Trabajo monográfico

Estadísticos F en Genética de Poblaciones

Micaela Long

Orientadora:
Dra. María Inés Fariello

Noviembre de 2023

Licenciatura en Matemática
Facultad de Ciencias
Universidad de la República
Montevideo, Uruguay

Resumen

El estudio de los patrones de variación genética compartida entre conjuntos de poblaciones contribuye a la comprensión de la historia de nuestra especie, pues cada gran evento demográfico deja una huella en la diversidad genética de una población.

Un marco metodológico útil para este análisis es el de los estadísticos F, que miden correlaciones en las frecuencias alélicas entre conjuntos de dos (F_2), tres (F_3) o cuatro (F_4) poblaciones. Los valores observados reflejan grados de ascendencia compartida, que pueden utilizarse para probar hipótesis sobre el árbol o grafo que relaciona a las poblaciones, los órdenes de división y los eventos de flujo de genes en el pasado, bajo distintos modelos históricos.

El objetivo de esta monografía es estudiar los estadísticos F, su relación con algunas medidas de divergencia entre poblaciones y su representación en un árbol o grafo poblacional, con énfasis en las herramientas proporcionadas por la teoría del coalescente. Además, estudiaremos su interpretación geométrica en el espacio de frecuencias alélicas y en el de componentes principales, lo que lleva a resultados útiles en el contexto de análisis y visualización de datos genómicos.

Este trabajo está basado principalmente en los artículos *"Admixture, Population Structure and F-Statistics"* y *"A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis"*, cuyo autor es Benjamin M. Peter.

Índice general

Capítulo 1

Introducción

Dos humanos difieren, en promedio, en aproximadamente 1 de cada 1.000 pares de bases de ADN (0.1 %) [35]. Alrededor del 85 % de esta variación se debe a diferencias entre individuos dentro de las poblaciones, mientras que el 15 % puede ser explicada por la estructura poblacional.

La homogeneidad de los humanos a nivel de ADN, así como la escasa variación entre poblaciones implican que las diferencias genéticas entre individuos y poblaciones no proporcionan una base biológica o taxonómica para ninguna forma de discriminación [14]. Sin embargo, el procentaje de variación presente entre poblaciones puede aprovecharse para estudiar en detalle nuestra historia y diversidad.

La forma más básica de representar la historia evolutiva de un conjunto de poblaciones es a través de un árbol filogenético, un modelo que en su sentido estricto supone que no hay flujo génico entre poblaciones después de los eventos de divergencia [15]. Sin embargo, en muchas ocasiones los grupos que se han separado pueden intercambiar genes. Este es ciertamente el caso de la historia de la especie humana, durante la cual las poblaciones han divergido de forma incompleta, o bien divergieron y posteriormente se mezclaron.

Un modelo alternativo es un grafo de mezcla, que generaliza el árbol poblacional al permitir aristas que representan la fusión de poblaciones o un intercambio significativo de migrantes.

Estos modelos proveen una descripción concisa de las relaciones histórico-demográficas entre poblaciones, asumiendo que estas relaciones son producto de eventos instantáneos y discretos de divisiones y mezcla.

Las últimas dos décadas han sido testigos de una explosión de datos genéticos, que proporcionan información útil para la comprensión de los patrones de diversidad entre individuos y poblaciones, y han motivado el desarrollo de herramientas informáticas para su análisis.

Si bien existen varios enfoques para construir árboles de poblaciones que incorporen eventos de mezcla a partir de los datos, muchos de estos métodos (como los basados en máxima verosimilitud) tienen dificultades en relación al costo computacional y la cantidad de poblaciones que pueden manejar. Además, muchas veces requieren de la especificación previa de la topología del árbol, por lo que sólo se pueden inferir de forma automática los valores de los parámetros, no la disposición de las poblaciones en el árbol [15].

Dado un conjunto de poblaciones, con el objetivo de inferir la topología del árbol o grafo que

las relaciona nos podemos preguntar: ¿están las poblaciones relacionadas en forma de árbol?, ¿desciende una población particular de varias poblaciones ancestrales?, ¿en qué proporciones contribuyen diferentes poblaciones a una determinada población?, ¿cuál es la población más cercana a una población dada? Estas preguntas nos llevan a interrogantes más fundamentales: ¿cómo definimos la divergencia genética entre poblaciones?, ¿cómo la medimos?, ¿qué noción de distancia utilizamos?

