+v}\right)^2 =$$

$$\frac{1}{1+4N\mu}\frac{u}{u+v} + \frac{4N\mu}{1+4N\mu}\left(\frac{u}{u+v}\right)^2$$

Reescribiendo $(EX_\infty)^2$ se obtiene:

$$(EX_\infty)^2 = \left(2N\frac{u}{u+v}\right)^2 = (2N + 2N(2N-1))\left(\frac{u}{u+v}\right)^2 =$$

$$2N\left(\frac{u}{u+v}\right)^2 + 2N(2N-1)\left\{\frac{1}{1+4N\mu} + \frac{4N\mu}{1+4N\mu}\right\}\left(\frac{u}{u+v}\right)^2$$

Haciendo $\mathbb{E}(X_\infty^2) - (EX_\infty)^2$ se obtiene el resultado deseado.

♠

4.2. Modelo de Moran

Para desarrollar el modelo Fisher-Wright, una de las hipótesis imprescindibles es que las generaciones que no se solapan; esto es lo que sucede, por ejemplo, en algunas plantas. Pero existen muchas otras especies en las que esto no ocurre. Para esas especies el modelo estudiado en la sección anterior no se puede aplicar; es conveniente en estos casos aplicar un modelo de tiempo continuo. El modelo de Moran se caracteriza por permitir que un individuo, pero solamente uno, mute en cualquier momento.

Es importante destacar que en el modelo de Fisher-Wright se consideran N individuos y $2N$ copias de un gen, o $2N$ individuos haploides. En este nuevo contexto se considerarúque todos son haploides. De todas maneras la

cantidad de individuos será $2N$ para poder comparar los resutlados obtenidos entre los diferentes modelos más fácilmente.

El modelo se caracteriza de la siguiente manera:

(i) Cada individuo es reemplazado con tasa 1. O sea, cada individuo x vive un tiempo cuya distribución es exponencial con esperanza 1 y luego es "‘reemplazado".

(ii) Para reemplazar un individuo x se elige al azar un indviduo de un conjunto que incluye al mismo x.

(iii) Un individuo a que es elegido muta a A con probabilidad u y un individuo A muta a a con probabilidad v.

(iv) El nuevo individuo, posiblemente mutado, reemplazará al individuo x.

La probabilidad p_i de reemplazar al individuo seleccionado por una A, dado que en el momento de la selección de ese individuo hay i individuos A es:

$$p_i = \frac{i}{2N}(1-v) + \frac{2N-i}{2N}u$$

Por lo tanto las posibilidades son:
ganar una A: $\quad i \to i+1 \quad$ con tasa de transición $\quad b_i = \frac{2N-i}{2N}p_i$

perder una A: $\quad i \to i-1 \quad$ con tasa de transición $\quad d_i = \frac{i}{2N}(1-p_i)$

A partir de las ecuaciones anteriores se deduce que la única manera de ganar una A es eligiendo una a y que ésta mute; de la misma manera, para perder una hay que elegir un A y que mute a a.

Para calcular la probabilidad $\pi(i)$, de que haya i individuos A en equilibrio, las transiciones $i \to i-1$ e $i-1 \to i$ deben ocurrir en la misma medida, por lo tanto:

$$\pi(i)d_i = \pi(i-1)b_{i-1}$$

Despejando $\pi(i) = \pi(i-1)\frac{b_{i-1}}{d_i}$ e iterando hasta el momento $k < i$, se obtiene:

$$\pi(i) = \pi(k) \prod_{j=k+1}^{i} \frac{b_{j-1}}{d_j} \tag{4.4}$$

Esto implica que si se tiene $\pi(k)$, para algún k, es posible calcular todas las probabilidades siguientes. Finalmente usando que la suma de todas las probabilidades es 1, quedan todos los parámetros determinados.

El siguiente ejemplo servirá como motivación para introducir el concepto de homocigosis esperada.

Ejemplo 4.1. Cuello de Botella

Dada una población se dice que en su historia demográfica hubo un cuello de botella si la misma tenía tamaño constante, repentinamente se redujo a pocos individuos y luego creció lentamente a un posible nuevo estado de equilibrio.

Una manera de que esto ocurra es que una hembra fecundada de una población grande emigre hacia un nuevo lugar geográfico y forme allí una nueva colonia. Ésta crecerá en un entorno, que posiblemente será diferente del original. Se cree que esto fue lo que pasó repetidamente en el proceso evolutivo de la Drosóphila Hawaiana.

Es de esperar que este proceso genere una gran endogamia [1]. Para poder medir la endogamia se estudiará la homocigosis esperada.

Definición 4.2. $J_t :=$ Homocigosis *esperada en la generación* t.

Para calcular la homocigosis esperada se calculará primero la probabilidad de que un individuo sea homocigoto y a partir de esta probabilidad se calculará el valor esperado. Notar que esta definición sólo tiene sentido si los individuos estudiados son diploides.

Notación: Se introducen las siguientes variables y sucesos.

X_t es la variable aleatoria que toma el valor 1 si los alelos son los mismos por descendencia y 0 si no lo son.

ν es la tasa de mutación. Como se considera un modelo de tiempo discreto, es la probabilidad de que el gen en estudio mute en un alelo, en una generación.

$A=$ {dos alelos seleccionados tienen el mismo ancestro} $\Rightarrow \mathbb{P}(A) = \frac{1}{2N}$

$B=$ {dos alelos seleccionados mutan} $\Rightarrow \mathbb{P}(B) = (1-\nu)^2$

[1] Para estudiar los efectos de los cuellos de botella sobre la variabilidad genética de una población Durret siguió a Nei, Maruyama y Chakraborty (1975).

$C=$ {los ancestros de los alelos considerados son idénticos descendencia} $\Rightarrow$
$\mathbb{P}(C) = \mathbb{P}(X_{t-1} = 1)$

Con esto entonces resulta que la probabilidad de que un individuo sea homocigoto es:
$\mathbb{P}((A \cap B) \cup (A^C \cap B \cap C))$
Como los sucesos son independientes y la intersección de los conjuntos de la unión es nula, debido a que $A \cap A^C = \emptyset$, se obtiene:

$$\begin{aligned} \mathbb{P}(X_t = 1) = & \quad \mathbb{P}(A)\mathbb{P}(B) + (1 - \mathbb{P}(A))\mathbb{P}(B)\mathbb{P}(C) \\ = & \quad \tfrac{1}{2N}(1-\nu)^2 + \left(1 - \tfrac{1}{2N}\right)(1-\nu)^2\mathbb{P}(X_{t-1} = 1) \end{aligned}$$

Por lo tanto:

$$\mathbb{E}(X_t|X_{t-1}) = (1-\nu)^2\frac{1}{2N} + (1-\nu)^2\left(1 - \frac{1}{2N}\right)X_{t-1}$$

$$J_t = \mathbb{E}(X_t) = \mathbb{E}(\mathbb{E}(X_t|X_{t-1})) = (1-\nu)^2\frac{1}{2N} + (1-\nu)^2\left(1 - \frac{1}{2N}\right)\mathbb{E}(X_{t-1})$$

$$\Rightarrow J_t = (1-\nu)^2\left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)J_{t-1}\right) \tag{4.5}$$

Iterando esta relación y reagrupando se obtiene:

$$J_t = \sum_{j=1}^{t} \frac{(1-\nu)^{2j}}{2N_{t-j}} \prod_{i=1}^{j-1}\left(1 - \frac{1}{2N_{t_i}}\right) + J_0(1-\nu)^{2t}\prod_{i=1}^{t}\left(1 - \frac{1}{2N_{t-i}}\right)$$

Por convención se toma $\prod_{i=1}^{0} z_i = 1$. Como ν es chico y las N_r son grandes, se puede utilizar la equivalencia $1 - x \sim e^{-x}$, si x está cerca de 0.

$$J_t = \sum_{j=1}^{t} \frac{1}{2N_{t-j}} \exp\left(-2\nu j - \sum_{i=1}^{j-1}\frac{1}{2N_{t-i}}\right) + J_0 \exp\left(-2\nu t - \sum_{i=1}^{t}\frac{1}{2N_{t-i}}\right)$$

Haciendo los cambios de variables $k = t - j$ y $l = t - i$ se obtiene:

$$J_t = \sum_{k=0}^{t-1} \frac{1}{2N_k} \exp\left(-2\nu(t-k) - \sum_{l=k+1}^{t-1}\frac{1}{2N_l}\right) + J_0 \exp\left(-2\nu t - \sum_{l=0}^{t-1}\frac{1}{N_l}\right)$$

Tomando las sumatorias como sumas de Riemann:

$$J_t = \int_0^t \frac{1}{2N_s} \exp\left(-2\nu(t-s) - \int_s^t \frac{1}{2N_u} du\right) ds + J_0 \exp\left(-2\nu t - \int_0^t \frac{1}{N_s} ds\right) \tag{4.6}$$

De esta ecuación se deduce que la homocigosis esperada, J_t converge exponencialmete a su valor límite.

Esta fórmula también sirve para calcular la homocigosis esperada para poblaciones de tamaño constante.

4.3. Fórmula de Muestreo de Ewens

En la década del 60 se terminó de descifrar el código genético y se logró evaluar la variación de un gran número de loci en diferentes poblaciones gracias a la sistematización de la electroforesis de proteínas, que además hace posible detectar mutaciones, sin importar si los nuevos alelos introducen fenotipos diferentes a los ya existentes o no. Las mutaciones que no producen efectos diferentes al ser traducidas son conocidas como mutaciones silenciosas o sinónimas, ya que si bien el nuevo nucleótido es diferente, y por lo tanto el codón del que es parte también, el aminoácido al que se traduce es el mismo que el que se traducía antes de mutar. Esto pasa muchas veces cuando la mutación se da en la tercera posición del codón.

Además esta técnica permitió obtener información representativa de las poblaciones, ya que hizo posible disponer de información de muchos individuos, lo que causó que las frecuencias alélicas de las muestras tomadas al azar sean representativas de las frecuencias de la población a la que pertenecen.

En esta sección se verá una fórmula explícita para la distribución estacionaria del modelo Fisher-Wright suponiendo que existen infinitos alelos disponibles. La consecuencia directa que tiene dicha suposición es que cada mutación resultará en un alelo nuevo. Se seguirá el razonamiento de Kimura, quien junto a Crow (1964) desarrolló el modelo que sigue.[7]

Si un gen consiste en 500 nucleótidos (en general las secuencias que los codifican pueden tener muchísimos más) el número de secuencias que se pueden formar es $4^{500} = 10^{500log4} = 10^{301}$. Dada una secuencia de 500 nucleótidos se pueden alcanzar $3 \cdot 500 = 1500$ secuencias. Por lo tanto, la probabilidad de obtener con 2 mutaciones la secuencia original es $\frac{1}{1500}$, asumiendo que los eventos son equiprobables.

Es importante aclarar que en este caso el concepto de alelo se refiere a secuencias diferentes solamente y no a la definición clásica, ya que no se tiene en cuenta la expresión de la secuencia, como se vio anteriormente.

A continuación se considerará una población de tamaño N, con tasa de mutación por individuo y por generación μ. Para k linajes las probabilidades de coalescencia y de mutación son, respectivamente:

$$\frac{k(k-1)}{2}\frac{1}{2N} \quad \text{y} \quad k\mu.$$

Reescalando el tiempo a $2N$ generaciones las tasas quedan:

$$\frac{k(k-1)}{2} \quad \text{y} \quad k\frac{\theta}{2},$$

donde $\theta = 4N\mu$.

Para calcular la probabilidad que ocurra una mutación antes que un evento de coalescencia se hace un cálculo similar que el que se hizo en el modelo con los 2 alelos:

$$\mathbb{P}(\textit{mute antes de coalescer}) = \frac{k\frac{\theta}{2}}{k\frac{\theta}{2} + \frac{k(k-1)}{2}} = \frac{\theta}{\theta + k - 1}$$

Como en el transcurso de este proceso no importa en qué momento ocurren los eventos sino el orden en que lo hacen, se puede utlizar un modelo urna siguiendo las ideas de Hoppe (1984).

Esta urna contiene una *bola negra* con masa θ y varias *coloreadas* con masa 1. En cada paso se elige una bola: si la bola es de color, ella y otra de ese color se vuelven a agregar en la urna; si la bola elegida es negra, se mete junto a la negra una bola de un nuevo color con masa 1.

Es claro que el hecho de sacar una bola de color corresponde a un evento de coalescencia y el de sacar una negra a un evento de mutación.

Observación 4.3.1. Si se revierte el tiempo en la urna de Hoppe, la probabilidad de perder un color en la transición de la generación k a la $k-1$ es $\frac{\theta}{\theta+k-1}$, y la de tener un evento de coalescencia es $\frac{k-1}{\theta+k-1}$.

Observación 4.3.2. La genealogía de k partículas puede ser representada corriendo el modelo de la urna de Hoppe k pasos hacia adelante.

Esta última observación es útil para calcular la distribución estacionaria. Para hacerlo se definirá una variable aleatoria que cuenta la cantidad de alelos diferentes presentes en la muestra y otra variable que indica en cada paso si se agregó o no un nuevo alelo.

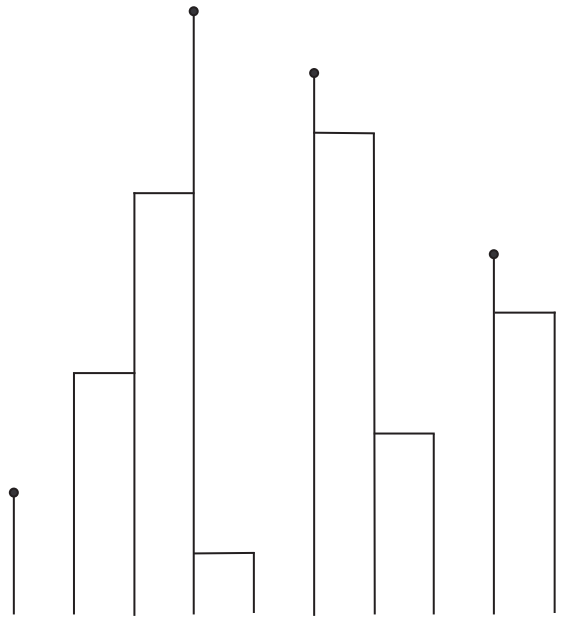


Figura 4.2: Simulación de la urna de Hoppe. Los puntos corresponden a nuevos colores introducidos (bolas negras obtenidas) y las líneas horizontales a la obtención de una bola de color.

Definición 4.3. *La* variable aleatoria K_n *es la cantidad de alelos diferentes presentes en una muestra de n individuos.*

Definición 4.4. *Se define la* variable aleatoria η_i*, como* $\eta_i = 1$ *si la bola que se agrega en el i-ésimo paso introduce un color diferente a los presentes en la urna y 0 si no.*

Las definiciones anteriores se relacionan a través de la siguiente igualdad:

$$K_n = \eta_1 + \ldots + \eta_n$$

donde $\eta_1, \ldots, \eta_n$ son independientes con $\mathbb{P}(\eta_i = 1) = \frac{\theta}{\theta+i-1}$

Para estudiar la distribución de K_n se aplicará la versión de Lindeberg del Teorema Central del Límite, teniendo en cuenta que:

$$\mathbb{P}(\eta_i = 1) = p \text{ y } \mathbb{P}(\eta_i = 0) = 1 - p \Rightarrow E\eta_i = p \text{ y } \text{var}(\eta) = p(1-p)$$

Teorema 4.4. *Para un θ fijo, si el tamaño de la muestra $n \to \infty$ entonces:*

$$EK_n \sim \theta \ln n \ \ y \ \ var(K_n) \sim \theta \ln n, \tag{4.7}$$

donde $a_n \sim b_n$ significa que $\frac{a_n}{b_n} \to 1$ cuando $n \to \infty$.

Además

$$\mathbb{P}\left(\frac{K_n - EK_n}{\sqrt{var(K_n)}} \leq x\right) \to \mathbb{P}(\chi \leq x), \tag{4.8}$$

donde χ es una variable aleatoria con distribución $\mathcal{N}(0,1)$.

Demostración de (4.8): Como las η_i son v.a.i., asumiendo que se cumple (4.7), se verá que las variables aleatorias $\xi_i = \eta_i - \mathbb{E}(\eta_i)$ están en las hipótesis de la versión de Lindeberg del Teorema Central del Límite (ver apéndice).

$\mathbb{E}(\xi_i) = \mathbb{E}(\eta_i - \mathbb{E}(\eta_i)) = 0$

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

$$\begin{aligned} Var(\xi_i) = &\ \mathbb{E}(\xi_i^2) - E^2(\xi_i) \\ = &\ E\left((\eta_i - \mathbb{E}(\eta_i))^2\right) \\ = &\ Var(\eta_i) \end{aligned}$$

Se define $V_n = \sqrt{Var\ S_n} = \sqrt{\theta \ln n}$

Las hipótesis de la versin de Lindeberg del Teorema Central del Límite piden que

$$\text{Dado } \gamma > 0 \quad \Theta_n(\gamma) = \frac{1}{V_n^2} \sum_{i=1}^n E\left(\xi_i^2 \mathbb{I}_{\{|\xi_i| \geq \gamma V_n\}}\right) \to 0 \text{ si } n \to \infty$$

$|\xi_i| = \left|\eta_i - \frac{\theta}{\theta+i-1}\right| \leq 1$ ya que las η_i valen 0 o 1 y $\frac{\theta}{\theta+i-1} < 1$.

Por lo tanto dado $\gamma > 0$ al elegir n_0 tal que $\gamma V_n > 1$, las indicatrices valen 0 $\forall n \geq n_0$.

La tesis de dicho teorema dice que $\frac{S_n}{V_n} \Rightarrow \mathcal{N}(0,1)$

Como $S_n = K_n - \mathbb{E}(K_n)$ sustituyendo se obtiene el resultado esperado.

♠

Demostración de (4.7):

$$EK_n = E\left(\sum_{i=1}^{n}\eta_i\right) = \sum_{i=1}^{n} E\left(\eta_i\right) = \sum_{i=1}^{n}\frac{\theta}{\theta+i-1}$$

La última parte de la igualdad se puede ver como una suma de Riemann aproximando una integral, entonces:

$$\sum_{i=1}^{n}\frac{\theta}{\theta+i-1} \sim \int_{\theta}^{n+\theta}\frac{1}{x}\,dx = \ln(n+\theta) - \ln(\theta) \sim \ln n.$$

Por otra parte:

$$\mathrm{Var}(K_n) \overset{indep.}{=} \sum_{i=1}^{n}\mathrm{Var}(\eta_i) = \sum_{i=2}^{n}\frac{\theta(i-1)}{(\theta+i-1)^2}$$

$$\text{Como } \lim_{i\to\infty}\frac{\theta(i-1)}{(\theta+i-1)} = 1, \text{ se tiene}$$

$$\mathrm{Var}(K_n) \sim \sum_{i=1}^{n}\frac{\theta}{\theta+i-1} \sim \ln n.$$

♠

Corolario 4.5. *$\frac{K_n}{\ln n}$ es un estimador asintótico de la tasa de mutación θ.*

Observación 4.3.3. La desviación standard de $K_n/\ln n$ es del orden de $1/\sqrt{\ln n}$. Por lo tanto, si el verdadero θ es 1 y se quiere estimar θ con un error standard de 0.1, se requiere un muestra de tamaño e^{100}.

Con la Ewen's Sampling Formula se verá que K_n es un estadístico suficente, o sea, que utiliza toda la información brindada por la muestra para estimar θ.[2]

El resultado anterior describe el comportamiento asintótico del número de alelos; el siguiente utiliza más informacón, ya que, incluye la cantidad de individuos por alelo.

[2]En [12] Simon Tavare estudia otros estimadores para θ, pero concluye que el mejor estimador asintótico es el que se acaba de ver.

Notación: a_i es el número de alelos presentes i veces en la población.

Teorema 4.6. *Fórmula de Muestreo de Ewens [2]*
Si la tasa de mutación reescalada es $\theta = 4N\mu$, entonces:

$$\mathbb{P}(a_1, \ldots, a_n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!},$$

donde $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$

Demostración: Por la observación (4.3.2), alcanza con probar que la distribución de colores en la urna de Hoppe en el tiempo n, cumple con la Formula de Muestreo de Ewens.

La demostración de la fórmula se hará por inducción en el tamaño de la muestra n.

$n = 1$**:** La partición $a_1 = 1$ tiene probabilidad 1.

Se supone que en el tiempo $\boldsymbol{n}$ se tiene el estado $a = (a_1, a_2, \ldots, a_n)$ y que en el tiempo $\boldsymbol{n-1}$ la configuración es $\bar{a}$. Se supone además que la fórmula vale para poblaciones de tamaño $n-1$. Se obtendrá entonces la probabilidad de la configuración a.

En la transición de $n-1$ a n individuos pueden haber pasado 2 cosas:

1. Que haya salido la bola negra, por lo que se le agregó un nuevo color a la urna, de manera que en la población hay un nuevo alelo con un solo representante.

2. Que haya salido una bola de color, por lo tanto se agregó otra de ese color, de manera que ese alelo que tenía j individuos tiene ahora $j+1$ representantes en total.

Se probará que

$$\sum_{\bar{a}} \frac{\mathbb{P}(\bar{a})}{\mathbb{P}(a)} p(\bar{a}, a) = 1,$$

donde $p(\bar{a}, a)$ es la probabilidad de transición del estado $\bar{a}$ al estado a.
En cada caso estas probabilidades son:

1. $\bar{a}_1 = a_1 - 1$

$$p(\bar{a}, a) = \frac{\theta}{\theta + n - 1},$$

y además la razón de las probabilidades en la Formula de Muestro de Ewens es

$$\frac{\mathbb{P}(a)}{\mathbb{P}(\bar{a})} = \frac{n}{\theta + n - 1} \cdot \frac{\theta}{a_1}.$$

2. Para algún $1 \leq j < n \quad \bar{a}_j = a_j + 1, \bar{a}_{j+1} = a_{j+1} - 1$, por lo tanto:

$$p(\bar{a}, a) = \frac{j\bar{a}_j}{\theta + n - 1},$$

donde el subíndice j corresponde a las bolas de un mismo color que se pueden elegir y $\bar{a}_j$ son los colores que tienen exactamente j bolas.