Un enfoque que ha ganado popularidad es el de los estadísticos F, presentados por David Reich et. al. en 2009 [28] y Nick Patterson et. al en 2012 [22].

Los estadísticos F son un conjunto de herramientas basadas en momentos, que sólo utilizan medias y varianzas de divergencias, donde la divergencia aquí se define como la diferencia en las frecuencias alélicas de distintas poblaciones. Específicamente, los estadísticos miden correlaciones en las frecuencias alélicas entre conjuntos de dos (F_2), tres (F_3) o cuatro (F_4) poblaciones. Los valores observados reflejan grados de ascendencia compartida, que pueden utilizarse para probar hipótesis sobre los órdenes de división de las poblaciones y los eventos de flujo de genes en el pasado bajo distintos modelos históricos.

Un supuesto implícito en el desarrollo de los estadísticos F ha sido que las poblaciones son discretas, y que el flujo génico es raro. Recientemente se mostró que eso no es necesario, y que es posible interpretar a las poblaciones como vectores en el espacio S-dimensional de frecuencias alélicas $\mathbb{R}^S$. Los estadísticos F tendrán una interpretación geométrica en este espacio, que está dada por las proyecciones ortogonales [20].

Una de las dificultades es que los conjuntos de datos suelen ser de gran tamaño, y en muchos casos están compuestos por unos pocos miles de individuos (de unas pocas poblaciones), genotipados para cientos de miles de sitios en el ADN, por lo que los datos se vuelven dispersos. Sin embargo, los procesos históricos que generan varación genética no escapan de la hipótesis de la variedad: si bien la dimensión original en la que viven los datos es alta, usualmente resultan en matrices de datos de bajo rango. En este sentido, una de las técnicas más utilizadas es el análisis de componentes principales (PCA), pues permite aproximar el conjunto de datos a través de un subespacio de dimensión menor, minimizando la pérdida de información. Las distintas propiedades y aplicaciones de los estadíticos F podrán derivarse en el espacio de componentes principales, y en particular en el de dimensión 2 (plano), lo que lleva a resultados útiles en el contexto de análisis y visualización de datos genómicos.

El presente trabajo tiene como objetivo estudiar los estadísticos F, y está basado en los artículos *"Admixture, Population Structure and F-Statistics"*, publicado en 2016 [23] y *"A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis"*, publicado en 2022 [24]. En ambos casos el autor es el biólogo evolutivo Benjamin M. Peter.

Comenzaremos desarrollando la teoría necesaria para definir los estadísticos F. Para esto, en el capítulo 2 introduciremos conceptos básicos de genética seguidos por los modelos clásicos de genética de poblaciones: Hardy-Weinberg, Wright-Fisher y la teoría del Coalescente. Por último presentaremos el estadístico más conocido y utilizado para medir la divergencia entre poblaciones: el índice de fijación F_{ST}.

El capítulo 3 está dedicado a presentar las principales propiedades y aplicaciones de los estadísticos F, y su relación con algunas medidas de divergencia entre poblaciones: varianza de frecuencias alélicas, heterocigosidad y tiempos esperados de coalescencia.

Por último, en el capítulo 4 veremos la interpretación geométrica de los estadísticos en el espacio de frecuencias alélicas y en el de componentes principales. En particular, veremos cómo se

manifiestan algunas de las propiedades teóricas presentadas en el capítulo 3, con énfasis en el plano.

Capítulo 2

Genética de poblaciones

2.1. Conceptos de genética

La información hereditaria de la mayoría de los seres vivos está contenida en el ácido desoxirribonucleico (ADN) [7]. El ADN consiste en dos cadenas complementarias que forman una estructura de doble hélice. Cada cadena es una secuencia de *nucleótidos*, moléculas pequeñas sintetizadas por todos los organismos, formadas por la unión de tres elementos: una base nitrogenada, un azúcar simple (desoxirribosa, dR) y un grupo fosfato (P). Las bases nitrogenadas presentes en el ADN son adenina (A), guanina (G), citosina (C) y timina (T).