Con esto queda

$$\frac{\mathbb{P}(a)}{\mathbb{P}(\bar{a})} = \frac{n}{\theta + n - 1} \cdot \frac{j\bar{a}_j}{(j+1)a_{j+1}}.$$

Por lo tanto:

$$\begin{aligned} \sum_{\bar{a}} \frac{\mathbb{P}}{\mathbb{P}(a)} p(\bar{a}, a) = & \quad \tfrac{\theta}{\theta+n-1} \cdot \tfrac{\theta+n-1}{n} \cdot \tfrac{a_1}{\theta} \\ + & \quad \textstyle\sum_{j=1}^{n-1} \tfrac{j\bar{a}_j}{\theta+n-1} \cdot \tfrac{\theta+n-1}{n} \cdot \tfrac{(j+1)a_{j+1}}{j\bar{a}_j} \end{aligned}$$

Finalmente:

$$\sum_{\bar{a}} \frac{\mathbb{P}}{\mathbb{P}(a)} p(\bar{a}, a) = \frac{a_1}{n} \sum_{j=1}^{n-1} \frac{(j+1)a_{j+1}}{n} = 1,$$

ya que por hipótesis $\sum_k k a_k = n$.

Como la distribución de la urna de Hoppe satisface la recursión y tiene las mismas condiciones iniciales que el coalescente con mutación, las distribuciones de ambos procesos son las mismas.

♠

Ejemplo 4.2. *A continuación se calculará la distribución del número de individuos por alelo en una población en la que hay solamente dos alelos presentes, condicionando a que hubo una mutación.*

Se considerará primero la situación en que un alelo tiene m representantes y el otro $n-m > m$. A esta partición se le llamará $a^{m,m-n}$, donde $a_m^{m,m-n} = 1$, $a_{m-n}^{m,m-n} = 1$ y $a_p^{m,m-n} = 0 \ \forall a_p^{m,m-n} \neq a_m^{m,m-n}, a_{m-n}^{m,m-n}$.

Definiendo $q(m, m-n) = \mathbb{P}(a^{m,m-n})$ y aplicando (4.6) se obtiene:

$$q(m, m-n) = \frac{n!}{\theta_{[n]}} \cdot \frac{\theta}{m^1 \cdot 1!} \cdot \frac{\theta}{(n-m)^1 \cdot 1!} = \frac{n!\theta^2}{\theta_{(n)}} \cdot \frac{1}{m(n-m)}$$

Y en el caso que $m = n - m$*:*

$$q(m, m) = \frac{n!}{\theta_{(n)}} \cdot \frac{\theta^2}{m^2 2!}$$

Joyce y Tavaré (1987) utilizaron el modelo de la urna de Hoppe, pero teniendo en cuenta además las relaciones entre los individuos. Para hacerlo le fueron asignando números a las bolas a medida que iban entrando a la urna; y para registrar las relaciones ancestrales entre ellas anotaron como ciclos a las bolas que pertenecen a una misma familia. La notación sigue las siguientes reglas:

- Un color nuevo inicia un ciclo nuevo.
- Si el color de la k-ésima bola se debe a que en el k-ésimo paso salió la bola introducida en el j-ésimo paso, la k se escribe a la izquierda de la j.

Ejemplo 4.3.

(1)	1 siempre es de un color nuevo
(1)(2)	salió una bola negra entonces 2 es de un color nuevo
(1)(32)	salió la bola de 2, entonces 3 es un hijo de 2
(41)(32)	4 es un hijo de 1
(41)(32)(5)	5 es de un color nuevo
(41)(32)(65)	6 es hijo de 5
(741)(32)(65)	7 es hijo de 6
(741)(832)(65)	8 es hijo de 3

Se denominará σ_n a la permutación de n individuos.

Dada la una permutación que representa a la corrida del modelo k pasos se obtiene directamente la historia de ese grupo. Además la manera de llegar a cada permutación es única, por lo que se obtiene el resultado siguiente.

Teorema 4.7. *Si* σ *es una permutación con* k *ciclos* $\Rightarrow \mathbb{P}(\sigma_n = \sigma) = \frac{\theta^k}{\theta_n}$

Justificación: Para ver que el resultado es cierto se calculará la probabilidad de obtener la permutación del ejemplo anterior. La generalización es directa.

$$\mathbb{P}(\sigma_8 = ((741)(832)(65)) = \frac{\theta}{\theta}\frac{\theta}{\theta+1}\frac{1}{\theta+2}\frac{1}{\theta+3}\frac{\theta}{\theta+4}\frac{1}{\theta+5}\frac{1}{\theta+6}\frac{1}{\theta+7}$$

Cada vez que θ aparece en el numerador significa que salió una bola negra (o que un nuevo color fue introducido) mientras que el hecho que esté el factor 1 en el numerador significa que la bola es hija de un padre determinado.

Intuitivamente al generalizar a n individuos, si la permutación tiene k ciclos, θ aparecerá k veces como numerador y θ_n será el denominador.

♠

Utilizando la notación que se introdujo en la Fórmula de Muestreo de Ewens se tiene que el número de ciclos $k = \sum a_j$ y el número de permutaciones del conjunto $\{1, 2, \ldots, n\}$ con a_1 ciclos de tamaño 1, a_2 de tamaño 2, ... es:

$$\frac{n!}{\prod_{j=1}^{n}(j^{a_j} j!)}$$

Sea $|S_n^k|$ el número de permutaciones de $\{1, 2, \ldots, n\}$ con exactamente k ciclos y K_n el número de alelos presentes en la muestra de tamaño n. Del Teorema 4.7 se obtiene:

$$\mathbb{P}(K_n = k) = \frac{\theta^k}{\theta_{(n)}} \cdot |S_n^k| \tag{4.9}$$

Combinando la ecuación anterior con (4.6) se obtiene:

$$\begin{aligned} \mathbb{P}(a_1, a_2, \ldots, a_n | K_n = k) &= \frac{\mathbb{P}(\{a_1,a_2,\ldots,a_n\} \cap \{K_n=k\})}{\mathbb{P}(K_n=k)} \\ &= \frac{\mathbb{P}(a_1,a_2,\ldots,a_n)}{\mathbb{P}(K_n=k)} \\ &= \frac{n!}{|S_n^k|} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{a_j} \frac{1}{a_j!} \end{aligned} \tag{4.10}$$

Observación 4.3.4. Al condicionar la partición alélica de la población a la cantidad de alelos presentes en la muestra, la probabilidad no depende de θ. Por lo tanto K_n es un estadístico suficiente para estimar θ.

Observación 4.3.5. $|S_n^k|$ es el número de Stirling de segundo tipo. Para calcularlos alcanza con la siguiente relación: $|S_n^k| = (n-1)|S_{n-1}^k| + |S_{n-1}^{k-1}|$.

Esto quiere decir que se puede construir un ciclo $\sigma \in S_n^k$ a partir de un miembro de S_n^{k-1} agregándole (n) como un nuevo ciclo o a partir de un ciclo $\pi \in S_n^{k-1}$ eligiendo un entero $1 \leq j \leq n-1$ y estableciendo $\sigma(j) = n, \sigma(n) = \sigma(j)$.

Capítulo 5

Estimadores de parámetros de mutación

En este capítulo se considerará, como se hizo en el capítulo anterior que hay infinitos alelos y que cada mutación se traduce en un alelo nuevo; por lo tanto el hecho que dos individuos sean del mismo tipo se debe a que ambos obtuvieron sus alelos por descendencia. La gran diferencia con el modelo de Fisher-Wright es que en el modelo que se vio en el Capítulo II los hijos son del mismo tipo que su padre (o madre); esto lleva a que se pierda toda la variabilidad, mientras que en el modelo de los alelos infinitos no se sabe con exactitud qué alelo portarán los hijos; esto provocará evolución sostenida en el tiempo, ya que la mutación es una fuente inagotable de variabilidad. La ausencia de selección hace que el modelo sea un modelo neutral de evolución.

En el modelo que sigue aparte de tener en cuenta el número de alelos diferentes, como se hizo para calcular la fórmula de muestreo de Ewens, se contará la cantidad de diferencias que existen entre las secuencias que los originan, por lo que se estará teniendo en cuenta más información que antes.

El objetivo de este capítulo es buscar estimadores de parámetros de mutación. Para hacerlo en algunos casos se tendrá en cuenta la historia de la muestra, su proceso de ramificación y en qué momento ocurrieron las mutaciones; en otros, al igual que se hizo en el capítulo anterior, solamente la partición alelélica de la muestra.

Definición 5.1. *Se llamarán* sitios segregantes *a las posiciones nucleotídicas en las que existe alguna diferencia entre por lo menos un par de individuos del conjunto que se esté estudiando.*

5.1. Organización de los datos y medidas.

La organización de los datos es muy útil para facilitar el trabajo, ya que si se tiene una muestra de n secuencias con largo s[1], habrá que almacenar y comparar $n \times s$ datos (letras).

En principio por sitio hay solamente dos letras posibles, ya que la probabilidad de que el mismo sitio mute más de una vez es muy baja. Este hecho sucede si el largo de las secuencias es muy grande, por lo que además de estar bajo las hipótesis de alelos infinitos, se va a hablar de sitios infinitos.

Notación: Para almacenar los datos se le adjudicará un 0 a la letra más numerosa del sito y un 1 al otro (suponiendo que la letra a la que se le adjudicó el 0 es el alelo salvaje). A cada una de las secuencias de largo s se la denotará $y_i = (y_{i1}, y_{i2}, \ldots, y_{is})$.

A continuación se definen diferentes medidas de diferencias entre secuencias:

Definición 5.2. *La* distancia *entre dos secuencias* $\Pi(i,j)$ *es el número de sitios en los que las secuencias* i *y* j *difieren*[2]*.*

$$\Pi(i,j) = \sum_{l=1}^{s} \mathbb{I}_{\{y_{il} \neq y_{jl}\}}, \ i \neq j \tag{5.1}$$

Trabajando con las secuencias como sucesiones de 0's y 1's esta distancia se puede calcular también como

$$\Pi(i,j) = \sum_{l=1}^{s} |y_{il} - y_{jl}|, \ i \neq j \tag{5.2}$$

Definición 5.3. *La* diversidad nucleotídica *de una muestra ,* Π_n*, es el promedio de las diferencias entre todos los pares de la misma.*

$$\Pi_n = \frac{1}{n(n-1)} \sum_{i \neq j} \Pi(i,j) \tag{5.3}$$

Definición 5.4. *La* diversidad nucleotídica por sitio*,* π_n*, es el promedio entre todos los sitios de la diversidad nucleotídica.*

[1]Dada la gran cantidad de datos que es posible obtener hoy en día y dependiendo del gen que se esté estudiando s puede ser muy grande.

[2]Esta distancia es conocida como la distancia de Hamming

$$\pi_n = \frac{\Pi_n}{s} \tag{5.4}$$

Para obtener la heterocigosidad de la muestra en el sitio l se verá otra manera de calcular la diversidad nucleotídica.

En cada sitio se encuentra una de las 4 letras posibles del alfabeto $\mathcal{A} = \{A, C, G, T\}$. Se denotará $n_{l\alpha}$ a la cantidad de veces que aparece la letra α en el sitio l de todos los elementos del conjunto.