Lo que distingue una molécula de ADN de otra es la secuencia de las bases nitrogenadas, razón por cual se hace referencia a una secuencia de ADN nombrando sólo sus bases. La disposición secuencial de estas cuatro bases a lo largo de la cadena es la que codifica la información genética.

En su forma de doble hélice, las dos hebras están unidas por puentes de hidrógeno entre pares de bases. Debido a la estructura química, la adenina (A) se aparea con la timina (T) mediante dos puentes de hidrógeno, mientras que la guanina (G) se aparea con la citosina (C) mediante un enlace triple.

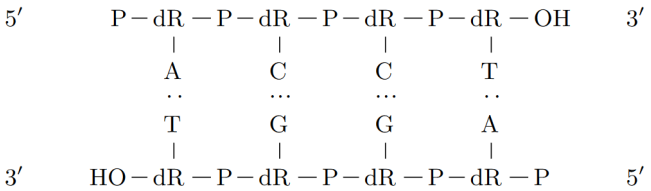

Figura 2.1: Estructura del ADN. Imagen tomada de [7].

Una consecuencia de la estructura de los nucleótidos es que las cadenas de ADN tienen orientación (direccionalidad). Esto se debe a que la desoxirribosa ($C_5H_{10}O_4$) contiene cinco átomos de carbono, numerados como 1´, 2´, 3´, 4´y 5´. Al inicio de la cadena (extremo 5´) sobresale el grupo fosfato (P) unido al carbono 5´ del primer nucleótido. En el otro extremo (extremo 3´) está expuesto el hidroxilo (OH) unido al carbono 3´ del último nucleótido. En la doble hélice, la

dirección de los nucleótidos en una hebra ($3´ \rightarrow 5´$) es opuesta a la dirección en la otra hebra ($5´ \rightarrow 3´$). Es decir, las cadenas son antiparalelas (paralelas pero con direcciones opuestas). Por tanto, el ADN no es una secuencia de palabras elegidas al azar de un alfabeto de cuatro letras, la química juega un rol importante.

En muchos organismos la mayor parte del ADN se encuentra en el núcleo de las células y se organiza en estructuras llamadas *cromosomas*. Las células de cada especie poseen un número característico de cromosomas.

Cada cromosoma contiene una copia del material genético. Los organismos *haploides* son aquellos que tienen sólo una copia de su material genético, es decir, un juego de cromosomas. Los organismos *diploides* (como los seres humanos) son aquellos que tienen dos copias. Existen organismos tetraploides (cuatro copias), hexaploides (seis copias) y poliploides (muchas copias). Por ejemplo, el *genoma humano* es la secuencia de ADN contenida en 23 pares de cromosomas en el núcleo de cada célula humana diploide. Y cada juego de 23 cromosomas contiene más de 3.000 millones de pares de bases de ADN.

La *herencia genética* es el proceso mediante el cual las características de los progenitores se transmiten a sus descendientes. La unidad fundamental de la herencia es el *gen*. Si bien su definición varía según el contexto, llamaremos gen al factor hereditario que determina una característica. Molecularmente, un gen es una secuencia de nucleótidos contiguos en la molécula de ADN, o en la molécula de ácido ribonucleico (ARN) en el caso de algunos virus.

Para que los genes se transmitan a los descendientes es necesaria una reproducción idéntica que dé lugar a una réplica de cada uno de ellos. En la reproducción celular los cromosomas se separan a través de los procesos de *mitosis* y *meiosis*. Estos garantizan que cada célula hija de un organismo determinado reciba una dotación cromosómica completa. La mitosis consiste en la separación de los cromosomas duplicados durante la división de las células *somáticas* (células no sexuales). La meiosis consiste en el apareamiento y la separación de los cromosomas duplicados durante la división de las células *sexuales*, para producir gametos (células reproductoras).

Las variaciones que se producen en el ADN de un individuo se denominan *variaciones genotípicas*. Estas generan versiones distintas de un mismo gen, denominadas *alelos*. Por ejemplo, un gen para el color del pelaje de los gatos puede existir como alelos que codifiquen el pelaje negro o el pelaje naranja. Las variaciones genotípicas surgen por *mutaciones* en el ADN. Una mutación es un cambio aleatorio en la secuencia de nucleótidos o en la organización del ADN (o ARN). Las mutaciones que se producen en los genes de las células sexuales pueden transmitirse de una generación a otra.

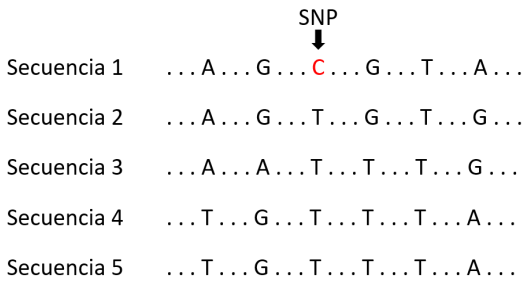

Figura 2.2: Cinco secuencias de ADN.

Los *polimorfismos de un solo nucleótido* (SNPs) son las variaciones genéticas más usuales entre

humanos (figura 2.2). Cada SNP es una variación en la secuencia de ADN que afecta a una sola base (A, C, G o T). Una variación puntual debe darse en al menos un 1 % de la población para ser considerada SNP.

Todos los alelos de un gen particular se localizan en un sitio específico en el cromosoma, denominado *locus*, cuyo plural es *loci*. El *genotipo* es el conjunto de alelos que posee un individuo. Un organismo diploide que posee los dos alelos idénticos se dice *homocigota* para ese locus. Uno que posee dos alelos diferentes es *heterocigota* para ese locus.

La *expresión génica* es el proceso mediante el cual todos los organismos transforman la información codificada por los ácidos nucleicos (ADN o ARN) en las proteínas necesarias para su desarrollo, funcionamiento y reproducción. Este proceso consta de una serie de pasos entre los que se encuentran la *transcripción* y *traducción*.

La transcripción consiste en la producción de copias de ARN mensajero (ARNm) a partir de la secuencia de ADN, y es llevada a cabo por la ARN polimerasa. En este paso la doble hélice de ADN se separa y cada cadena sirve como molde para la síntesis de una nueva cadena complementaria. Esta copia de ARNm transporta la información necesaria para la elaboración de proteínas desde el núcleo de la célula hacia el citoplasma, donde ocurre la traducción.

La traducción es el proceso que convierte una secuencia de ARNm en una cadena de *aminoácidos*, moléculas que se combinan para formar las proteínas. Aunque existen excepciones, este es el dogma central de la biología molecular: el ADN se transcribe en ARN, que a su vez se traduce en proteínas.

Los genes codifican *fenotipos*: manifestaciones o apariciones de una característica. Un fenotipo puede referirse a una característica física, fisiológica, bioquímica o conductual. El genotipo determina el potencial para el desarrollo y establece ciertos límites o fronteras al mismo. Un fenotipo determinado surge de un genotipo que se desarrolla dentro de un ambiente en particular.

La genética de poblaciones explora la composición genética de grupos de la misma especie llamados *poblaciones*, y cómo esta composición varía con el tiempo y espacio geográfico. Una población es un grupo de individuos de la misma especie que habitan en un área geográfica restringida que les permite reproducirse entre sí.

La diversidad y la adaptación de la vida son producto de la *evolución*: el cambio genético a través del tiempo. La evolución puede verse como un proceso de dos pasos. Primero surgen variantes genéticas al azar, y luego, la frecuencia de estas variantes aumenta o disminuye. Dado que la evolución es el cambio genético, la genética de poblaciones es fundamentalmente el estudio de la evolución.

Esta rama de la genética se diferencia de gran parte de la biología en que sus conocimientos son teóricos más que experimentales u observacionales. La evolución es la variación de las frecuencias genotípicas a través del tiempo, y mientras que las frecuencias genotípicas son fácilmente medibles, su variación no lo es. La escala de tiempo es del orden de decenas de miles a millones de años, por lo que los cambios son imposibles de ver de forma directa. Si bien podemos observar el estado de una población, no tenemos cómo explorar directamente su evolución.

Es por esto que entre las herramientas fundamentales de la genética de poblaciones se encuentran los modelos matemáticos, representaciones simplificadas que en general describen los procesos en términos de ecuaciones, y donde los factores que influyen en un proceso están representados por variables en estas ecuaciones.