Si hay solamente 2 letras posibles por sitio $n_{l\alpha}(n - n_{l\alpha})$ es la cantidad de pares sin ordenar de secuencias que difieren en ese sitio.

$$\Pi_n = \frac{1}{n(n-1)} \sum_{l=1}^{s} \sum_{\alpha \in \mathcal{A}} n_{l\alpha}(n - n_{l\alpha}) := \frac{n}{n-1} \sum_{l=1}^{s} H_l, \tag{5.5}$$

donde H_l es la heterocigosidad en el sitio l, que se define como

$$H_l = \sum_{\alpha \in \mathcal{A}} \frac{n_{l\alpha}}{n} \left(1 - \frac{n_{l\alpha}}{n}\right).$$

La diversidad nucleotídica será entonces

$$\pi_n = \frac{n}{n-1} \frac{1}{s} \sum_{l=1}^{s} H_l.$$

Observación 5.1.1. π_n da una medida de la heterocigosidad promedio en la región; teniendo en cuenta que en la fórmula hay un factor de corrección.

Cabe esperar entonces que los resultados de todas estas medidas estén estrechamente relacionadas con el parámetro de mutación de la muestra, por lo tanto sería lógico que podamos relacionar θ[3] con las mismas.

5.2. Curvas de diferencias pareadas

El objetivo de esta sección es averiguar en promedio qué fracción del total de pares estarán separadas por exactamente k sitios segregantes en una muestra de n individuos. Para hacerlo se trabajará con las variables aleatorias $\Pi(i, j)$ que si bien son idénticamente distribuídas, no son independientes. La distribución se puede hallar de la siguiente manera:

$$\mathbb{P}(\Pi(i,j) = k) = E(\mathbb{P}(\Pi(i,j) = k | T_2))$$

[3] θ es la tasa de mutación esperada por individuo por cada $2N$ generaciones multiplicada por 2, esto es: $4N\mu$

Observación 5.2.1. Bajo las condiciones del modelo $\Pi(i,j)$ condicionado a T_2 se puede aproximar por una Poisson de parámetro θT_2.

$$\Rightarrow \mathbb{E}(\Pi(i,j)) = \mathbb{E}(\mathbb{E}(\Pi(i,j)|T_2)) = \mathbb{E}(\theta T_2) = \theta \mathbb{E}(T_2)$$

Si la cantidad de individuos varía según la función $\lambda(t)$ respecto de la población inicial se obtiene

$$\mathbb{P}(\Pi(i,j) = k) = \int_0^\infty e^{-\theta k} \frac{(\theta t)^k}{k!} \lambda(t) e^{-\Lambda(t)} dt, \quad \text{con } \Lambda(t) = \int_0^t \lambda(s) ds \quad (5.6)$$

En el caso en el que el tamaño de población es constante $\lambda(t) = 1$, por lo tanto $\Lambda(t) = t$, de donde, calculando la integral, se obtiene:

$$\mathbb{P}(\Pi(i,j) = k) = \frac{1}{1+\theta} \left(\frac{\theta}{1+\theta} \right)^k, \ k = 0, 1, 2, \ldots$$

Observación 5.2.2. Este resultado se podría haber obtenido usando las probabilidades calculadas en la observación 4.3.1. En este caso $k = 2$ por lo tanto la probabilidad de que ocurra un evento de mutación antes de que un evento de coalescencia es $\frac{\theta}{2-1+\theta}$ y la probabilidad que ocurra un evento de coalescencia es $\frac{1}{2-1+\theta}$. Entonces, para obtener exactamente k mutaciones, antes que ocurra el evento de coalescencia, es necesario que ocurran todas las mutaciones y luego se produzca la coalescencia, por lo tanto la probabilidad será $\left(\frac{\theta}{1+\theta}\right)^k \frac{1}{1+\theta}$.

La distribución obtenida es una geométrica con shift, $\mathbb{P}(X = k) = (1-p)^k p$, por lo tanto su esperanza es $\frac{1}{p} - 1$ y su variaza $\frac{1-p}{p^2}$. Se concluye entonces que en el caso en el que el tamaño de la población es constante

$$\mathbb{E}(\Pi(i,j)) = \theta \text{ y } Var(\Pi(i,j)) = \theta(1+\theta) = \theta + \theta^2.$$

La curva de diferencias pareadas se construye a partir de la distribución empírica del conjunto $\{\Pi(i,j),\ 1 \leq i \neq j \leq n\}$. Para estimar las probabilidades de la ecuación 5.6, se define una variable aleatoria, cuya función es contar los pares de secuencias separados por exactamente k sitios segregantes.

Definición 5.5.

$$\Pi_{nk} = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{\{\Pi(i,j)=k\}}$$

Π_{nk} es la fracción de pares de secuencias separadas por exactamente k sitios.

$$\begin{aligned}\mathbb{E}(\Pi_{nk}) &= E\left(\tfrac{1}{n(n-1)}\sum_{i\neq j}\mathbb{I}_{\Pi(i,j)=k}\right)\\ &= E\left(\mathbb{I}_{\Pi(i,j)=k}\right)\\ &= \mathbb{P}(\Pi(1,2)=k), \quad k=0,1,2,\ldots\end{aligned}$$

Observación 5.2.3. Para estimar θ a partir de la curva de diferencias pareadas se grafica la distribución empírica de las mismas y se busca el θ que al graficar las probabilidades mejor se ajuste a la curva empírica.

5.3. Diversidad nucleotídica

Para obtener un estimador de θ a partir de la diversidad nucleotídica se calculará la esperanza.

$$\begin{aligned}\mathbb{E}(\Pi_n) &= E\left(\tfrac{1}{n(n-1)}\sum_{i\neq j}\Pi(i,j)\right)\\ &= \mathbb{E}(\Pi(1,2))\\ &= \mathbb{E}(\text{cantidad de sitios segregantes})\\ &= \theta\mathbb{E}(T_2)\end{aligned}$$

De la primer igualdad a la segunda se pasa aplicando la linealidad de la esperanza, que la esperanza de la distancia entre cada par es la misma y que el total de pares posibles (teniendo en cuenta el orden) es $n(n-1)$.

Para pasar de la tercera a la cuarta hay que tener en cuenta el proceso por el cual se generan los sitios segregantes entre un par de secuencias. Partiendo del ancestro común entre ambas; cuando este individuo se convierte en dos, se generan dos secuencias exactamente iguales (porque la probabilidad de que en ese momento además haya una mutación es muy baja), a medida que pasa el tiempo se podrá producir una mutación (o sea, obtener un sitio segregante) con probabilidad θ por cada generación. El tiempo durante el cual pueden ocurrir entonces las mutaciones es T_2 [4].

En el caso en que el tamaño poblacional es constante se vio que $\mathbb{E}(T_2)=1$, por lo tanto *si el número de individuos se mantiene constante*:

$$\mathbb{E}(\Pi_n)=\theta \tag{5.7}$$

La varianza de Π_n fue encontrada por Tajima [9]:

[4] Recordar que T_2 es el tiempo hasta el MRCA medido de a 2N generaciones.

Teorema 5.1.

$$Var(\Pi_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \tag{5.8}$$

Demostración:

$$\Pi_n^2 = \left(\frac{1}{n(n-1)}\right)^2 \sum_{\{i_1,j_1\}} \sum_{\{i_2,j_2\}} \Pi(i_1,j_1)\Pi(i_2,j_2)$$

La suma se puede separar en tres sumandos diferentes:

- $\#\{i_1,i_2,j_1,j_2\} = 2$ si $i_1 = i_2$ y $j_1 = j_2$: C_2^n sumandos.
- $\#\{i_1,i_2,j_1,j_2\} = 3$ si $i_1 = i_2$ y $j_1 \neq j_2$ o al revés : $n(n-1)(n-2) = 2(n-2)\cdot C_2^n$ sumandos.
- $\{i_1 \neq i_2\} \cap \{j_1 \neq j_2\} = \emptyset$: $C_2^n \cdot C_2^{n-2}$ sumandos.

Para chequear que estos son todos los sumandos posibles, sumando se verifica que el total da $(C_2^n)^2$.

$C_2^n \cdot \left(1 + 2(n-2) + C_2^{n-2}\right) = C_2^n \cdot \left(1 + 2(n-2) + \frac{(n-2)(n-3)}{2}\right) =$
$C_2^n \cdot \left(\frac{2+4n-8+n^2-5n+6}{2}\right) = C_2^n \cdot \left(\frac{n^2-n}{2}\right) = (C_2^n)^2$

Notación: Para i,j,k,l índices diferentes, se define:

- $U_2 = E\left((\Pi(i,j))^2\right) - \theta^2$
- $U_3 = \mathbb{E}(\Pi(i,j)\Pi(i,k)) - \theta^2$
- $U_4 = \mathbb{E}(\Pi(i,j)\Pi(k,l)) - \theta^2$

Con la nueva notación definida la varianza de Π_n queda escrita de la siguiente manera:

$$(\star) \quad Var(\Pi_n) = (C_2^n)^{-1}\left(U_2 + 2(n-2)U_3 + C_2^{n-2}U_4\right)$$

Se calcularán U_2, U_3 y U_4 calculando las varianzas de Π_2, Π_3 y Π_4 respectivamente y luego se sustituirá en $(\star)$.

Observación 5.3.1. De la observación 5.2.1 se tiene que $\mathbb{E}(\Pi_2) = \theta$ y que $Var(\Pi_2) = \theta + \theta^2$, por lo tanto:

$$U_2 = \theta + \theta^2$$

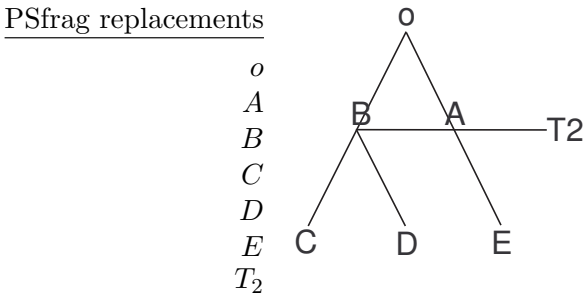


Figura 5.1: Relación ancestral entre 3 individuos.

n=3: Se denotará $n_{AB} = n_{BA}$ al número de mutaciones ocurridas en el camino AoB, $n_{BC} = n_{CB}$ a las ocurridas en la rama BC, etc..