2.2. Modelo de Hardy-Weinberg

En 1908 el matemático inglés Godfrey H. Hardy y el genetista y médico alemán Wilhelm Weinberg formularon de forma independiente un modelo genético poblacional que lleva sus nombres y ocupa un lugar central en la genética de poblaciones [9]. La ley de Hardy-Weinberg es un principio que establece que, bajo ciertas condiciones, las frecuencias alélicas y genotípicas de un locus particular en una población se mantendrán constantes a lo largo del tiempo.

El modelo asume los siguientes dos grupos de hipótesis en la población:

Grupo 1.

- Organismos diploides
- Reproducción sexuada
- Generaciones no solapantes
- Gen autosómico
- Frecuencias alélicas no dependientes del sexo

Grupo 2.

- Población de tamaño infinito
- Apareamientos al azar
- Ausencia de migración desde otras poblaciones
- Ausencia de mutación
- Selección natural no opera sobre el gen considerado

El motivo de separación de los supuestos en dos grupos obedece a que si las hipótesis del Grupo 1 son modificadas, podemos obtener de forma directa un modelo que se adapte a la nueva situación [13]. Sin embargo, la modificación de las hipótesis del Grupo 2 lleva a otros modelos, algunos de los cuales veremos más adelante en este trabajo.

Una importante propiedad de este modelo es la separación de cada generación en dos fases:

- Fase gamética, en la que cada locus está representado por un solo alelo.
- Fase diploide u organísmica, en la que cada locus está representado por dos alelos.

Consideremos un locus con dos alelos A_1 y A_2 en una población que cumple las hipótesis del modelo. Los genotipos posibles son A_1A_1, A_1A_2, A_2A_2. Supongamos que la frecuencia del alelo A_1 es p y la frecuencia del alelo A_2 es $q = 1 - p$.

Para formar un cigoto en una generación siguiente, la hipótesis de apareamiento al azar implica que elegimos aleatoriamente dos gametos de la generación parental. La probabilidad de que un cigoto sea homocigota A_1A_1 es la probabilidad de elegir A_1 dos veces. Como el apareamiento es aleatorio y el tamaño poblacional es infinito, esta probabilidad es p^2. Análogamente, la probabilidad de que se forme un cigoto A_2A_2 es q^2.

Ahora bien, hay dos maneras de formar un heterocigota A_1A_2. La primera es que A_1 corresponda al gameto femenino, y A_2 al gameto masculino, evento con probabilidad pq. La segunda es que A_2 corresponda al gameto femenino, y A_1 al masculino, evento con probabilidad qp. Por tanto, la probabilidad de obtener A_1A_2 es $2pq$.

Luego de una ronda de apareamiento tenemos las siguientes frecuencias genotípicas

Genotipo:	A_1A_1	A_1A_2	A_2A_2
Frecuencia (H-W):	p^2	$2pq$	q^2

Las frecuencias genotípicas esperadas bajo el modelo de Hardy-Weinberg dependen sólo de las frecuencias alélicas.

A partir de las frecuencias genotípicas podemos calcular las frecuencias alélicas

$$\text{frec}(A_1) = \frac{2\ \#\text{individuos } A_1A_1 + \ \#\text{individuos } A_1A_2}{2\ \#\text{individuos}} = \text{frec}(A_1A_1) + \frac{1}{2}\text{frec}(A_1A_2)$$

$$\text{frec}(A_2) = \frac{2\ \#\text{individuos } A_2A_2 + \ \#\text{individuos } A_1A_2}{2\ \#\text{individuos}} = \text{frec}(A_2A_2) + \frac{1}{2}\text{frec}(A_1A_2)$$

Por lo que luego de la ronda de apareamiento, las frecuencias alélicas en la fase gamética serán

$$\text{frec}(A_1) = p^2 + \frac{1}{2}2pq = p^2 + p(1-p) = p$$

$$\text{frec}(A_2) = q^2 + \frac{1}{2}2pq = q^2 + (1-q)q = q$$

Por tanto, las frecuencias alélicas se mantienen constantes. Como las frecuencias alélicas determinan las genotípicas, éstas también se mantienen invariantes.