Con esta notación las diferencias entre pares verifican las ecuaciones siguientes:

- $\Pi(C, D) = n_{CB} + n_{BD}$
- $\Pi(C, E) = n_{CB} + n_{BA} + n_{AE}$
- $\Pi(D, E) = n_{DB} + n_{BA} + n_{AE}$

Sumando las 3 igualdades se obtiene:

$$3\Pi_3 = 2n_{AB} + 2n_{BC} + 2n_{BD} + 2n_{AE}. \tag{5.9}$$

De 5.2.1

$$\mathbb{E}(n_{AB}) = \theta \text{ y } Var(n_{AB}) = \theta + \theta^2. \tag{5.10}$$

n_{AB} es independiente de n_{BC}, n_{BD} y n_{AE}, a su vez las tres últimas tienen la misma distribución, ya que son el número de mutaciones acumuladas en un linaje a lo largo del tiempo T_2. Por lo tanto las covarianzas entre los pares serán también las mismas ($cov(n_{BC}, n_{BD}) = cov(n_{BD}, n_{AE}) = cov(n_{BC}, n_{AE})$).

$$\Rightarrow \frac{9}{4} \cdot Var(\Pi_3) = Var(n_{AB}) + 3 \cdot Var(n_{BC}) + 6 \cdot cov(n_{BC}, n_{BD}) \tag{5.11}$$

n_{BC} cuenta la cantidad de diferencias entre B y C ocurridas durante el tiempo T_2. Razonando de la misma manera que se hizo en la observación

5.2.1 se concluye que la distribución de esta variable es la de Poisson con parámetro $T_2\frac{\theta}{2}$.

$$\Rightarrow \mathbb{E}(n_{BC}) = E\left(\mathbb{E}(n_{BC}|T_2)\right) = E\left(T_2\frac{\theta}{2}\right) = \frac{\theta}{2}\mathbb{E}(T_2) = \frac{\theta}{2}\frac{1}{C_2^3} = \frac{\theta}{2}\frac{1}{3} = \frac{\theta}{6} \tag{5.12}$$

Para calcular $E\left(n_{BC}^2\right)$, basta recordar que si una variable aleatoria Z tiene distribución de Poisson de parámetro λ, entonces $\mathbb{E}(Z^2) = \lambda+\lambda^2$ y que si una variable aleatoria X tiene distribución exponencial con parámetro μ, entonces $\mathbb{E}(X^2) = 2\mu^2$.

$$\begin{aligned}
\Rightarrow \mathbb{E}(n_{BC}^2) = &\quad E\left(\mathbb{E}(n_{BC}^2|T_2)\right) \\
= &\quad E\left(T_2\tfrac{\theta}{2} + \left(T_2\tfrac{\theta}{2}\right)^2\right) \\
= &\quad \tfrac{\theta}{2}\mathbb{E}(T_2) + \tfrac{\theta^2}{4}\mathbb{E}(T_2^2) = \\
= &\quad \tfrac{\theta}{2}\tfrac{1}{3} + \tfrac{\theta^2}{4}\cdot 2\cdot\tfrac{1}{9} \\
= &\quad \tfrac{\theta}{6} + \tfrac{\theta^2}{18}
\end{aligned}$$

$$\Rightarrow Var(n_{BC}) = \mathbb{E}(n_{BC}^2) - E^2(n_{BC}) = \frac{\theta}{6} + \frac{\theta^2}{36} \tag{5.13}$$

Como n_{BC} y n_{BD} son independientes al condicionar a T_2, se tiene que

$$\mathbb{E}(n_{BC}n_{BD}|T_2) = \mathbb{E}(n_{BC}|T_2)\mathbb{E}(n_{BD}|T_2) = \left(T_2\frac{\theta}{2}\right)^2.$$

$$\Rightarrow \mathbb{E}(n_{BC}n_{BD}) = \frac{\theta^2}{18}$$

Dadas dos variables aleatorias X e Y, la covarianza se calcula de la siguiente manera: $cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

$$\Rightarrow cov(n_{BC}, n_{BD}) = \frac{\theta^2}{18} - \left(\frac{\theta}{6}\right)^2 = \frac{\theta^2}{36} \tag{5.14}$$

Combinando (5.10), (5.11), (5.13) y (5.14) se tiene que:

$$Var(\Pi_3) = \frac{4}{9}\left(\theta + \theta^2 + 3\left[\frac{\theta}{6} + \frac{\theta^2}{36}\right] + 6\cdot\frac{\theta^2}{36}\right).$$

$$\Rightarrow Var(\Pi_3) = \frac{2}{3}\theta + \frac{5}{9}\theta^2 \tag{5.15}$$

Por lo tanto para $n = 3$ se cumple la tesis.

PSfrag replacements

n=4: En este caso hay que considerar más tipos de árboles y a su vez más segmentos.

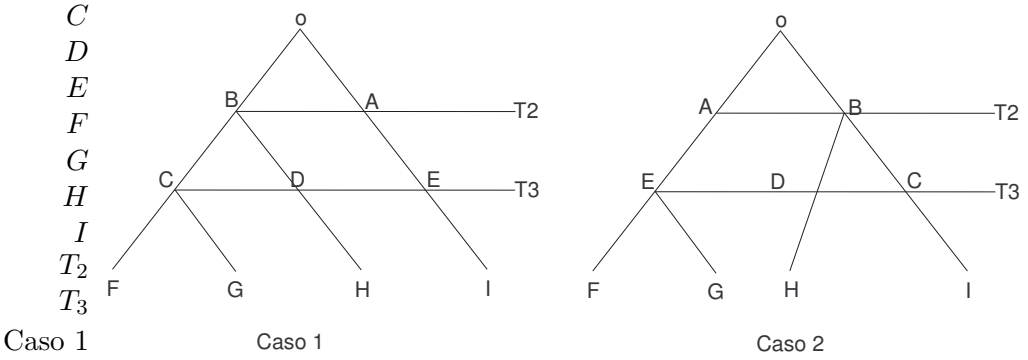


Figura 5.2: Relaciones ancestrales entre 4 individuos.

Las probabilidades de que ocurra cada uno de los casos es $\frac{2}{3}$ para el primero y $\frac{1}{3}$ para el segundo. Esto se debe a que luego de ocurrido el primer evento de coalescencia entre los individuos F y G, quedan tres individuos de los cuales 2 de ellos tienen un ancestro común. Los posibles pares de hermanos son el ancestro común de F y G con H, el de F y G con I o H con I. De estas tres posibilidades se obtienen 2 topologías posibles, la del primer caso corresponde a los primeros 2 pares y la del segundo caso corresponde al tercer par.

Caso 1

$$\Pi_4 = \Pi(F,G) + \Pi(F,H) + \Pi(F,I) + \Pi(G,H) + \Pi(G,I) + \Pi(H,I)$$

- $\Pi(F,G) = n_{FC} + n_{CG}$
- $\Pi(F,H) = n_{FC} + n_{CB} + n_{DB} + n_{DH}$
- $\Pi(F,I) = n_{FC} + n_{CB} + n_{BA} + n_{AE} + n_{EI}$

- $\Pi(G,H) = n_{GC} + n_{CB} + n_{DB} + n_{DH}$
- $\Pi(G,I) = n_{GC} + n_{CB} + n_{BA} + n_{AE} + n_{EI}$
- $\Pi(H,I) = n_{HD} + n_{DB} + n_{BA} + n_{AE} + n_{EI}$

$$\Rightarrow 6\Pi_4 = 3n_{AB}+4n_{BC}+3n_{BD}+3n_{AE}+3n_{FC}+3n_{GC}+3n_{HD}+3n_{EI} \quad (5.16)$$

Caso 2

En este caso haciendo el mismo razonamiento que en el anterior se obtiene:

$$\Rightarrow 6\Pi_4 = 4n_{AB}+3n_{BC}+3n_{BD}+4n_{AE}+3n_{FC}+3n_{GC}+3n_{HD}+3n_{EI} \quad (5.17)$$

Se dividirán las variables en tres conjuntos: el primero será $\{n_{AB}\}$, el segundo $\{n_{BC}, n_{BD}, n_{AE}\}$ y el tercero $\{n_{FC}, n_{GC}, n_{HD}, n_{IE}\}$.

Observar que las variables que se encuentran en diferentes conjuntos son independientes. Para las variables de los primeros 2 conjuntos los cálculos se hicieron en el caso para n=3. Para hacer los cálculos para las variables que quedan, se razonará de la misma manera que antes. Es importante destacar que

$$\mathbb{E}(T_3) = \frac{1}{C_2^4} = \frac{1}{6}$$

$$\mathbb{E}(n_{CF}) = \mathbb{E}(\mathbb{E}(n_{CF}|T_3)) = E\left(\frac{\theta}{2}T_3\right) = \frac{\theta}{2}\cdot\frac{1}{6} = \frac{\theta}{12} \quad (5.18)$$

$$\mathbb{E}(n_{CF}^2) = \mathbb{E}(\mathbb{E}(n_{CF}^2|T_3)) = \mathbb{E}(T_3)\tfrac{\theta}{2} + E(T_3\tfrac{\theta}{2})^2 = \tfrac{1}{6}\tfrac{\theta}{2} + \tfrac{2}{36}\tfrac{\theta^2}{4} = \tfrac{\theta}{12} + \tfrac{\theta^2}{72}$$

$$\Rightarrow Var(n_{CF}) = \mathbb{E}(n_{CF}^2) - E^2(n_{CF}) = \frac{\theta}{12} + \frac{\theta^2}{144} \quad (5.19)$$

por lo tanto

$$Var(n_{CF}) = \mathbb{E}(n_{CF}^2) - (\mathbb{E}(n_{CF}))^2 = \frac{\theta}{12} + \frac{\theta^2}{144} \quad (5.20)$$

Como n_{CF} y n_{CG} son independientes, condicionando a T_3 se tiene:
$\mathbb{E}(n_{CF}n_{CG}|T_3) = \mathbb{E}(n_{CF}|T_3)\mathbb{E}(n_{CG}|T_3) = \left(\frac{\theta}{2}T_3\right)^2$

$$\Rightarrow \mathbb{E}(n_{CF}n_{CG}) = \frac{\theta^2}{4}\frac{2}{36} = \frac{\theta^2}{76} \quad (5.21)$$

De este resultado y (5.18) se obtiene:

$$cov(n_{CF}, n_{CG}) = \frac{\theta^2}{72} - \left(\frac{\theta}{12}\right)^2 = \frac{\theta^2}{144}. \tag{5.22}$$

Denotando C_i al i-ésimo caso y usando (5.10), (5.12) y (5.18), se obtiene:

$$\mathbb{E}(\Pi_4|C_1) = \tfrac{1}{6}\left(3\mathbb{E}(n_{AB}) + 10\mathbb{E}(n_{BC}) + \mathbb{E}(n_{CF})\right) = \tfrac{\theta}{12}\left(2\cdot 3 + 10\cdot\tfrac{1}{3} + 12\cdot\tfrac{1}{6}\right)$$

$$\mathbb{E}(\Pi_4|C_1) = \frac{17}{18}\theta$$

$$\mathbb{E}(\Pi_4|C_2) = \tfrac{1}{6}\left(4\mathbb{E}(n_{AB}) + 10\mathbb{E}(n_{BC}) + \mathbb{E}(n_{CF})\right) = \tfrac{\theta}{2}\left(2\cdot 4 + 10\cdot\tfrac{1}{3} + 12\cdot\tfrac{1}{6}\right)$$

$$\mathbb{E}(\Pi_4|C_2) = \frac{10}{9}\theta$$

$$\Rightarrow \mathbb{E}(\Pi_4) = \frac{2}{3}\mathbb{E}(\Pi_4|C_1) + \frac{1}{3}\mathbb{E}(\Pi_4|C_2) = \left(\frac{2}{3}\frac{17}{18} + \frac{1}{3}\frac{20}{18}\right)\theta = \theta$$

El resultado concuerda con (5.7).