Generalización para k alelos.

El razonamiento anterior puede ser extendido al caso en el que tenemos un locus con k alelos distintos $A_1, \cdots, A_k$. Si p_i es la frecuencia del alelo A_i en la población, tenemos que

$$p_1 + \cdots + p_k = 1$$

En una ronda de reproducción, la probabilidad de formar un cigoto A_iA_i es p_i^2, con $i = 1, \cdots, k$.

Por otro lado, siguiendo el razonamiento realizado para 2 alelos, la probabilidad de formar un cigoto A_iA_j con $i \neq j$ es $2p_ip_j$. Es decir

$$\text{frec}(A_iA_j) = (2 - \delta_{ij})p_ip_j \quad \text{con } \delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$$

Luego

$$\begin{aligned} \text{frec}(A_1) &= \text{frec}(A_1A_1) + \frac{1}{2}\text{frec}(A_1A_2) + \cdots + \frac{1}{2}\text{frec}(A_1A_k) \\ &= p_1^2 + \frac{1}{2}2p_1p_2 + \cdots + \frac{1}{2}2p_1p_k \\ &= p_1 \underbrace{(p_1 + \cdots + p_k)}_{1} = p_1 \end{aligned}$$

El cálculo es análogo para A_i con $i = 2, \cdots, k$. Es decir, en el caso general las frecuencias alélicas se mantienen constantes y las genotípicas quedan determinadas por las alélicas.

El modelo de Hardy-Weinberg es un modelo matemático que simplifica el comportamiento reproductivo en una población ideal. Es claro que los supuestos del modelo son violados en las poblaciones naturales. Sin embargo, algunas violaciones pueden tener un efecto mínimo, lo que hace que las poblaciones estén en equilibrio de Hardy-Weinberg.

Por otro lado, la utilidad del modelo se debe precisamente a su falta de realismo, ya que es utilizado como hipótesis nula. Constatar desviaciones significativas de lo esperado según el modelo de Hardy-Weinberg es indicio de que algo, que no se ajusta a lo asumido en el modelo, está sucediendo en la población de estudio [13]. Por ejemplo, el modelo de Hardy-Weinberg provee predicciones sobre el comportamiento de un gen sobre el que la selección no opera. Si sospechamos que la selección actúa sobre cierto gen, podemos comenzar mostrando que dicho gen se aparta de las predicciones del modelo.

La ley de Hardy-Weinberg indica que, cuando se cumplen los supuestos del modelo, la reproducción sola no altera las frecuencias alélicas ni genotípicas, y las frecuencias alélicas determinan las frecuencias de los genotipos. Como la evolución es la variación en las frecuencias alélicas, la ley de Hardy-Weinberg nos dice que la reproducción sola no provocará evolución. Para que las poblaciones evolucionen se requieren otros procesos: *selección*, *mutación*, *migración* y *azar*. El último de ellos actúa de forma inmediata cuando quitamos el supuesto de tamaño poblacional infinito, y lleva el nombre de *deriva genética*.

2.3. Deriva genética

Uno de los supuestos del modelo de Hardy-Weinberg es que el tamaño de la población es infinito. Si bien este supuesto puede ser razonable para aproximar el comportamiento en poblaciones de gran tamaño, no se ajusta a todas [25].

En poblaciones finitas, la variación en la cantidad de descendencia entre individuos resulta en variaciones aleatorias en las frecuencias alélicas, fenómeno al que llamamos *deriva genética*.

A continuación veremos el modelo de Wright-Fisher, que incorpora la deriva genética al asumir que la población es finita.

2.3.1. Modelo de Wright-Fisher

Uno de los modelos más importantes de la genética de poblaciones fue introducido por el biólogo y matemático Ronald Fisher en 1930, y por el genetista Sewall Wright en 1931.

Este modelo de reproducción provee una descripción dinámica de la transmisión de alelos de una generación a la siguiente, y de la evolución de una población bajo los mismos supuestos que el modelo de Hardy-Weinberg, excepto el del tamaño poblacional. En este contexto la población tendrá tamaño finito y constante [13].

Consideremos un locus con dos alelos A_1 y A_2 en una población de tamaño constante N, en la que se cumplen el resto de los supuestos del modelo de Hardy-Weinberg.