Para calcular la varianza, se tendrá en cuenta, como se hizo para calcular la esperanza, que variables en diferentes niveles son independientes y que hay 2 casos diferentes de relaciones de ancestralidad.

$$\begin{aligned} 36\ Var(\Pi_4|C_1) = &\ 9\ Var(n_{AB}) + 34\ Var(n_{BC}) + 66\ cov(n_{BC}, n_{BD}) \\ + &\ 36\ Var(n_{CF}) + 108\ cov(n_{CF}). \end{aligned}$$

Sutituyendo los valores obtenidos en (5.10), (5.13), (5.14),(5.20) y (5.22)

$$\begin{aligned} 36\ Var(\Pi_4|C_1) = &\ 9(\theta+\theta^2) + 34\left(\tfrac{\theta}{6} + \tfrac{\theta^2}{36}\right) + 66\tfrac{\theta^2}{36} + 36\left(\tfrac{\theta}{12} + \tfrac{\theta^2}{144}\right) + 108\tfrac{\theta^2}{144} \\ = &\ \left(\tfrac{53}{3}\right)\theta + \left(\tfrac{115}{9}\right)\theta^2. \end{aligned}$$

Haciendo el mismo razonamiento:

$$\begin{aligned} 36\ Var(\Pi_4|C_2) &= 16\ Var(n_{AB}) + 34\ Var(n_{BC}) + 66\ cov(n_{BC}, n_{BD}) \\ &+36\ Var(n_{CF}) + 108\ cov(n_{CF}) \\ &= 36Var(\Pi_4|C_2) + 7Var(n_{AB}) \\ &= \left(\tfrac{74}{3}\right)\theta + \left(\tfrac{178}{9}\right)\theta^2. \end{aligned}$$

La varianza total se obtiene usando que

$Var(Y) = \mathbb{E}(Var(Y|X)) + Var(\mathbb{E}(Y|X))$, tomando $X = C_i,\ i = 1, 2$.

$$\begin{array}{rcl} \Rightarrow Var(\Pi_4) = & \frac{2}{3}Var(\Pi_4|C_1) + \frac{1}{3}Var(\Pi_4|C_2) + \frac{2}{3}\left(\mathbb{E}(\Pi_4|C_1) - \mathbb{E}(\Pi_4)\right)^2 \\ + & \frac{1}{3}\left(\mathbb{E}(\Pi_4|C_2) - \mathbb{E}(\Pi_4)\right)^2 \end{array}.$$

Sustituyendo por los resultados obtenidos anteriormente, se tiene:

$$= \frac{2}{3}\left[\frac{53}{108}\theta + \frac{115}{324}\theta^2\right] + \frac{1}{3}\left[\frac{74}{108}\theta + \frac{178}{324}\theta^2\right] + \frac{2}{3}\left[\frac{17}{18}\theta - \theta\right]^2 + \frac{1}{3}\left[\frac{20}{18}\theta - \theta\right]^2$$

$$\Rightarrow Var(\Pi_4) = \frac{5}{9}\theta + \frac{23}{54}\theta^2 \tag{5.23}$$

Esta varianza, al igual que la varianza calculada para el caso $n = 3$, coincide con la tesis para el caso $n = 4$.

Ahora que se obtuvieron las varianzas de Π_n para $n = 2, 3, 4$, es posible calcular U_2, U_3 y U_4.

Anteriormente se había calculado U_2:

$$U_2 = \theta + \theta^2 \tag{5.24}$$

Para calcular U_3 se usará $(\star)$. Sustituyendo n por 3 se obtiene que $3Var(\Pi_3) = U_2 + 2U_3$.

Sustituyendo por lo resultados obtenidos en (5.20) y (5.24) se cumple que $2\theta + \frac{5}{3}\theta^2 = \theta + \theta^2 + 2U_3$.

$$\Rightarrow U_3 = \frac{1}{2}\left(\theta + \frac{2}{3}\theta^2\right) = \frac{\theta}{2} + \frac{\theta^2}{3} \tag{5.25}$$

Para calcular U_4 se hará lo mismo que para $n = 3$. Usando $(\star)$ se obtiene que $6Var(\Pi_4) = U_2 + 4U_3 + U_4$, y sustituyendo por los resultados de (5.23), (5.24) y (5.25) se cumple que $\frac{10}{3}\theta + \frac{23}{9}\theta^2 = \theta + \theta^2 + 4\left(\frac{\theta}{2} + \frac{\theta^2}{3}\right) + U_4$.

$$\Rightarrow U_4 = \frac{\theta}{3} + \frac{2}{9}\theta^2 \tag{5.26}$$

Sustituyendo ahora (5.24), (5.25) y (5.26) en $(\star)$ se obtiene:

$$\begin{array}{rcl} Var(\Pi_n) = & (C_2^n)^{-1}\left[\theta + \theta^2 + 2(n-2)\left(\frac{\theta}{2} + \frac{\theta^2}{3}\right) + \frac{(n-2)(n-3)}{2}\frac{\theta}{2} + 2\frac{\theta^2}{9}\right] \\ = & (C_2^n)^{-1}\theta\left(\frac{6+6n-12+n^2-5n+6}{6}\right) + \theta^2\left(\frac{9+6n-12+n^2-5n+6}{9}\right) \end{array}$$

$$\Rightarrow Var(\Pi_n) = (C_2^n)^{-1}\left(\frac{n(n+1)}{6}\theta + \frac{n^2+n+3}{9}\theta^2\right).$$

♠

5.4. Número de sitios segregantes.

Dado un conjunto de individuos, el total de mutaciones S_n que se observan en el presente son todas las mutaciones acumuladas desde el MRCA (bajo las hipótesis del modelo de los alelos infinitos). Basándose en este hecho, Watterson (1975) obtuvo un estimador de θ.

Para calcular la esperanza del total de diferencias entre un par de secuencias se condicionó respecto a T_2. Siguiendo el mismo razonamiento al tratar con toda la muestra (n secuencias) habrá que tener en cuenta la cantidad de mutaciones acumuladas en cada linaje, por lo que se condicionará al largo total del árbol L_n.

Definición 5.6. *Se define el* largo total del árbol L_n *como la suma del largo de todas las ramas.*

$$L_n = \sum_{j=2}^{n} jT_j$$

El largo de cada rama es suma de T_j y la cantidad de sumandos depende del tiempo que existió dicha rama.

Al condicionar S_n a L_n se obtiene una Poisson con parámetro $\frac{\theta}{2}L_n$. Observar que el parámetro $\frac{\theta}{2}$ es la probabilidad de mutación de un individuo solamente por generación y que L_n cuenta los tiempos de convergencia de un par como el doble, ya que suma el largo de cada rama.

$$\begin{aligned}\mathbb{E}(S_n) &= E\left(\mathbb{E}(S_n|L_n)\right)\\ &= E\left(\tfrac{\theta}{2}L_n\right)\\ &= \tfrac{\theta}{2}\textstyle\sum_{j=2}^{n} j\mathbb{E}(T_j)\\ &= \tfrac{\theta}{2}\textstyle\sum_{j=2}^{n} j\,\frac{2}{j(j-1)}\\ &= \theta\textstyle\sum_{j=1}^{n-1}\frac{1}{j}\end{aligned} \tag{5.27}$$

Observación 5.4.1. Si n es grande $\mathbb{E}(S_n) \approx \theta\log(n)$, como se vio en el teorema 4.4 del capítulo anterior.

Definición 5.7. *Y_j es el número de mutaciones ocurridas en la muestra cuando la misma tiene exactamente j individuos.*

Como S_n es el número de mutaciones acumuladas a lo largo de toda la historia de la muestra, se puede escribir $S_n = Y_2 + \ldots + Y_n$. Dado que los T_j son independientes, los Y_j también lo son. Al igual que sucede con Y_2, las Y_j restrictas a T_j son Poisson de parámetro $\frac{\theta}{2} j T_j$, ya que durante el tiempo T_j la mutación puede ocurrir en j ramas diferentes.

Al escribir S_n como una suma de variables independientes es fácil encontrar su función generatriz.

Primero se calculará la función generatriz para Y_j, tomando $\alpha = \frac{\theta}{2} j T_j$, para simplificar la notación.

$$\begin{aligned} E\left(s^{Y_j}\right) &= E\left(E\left(s^{Y_j}|T_j\right)\right) \\ E\left(s^{Y_j}|T_j\right) &= \sum_{k=0}^{\infty} s^k \mathbb{P}(X=k) \\ &= \sum_{k=0}^{\infty} s^k \frac{\alpha^k}{k!} e^{-\alpha} \\ &= \sum_{k=0}^{\infty} \frac{(s\alpha)^k}{k!} e^{-\alpha} \\ &= e^{-\alpha} e^{s\alpha} \\ &= e^{\alpha(s-1)} \quad , \alpha = \frac{\theta}{2} j T_j \end{aligned}$$

En la sección 2.2 se vio que $T_j \sim \mathcal{E}(\lambda)$, con $\lambda = \frac{j(j-1)}{2}$ por lo tanto, tomando $k = \frac{\theta}{2} j(s-1)$ se tiene:

$$E\left(e^{kt}\right) = \int_0^{\infty} e^{kt} e^{-\lambda t} \lambda dt = \frac{\lambda}{k-\lambda} e^{(k-\lambda)t} \Big|_0^{\infty}.$$

Tomando $s \in (0,1)$:
$k - \lambda = \frac{\theta}{2} j(s-1) - \frac{j(j-1)}{2} = \frac{j}{2}(\theta(s-1) - (j-1)) < 0 \ \forall j.$

$$\Rightarrow E\left(s^{Y_j}\right) = E\left(e^{\alpha(s-1)}\right) = \frac{j-1}{j-1+\theta(1-s)}$$

En la última igualdad se usó que la función generatriz de una geométrica de parámetro p, es $\frac{p}{1-s(1-p)}$, por lo tanto tomando $p = \frac{\theta}{\theta+j-1}$, se obtiene el resultado anterior.

Observación 5.4.2. $\mathbb{P}(Y_j = k) = \left(\frac{\theta}{\theta+j-1}\right)^k \left(\frac{j-1}{\theta+j-1}\right)$

Observación 5.4.3. El resultado anterior se puede obtener realizando el mismo razonamiento que se hizo en la observación 5.2.2

Para calcular la varianza de las Y_j es posible hacerlo derivando la función generatriz o utilizando el hecho que tienen disribución geométrica y la varianza de una geométrica es $\frac{p}{(1-p)^2} = \frac{\theta^2}{(j-1)^2} + \frac{\theta}{j-1}$.

Como las Y_j son independientes:

$$\Rightarrow Var(S_n) = \sum_{j=2}^{n} Var(Y_j) = \theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2} \tag{5.28}$$

La función generadora de momentos de S_n es

$$E\left(s^{S_n}\right) = \Pi_{j=2}^{n} E\left(s^{Y_j}\right) = \Pi_{j=2}^{n} \frac{j-1}{j-1+\theta(1-s)}$$

Por definición de función generadora de momentos:

$$F(s) = \sum_{n=0}^{\infty} \mathbb{P}(S_n = m) s^m$$

Por lo tanto, si se transforma el resultado en un polinomio, los coeficientes de grado m del mismo serán las $\mathbb{P}(S_n = m)$

$$F(s) = \prod_{l=0}^{n-1} \frac{l}{l+\theta(1-s)}$$

Observando que:

$$\frac{l}{l+\theta(1-s)} = \frac{l}{(l+\theta)\left(1-\frac{\theta 2}{l+\theta}\right)} = \frac{l}{l+\theta} \sum_{k=0}^{\infty} \left(\frac{\theta s}{l+\theta}\right)^k$$

Se obtiene:

$$\begin{aligned} F(s) = & \prod_{l=1}^{n-1} \tfrac{l}{l+\theta} \sum_{k=0}^{\infty} \left(\tfrac{\theta s}{l+\theta}\right)^k \\ = & \left(\prod_{l=1}^{n-1} \tfrac{l}{l+\theta}\right) \sum_{k_1}^{n-1} \left(\tfrac{\theta s}{1+\theta}\right)^{k_1} \sum_{k_2}^{n-1} \left(\tfrac{\theta s}{2+\theta}\right)^{k_2} \cdots \sum_{k_{n-1}}^{n-1} \left(\tfrac{\theta s}{n-1+\theta}\right)^{k_{n-1}} \\ = & \prod_{l=1}^{n-1} \tfrac{l}{l+\theta} \sum_{k_1}^{n-1} \sum_{k_2}^{n-1} \cdots \sum_{k_{n-1}}^{n-1} (\theta s)^{k_1+\cdots+k_{n-1}} \left(\tfrac{1}{1+\theta}\right)^{k_1} \cdots \left(\tfrac{n-1}{1+\theta}\right)^{k_{n-1}} \end{aligned} .$$

Por lo tanto:

$$\begin{aligned} \mathbb{P}(S_n = m) = & \theta^m \prod_{l=1}^{n-1} \tfrac{l}{l+\theta} \sum_{\substack{k_1,\ldots,k_{n-1} \\ k_1+\cdots+k_{n-1}=m}} \left[\left(\tfrac{1}{1+\theta}\right)^{k_1} \cdots \left(\tfrac{1}{n-1+\theta}\right)^{k_{n-1}}\right] \\ = & \theta^m (n-1)! \sum_{\substack{j_1,\ldots,j_{n-1} \\ j_1+\cdots+j_{n-1}=m+n-1}} \left[\left(\tfrac{1}{1+\theta}\right)^{j_1} \cdots \left(\tfrac{1}{n-1+\theta}\right)^{j_{n-1}}\right]. \end{aligned}$$

5.5. Estimadores de θ

5.5.1. Estimador de Waterson: θ_W

De la ecuación (5.27) se puede estimar θ de la siguiente manera:

$$\theta_W = \frac{S_n}{\sum_{j=1}^{n} \frac{1}{j}} \tag{5.29}$$

Por lo que se vio anteriormente este es un estimador insesgado de θ. Del cálculo de $Var(S_n)$ se tiene que:

$$Var(\theta_W) = \left[\theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2}\right] \left[\sum_{j=1}^{n-1} \frac{1}{j}\right]^{-2}$$

$\Rightarrow \lim_{n\to\infty} Var(\theta_W) = 0$, por lo tanto θ_W es un estimador débilmente consistente de θ.

5.5.2. Estimador de Tajima: θ_T

Otro estimador de θ es $\theta_T = \Pi_n$, cuya varianza está dada por la ecuación (5.8).

La distribución límite de Π_n tiene límite no degenerado, por lo tanto θ_T no puede ser consistente.

Si bien K_n es un estadístico suficiente, se vio que para estimar θ se necesita una muestra muy grande, ya que depende del logaritmo del tamaño de la misma. Los nuevos estadísticos tornan el uso de K_n ineficiente.

5.5.3. Comparación de los estimadores de Tajima y Watterson

La diferencia entre θ_W y θ_T se da cuando se tienen alelos con frecuencias bajas ya que los sitios segregantes no tienen en cuenta las frecuencias, pero el promedio de diferencias nucleotídicas sí lo hace. Cuando hay efectos selectivos en contra de uno o varios alelos, las frecuencias de éstos serán mucho menores a las de los alelos sobre los cuales la selección no tiene efectos negativos, por lo que estos estimandores permiten testear hipótesis respecto a la neutralidad de un modelo para una población dada. A su vez, habrá que tener en cuenta la historia de los tamaños poblacionales, ya que si la población no está en equilibrio es probable que también se obtengan

estimaciones diferentes del parámetro de mutación y no se podría rechazar la hipótesis de neutralidad, aunque si la de que la población esté en equilibrio [11].

Al observar las varianzas de los estimadores se tiene un indicio de que la correlación será grande cuando el tamaño de la muestra es chico (ya que las mutaciones tienen mayor incidencia en la frecuecia) y decrecerá lentamente a medida que el tamaño de la población crece.

Tajima [10] calculó la covarianza entre ambos estimadores dependiendo del tamaño de la población n, que se verá en la sección siguiente. Para testear la hipótesis de neutralidad desarrolló un test a partir de la diferencia existente entre ambos estimadores.

Wlasiuk, Garza y Lessa utilizan el test deasarrollado por Tajima para testear la hipótesis de neutralidad en la diferenciación de los Tucu-tucus del Río Negro (Ctenomys Rionegrensis)[14].

5.5.4. Covarianza entre el número de sitios segregantes y el promedio de diferencias nucleotídicas

Teorema 5.2.

$$cov(S_n, \Pi_n) = \theta + \left(\frac{1}{2} + \frac{1}{n}\right)\theta^2 [10]$$

Para calcular la covarianza, primero hay que observar que

$$cov\left(S_n, \Pi_n\right) = cov\left(S_n, \frac{1}{n(n-1)}\sum_{i\neq j}\Pi(i,j)\right) = cov(S_n, \Pi(i,j)). \quad (5.30)$$

Al igual que se hizo para calcular la varianza de Π_n, para calcular la covarianza hay que tener en cuenta la genealogía y el tamaño de la población.

Caso $n = 2$

Cuando $n = 2$, se tiene que $S_n = \Pi(i,j)$, por lo tanto $cov(S_n, \Pi(i,j)) = Var(\Pi(i,j)) = Var(S_n)$. De (5.28) y que $n = 2$, $\Pi(1,2) = \theta + \theta^2$, por lo tanto:

$$cov(S_n, \Pi(i,j)) = \theta + \theta^2 \quad (5.31)$$

Caso $n = 3$

La relación genealógica entre los 3 individuos está ilustrada por la figura 5.3 de la demostración de la varianza de Π_n (5.3.1). Se utilizará la misma notación que se empleó en dicha demostración.

Al tener tres secuencias en algún momento un par de ellas sufrirá un evento de coalescencia; las maneras en que esto puede ocurrir es que B sea el MRCA de C y D o que O sea el MRCA de los otros dos pares restantes. Por lo tanto la probabilidad que B sea el primer ancestro común es $1/3$ y que O lo sea $2/3$, de donde se obtiene:

$$cov(S_3, \Pi_n) = \frac{1}{3} cov(S_3, n_{CD}) + \frac{2}{3} cov(S_3, n_{CE})$$

En este caso se tiene: $S_3 = n_{BA} + n_{BC} + n_{BD} + n_{EA}$.

Como n_{BC}, n_{BD} y n_{EF} tienen la misma distribución:

$$cov(S_3, n_{CE}) = Var(n_{BA}) + cov(S_3, n_{CD}).$$

Sustituyendo por los resultados obtenidos en las ecuaciones (5.10), (5.13) y (5.14) se obtiene:

$$\begin{aligned} cov(S_3, n_{CD}) = & \ \tfrac{2}{3} Var(n_{BF}) + cov(S_3, n_{CD}) \\ = & \ 2Var(n_{BC}) + 4cov(n_{BC}, n_{BD}) \\ = & \ \tfrac{\theta}{3} + \tfrac{\theta^2}{6} \end{aligned}$$

Juntando las ecuaciones anteriores, se obtiene que:

$$\begin{aligned} cov(S_3, \Pi_3) = & \ \tfrac{2}{3} Var(n_{BA}) + cov(\Pi, n_{CD}) \\ = & \ \theta + \tfrac{5}{6}\theta^2 \end{aligned} .$$

Caso general

De $n-1$ secuencias se obtienen n en el momento en que una de las $n-1$ secuencias se bifurca (o sea tiene 2 hijos). Suponiendo que dicha bifurcación ocurre en el punto A de la figura y sus descendientes son B y C la covarianza está dada por la siguiente ecuación:

$$cov(S_n, \Pi_n) = \frac{1}{C_2^n} cov(S_n, n_{BC}) + \left(1 - \frac{1}{C_2^n}\right) cov(S_n, n_{ij}). \tag{5.32}$$

Donde el par ij es diferente del par BC.

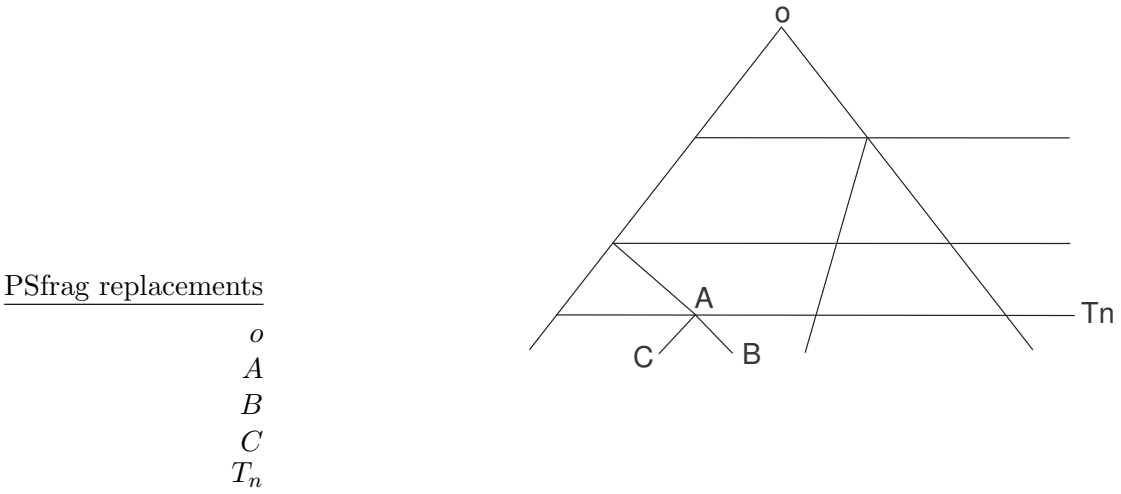


Figura 5.3: Esta es una genealogía posible para el caso $n = 5$

Para calcular $cov(S_n, n_{BC})$ se descompone S_n en segmentos, de la misma manera que se hizo en caso para $n = 3$. Como se vio anteriormente si hay segmentos en los caminos que corresponden a los pares de individuos (i, j) y (k, l) y éstos están en diferentes generaciones n_{ij} y n_{kl} son independientes, por lo tanto su covarianza es 0. De esto se deduce que para calcular la covarianza entre el número de sitios segregantes y el promedio de diferencias nucleotídicas sólo hay que tener en cuenta qué es lo que pasa con S_n a partir de la generación del individuo A. Por lo tanto lo que en realidad hay que calcular es la covarianza entre las ramas desde el momento T_n hasta el presente con n_{BC}. Haciendo esto y recordando que todos esos segmentos tienen la misma distribución se obtiene:

$$cov(S_n, n_{BC}) = Var(n_{BC}) + 2(n-2)cov(n_{AB}, n_{AC}) \tag{5.33}$$

y

$$cov(S_n, n_{ij}) = cov(S_{n-1}, \Pi_{n-1}) + Var(n_{BC}) + 2(n-2)cov(n_{AB}, n_{AC}) \tag{5.34}$$

Donde S_{n-1} y Π_{n-1} son el número de sitios segregantes y el promedio de diferencias nucloetídicas de $n-1$ secuencias. Sustituyendo (5.33) y (5.34) en (5.32) se obtiene:

$$cov(S_n, \Pi_n) = \frac{(n-1)(n-2)}{n(n-1)} cov(S_{n-1}, \Pi_{n-1}) + Var(n_{BC} + 2(n-2) cov(n_{AB}, n_{AC}) \tag{5.35}$$

Recordar entonces que:

$$\mathbb{E}(n_{AB}) = \mathbb{E}(\mathbb{E}(n_{AB}|T_n)) = \mathbb{E}(\frac{\theta}{2} T_n) = \frac{\theta}{2} \frac{1}{C_2^n} = \frac{\theta}{n(n-1)}$$

Combinando este razonamiento con (5.13) se tiene que:

$$\mathbb{E}(n_{AB}^2) = \frac{\theta}{2}\mathbb{E}(T_n) + \frac{\theta^2}{4}\mathbb{E}(T_n^2) = \frac{\theta}{n(n-1)} + \frac{\theta^2}{4} \cdot 2 \cdot \left(\frac{1}{C_n^2}\right)^2$$

$$\mathbb{E}(n_{AB}^2) = \frac{\theta}{n(n-1)} + \frac{2\theta^2}{(n(n-1))^2}$$

De donde:

$$Var(n_{AB}) = \frac{\theta}{n(n-1)} + \left(\frac{\theta}{n(n-1)}\right)^2$$

Siguiendo ese razonamiento también se obtiene:

$$cov(n_{AB}, n_{AC}) = \left(\frac{\theta}{n(n-1)}\right)^2 \tag{5.36}$$

Además:

$$Var(n_{BC}) = 2Var(n_{AB}) + 2cov(n_{AB}, n_{AC}) = \frac{2\theta}{n(n-1)} + \left(\frac{2\theta}{n(n-1)}\right)^2 \tag{5.37}$$

Sustituyendo (5.36) y (5.37) en (5.35):

$$cov(S_n, \Pi_n) = \frac{(n+1)(n-2)}{n(n-1)} cov(S_{n-1}, \Pi_{n-1}) + \frac{2\theta}{n(n-1)} + \frac{2\theta^2}{n(n-1)^2} \tag{5.38}$$

Utilizando el principio de inducción completa se termina la demostración. Para el caso $n = 2$ la fórmula se cumple ya que $cov(S_2, \Pi_2) = \theta + \left(\frac{1}{2} + \frac{1}{2}\right)\theta^2$ y en (5.31) se vio que $cov(S_2, \Pi_2) = \theta + \theta^2$

Además

$$H) \quad cov(S_{n-1}, \Pi_{n-1}) = \theta + \left(\frac{1}{2} + \frac{1}{n-1}\right)\theta^2$$

$$T) \quad cov(S_n, \Pi_n) = \theta + \left(\frac{1}{2} + \frac{1}{n}\right)\theta^2$$

Demostración del paso inductivo: De (5.38) se sabe que
$cov(S_n, \Pi_n) = \frac{(n+1)(n-2)}{n(n-1)} cov(S_{n-1}, \Pi_{n-1}) + \frac{2\theta}{n(n-1)} + \frac{2\theta^2}{n(n-1)^2}$
Usando la hipótesis
$cov(S_n, \Pi_n) = \frac{(n+1)(n-2)}{n(n-1)} \left(\theta + \left(\frac{1}{2} + \frac{1}{n-1}\right)\theta^2\right) + \frac{2\theta}{n(n-1)} + \frac{2\theta^2}{n(n-1)^2}$

$$\begin{aligned}
\Rightarrow cov(S_n, \Pi_n) = & \frac{(n+1)(n-2)}{n(n-1)} \left[\theta + \left(\tfrac{1}{2} + \tfrac{1}{n}\right)\theta^2\right] + \frac{2\theta}{n(n-1)} + \frac{2\theta^2}{n(n-1)^2} \\
= & \frac{n^2-n-2+2}{n(n-1)}\theta + \left[\frac{(n+1)(n-2)}{2n(n-1)} + \frac{(n+1)(n-2)}{n(n-1)^2} + \frac{2}{n(n-1)^2}\right]\theta^2 \\
= & \theta + \left[\frac{(n+1)(n-2)(n-1)}{2n(n-1)} + \frac{2(n+1)(n-2)}{n(n-1)^2} + \frac{4}{n(n-1)^2}\right]\theta^2 \\
= & \theta + \left[\frac{n^3-3n+2}{2n(n-1)^2}\right]\theta^2 \\
= & \theta + \left[\frac{n+2}{2n}\right]\theta^2 \\
= & \theta + \left[\frac{1}{2} + \frac{1}{n}\right]\theta^2
\end{aligned}$$

♠

Observación 5.5.1. A medida que n crece, la covarianza se acerca a $\theta + \frac{1}{2}\theta^2$. Esta covarianza es llamada *covarianza estocástica.*

Definición 5.8. $cov_{st}(S_n, \Pi_n) = \theta + \frac{1}{2}\theta^2$

Definición 5.9. Covarianza de la muestra

$$cov_s(S_n, \Pi_n) = cov(S_n, \Pi_n) - cov_{st}(S_n, \Pi_n) = \frac{1}{n}\theta^2$$

Definición 5.10. Coeficiente de correlación (r)

$$r = \frac{cov(S_n, \Pi_n)}{\sqrt{Var(S_n)Var(\Pi_n)}}$$

Según Tajima [10] numéricamente se ve que si la muestra es chica el coeficiente es grande y disminuye a medida que n crece.

Apéndice

.1. Teorema Central del Límite

Teorema .3. *Teorema Central del Límite, versión de Lindeberg*

Sean $\xi_1, \ldots, \xi_i, \ldots$ *variables aleatorias independientes tales que* $E(\xi_k) = 0$ *y* $var(\xi_k) = \sigma_k^2 < \infty$. *Se definen* $S_0 = 0$, $S_n = \sum_{i=1}^n \xi_i$ *y* $V_n = \sqrt{\sum_{k=1}^n \sigma_k^2}$
Si se cumple

$$\forall \gamma > 0,\ \theta_n(\gamma) = \frac{1}{V_n^2} \sum_{i=1}^{n} E(\xi_i^2 \mathbb{I}_{\{|\xi_i| \geq \gamma V_n\}}) \to 0 \ si\ n \to \infty$$

Entonces

$$\frac{S_n}{V_n} \to \mathcal{N}(0,1)$$

.2. Cadenas de Markov absorbentes

Definición .11. *Un* estado s_i *de una cadena de Markov se dice* absorbente *si es imposible de dejarlo (e.g.* $p_{ii} = 1$ *).*

Una cadena de Markov *se dice* absorbente *si teine por lo menos un estado absorbente y si desde cada uno es posible llegar a uno absorbente, aunque sea en más de un paso.*

Definición .12. *En una cadena de Markov absorbente, los estados que no lo son, se llaman estados* transitorios

Forma Canónica

Dada una cadena de Markov absorbente, si reenumeramos los estados de manera que los primeros t sean los transitorios y los últimos r los absorbentes, la matriz de transición tendrá la siguiente *forma canónica*:

$$P = \begin{array}{c} \\ \text{TR.} \\ \text{ABS.} \end{array} \begin{array}{c} \begin{array}{cc} \text{TR.} & \text{ABS.} \end{array} \\ \left(\begin{array}{c|c} \text{Q} & \text{R} \\ \hline 0 & I \end{array}\right) \end{array}$$

I es la identidad de dimensión $r \times r$, 0 es una matriz $r \times t$ de ceros, R es una matriz no nula $t \times r$ y Q es una matriz $t \times t$.

Haciendo cuentas, resulta que

$$P = \begin{array}{c} \\ \text{TR.} \\ \text{ABS.} \end{array} \begin{array}{c} \begin{array}{cc} \text{TR.} & \text{ABS.} \end{array} \\ \left(\begin{array}{c|c} \text{Q}^n & * \\ \hline 0 & I \end{array}\right) \end{array}$$

Donde * es una submatriz que se puede escribir en términos de Q y R.

Probabilidad de absorción

Teorema .4. *En una cadena de Markov absorbente, la probabilidad que el proceso sea absorvido es 1. O sea* $Q^n \to 0$, *cuando* $r \to \infty$

Dem.: Desde cada estado transitorio s_j es posible alcanzar cualquier estado absorbente, por definición. Sea m_j el menor número de pasos requeridos para el alcanzar un estado absorbente, empezando del s_j. Sea p_j la probabilidad de empezando el estado s_j el proceso no sea absorvido en m_j pasos. $p_j < 1$. Sea m el mayor de los m_j y sea p el mayor de los p_j.

La probabilidad de no ser absorvido en m pasos es menor o igual que p, en $2m$ pasos es menor o igual que p^2, etc. Como $p < 1$ estas probabilidad

tiende a 0. Como la probabilidad de no ser absorvido en n pasos es monótona decreciente, estas probabilidades también tienden a 0, por lo tanto $\lim_{n\to\infty} Q^n = 0$. ♠

La matriz fundamental

Teorema .5. *Para una cadena de Markov absorbente, la matriz $I - Q$ es inveritble.*

Dem.: Sea $x/(I-Q)x = 0$.
$\Rightarrow x = Qx \;\Rightarrow x = Q^n x$.
Como $Q^n \to 0$, se tiene que $Q^n x \to 0$, por lo tanto $x = 0$. ♠

Bibliografía

[1] Durret, R. *Probability Models for DNA Sequence Evolution*, Segunda edición, 2008.

[2] Ewens, W.J. *The sampling theory of selectively neutral alleles.* Theor. Pop. Biol. 3, 87112, 1972.

[3] Feller, W. *An Introduction to Probability Therory and its Aplications*, volumen I, Princeton, Tercera edición, 1968.

[4] Gillespie, J. H. *Population genetics- a concise guide.* Johns Hopkins Univ. Press, Baltimore, 1998.

[5] Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S., and Clegg, J.B. *Archaic African and Asian lineages in the genetic ancestry of modern humans.* Am. J. Human Genetics, 1997.

[6] Hartl, D. L. y Clark, A. G. *Principles of population genetics*, Sunderland Assoc., Sunderland, MA, 1997.

[7] Kimura, M. y Crow, J.F. *The number of allels that can be maintained in a finite population.* Genetics 49, 725-738, 1964.

[8] Lessa, E.P. *Guía de Estudio de Genética de Poblaciones.*, Laboratorio de Evolución de Facultad de Ciencias, Montevieo, Uruguay, 2001.

[9] Tajima, F. *Evolutionary relationship of DNA sequences in finite populations.* Genetics 105:437-460, 1983.

[10] Tajima, F. *Statistical Method for Testing the Neutral Mutation Hypotesis by DNA Polymorphism.* Genetics 123, 1989: 585-595.

[11] Tajima, F. *The Effect of Change in Population Size on DNA Polymorphism.* Genetics 123, 1989: 597-601.

[12] Tavare, S. *Ancestral Inference in Population Genetics.* Springer Lecture Notes in Mathematics. 2003.

[13] Watterson, G. A. *On the number of segregating sites in genetical models without recombination.* Theor. Pop. Biol., 1975.

[14] Wlasiuk, G., Garza, J.C. y Lessa, E.P. *Genetic and Geographic Diferentiation in the Rio Negro Tuco-Tuco (Ctenomys Rionegrensis): Inferring the Roles of Migration and Drift From Multiple Genetic Markers.* Evolucion 57(4), 2003: 913-926.