Los individuos diploides tienen dos copias de su material genético en cada célula. Trataremos a los N individuos como $2N$ copias de un locus (es decir, no juntaremos las copias para formar

individuos). Esto es análogo a considerar una población de $M = 2N$ individuos, con M individuos haploides.

Podemos representar el estado de la población en la generación n como una urna que contiene $2N$ bolillas: una cantidad i son de tipo A_1, y una cantidad $2N - i$ son de tipo A_2 [7]. Para construir la generación $n + 1$, sacamos $2N$ bolillas de la urna con repetición.

En el modelo haploide todos los alelos se eligen de forma independiente, mientras que en el diploide el segundo alelo elegido debe ser de un parental distinto que el primero. Sin embargo, los modelos son probabilísticamente similares para N grande, por lo que asumiremos independencia en la elección de alelos.

La probabilidad de que en la generación $n+1$ haya j alelos de tipo A_1, dado que en la generación n hay i, se calcula mediante la distribución binomial:

$$p(i,j) = \binom{2N}{j} {p_i}^j (1-p_i)^{2N-j} \tag{2.1}$$

donde $p_i = i/2N$ es la probabilidad de sacar un alelo de tipo A_1 en la urna dado que hay i, y el número de formas de elegir j elementos tomados de $2N$ es el coeficiente binomial

$$\binom{2N}{j} = \frac{(2N)!}{j!(2N-j)!}$$

Sea X_n la variable aleatoria que cuenta la cantidad de alelos de tipo A_1 en la generación n. El proceso estocástico $\{X_n : n \in \mathbb{N}\}$ es una *cadena de Markov* homogénea en el tiempo, con espacio de estados $E = \{0, \cdots, 2N\}$, y matriz de transición de entradas $p(i,j)$. En efecto, para todo $\{i_0, \cdots, i_{n+1}\} \subset E$ se tiene

$$\begin{aligned}
\mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n, \cdots, X_0 = i_0) &= p(i_n, i_{n+1}) \\
&= \binom{2N}{i_{n+1}} \cdot {p_{i_n}}^{i_{n+1}} (1-p_{i_n})^{2N-i_{n+1}} \\
&= \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i)
\end{aligned}$$

Como suponemos ausencia de mutación y migración, una vez que un alelo desaparece en la población, no vuelve a aparecer. Por tanto, los estados 0 y $2N$ son estados absorbentes para X_n

$$p(2N, 2N) = p(0,0) = 1$$

Como el espacio de estados E es finito, la probabilidad de absorción es 1 [18]. Para una prueba de este resultado ver el Apéndice A.

Es decir, eventualmente el número de alelos de tipo A_1 en la población será 0 (lo que indicaría la pérdida del alelo A_1), o bien $2N$ (pérdida del alelo A_2). Cuando esto último sucede, se dice que el alelo A_1 *se fijó* en la población.

Por otro lado, como $X_{n+1}|X_n \sim \text{Bin}(2N, X_n/2N)$ tenemos que

$$\mathbb{E}(X_{n+1} | X_n = i) = 2N \left(\frac{X_n}{2N} \right) = i = X_n \tag{2.2}$$

Luego, como $\mathbb{E}(|X_n|) = \mathbb{E}(X_n) = \sum_{i=1}^{2N} i\mathbb{P}(X_n = i) < \infty$ y todo proceso estocástico está adaptado a su filtración natural, resulta que $\{X_n : n \in \mathbb{N}\}$ es una *martingala* en tiempo discreto. En particular, usando 2.2 tenemos

$$\mathbb{E}(X_{n+1}) = \mathbb{E}\left(\mathbb{E}(X_{n+1}|X_n = i)\right) = \mathbb{E}(X_n) = \cdots = \mathbb{E}(X_0) \tag{2.3}$$

y

$$\mathbb{E}(X_n|X_0) = X_0 \tag{2.4}$$

Más detalles sobre martingalas y propiedades que utilizaremos de la esperanza condicional se pueden encontrar en el Apéndice B.

En la figura 2.3 simulamos 10 trayectorias de la frecuencia del alelo A_1 bajo el modelo de Wright-Fisher, para una población de tamaño $N = 50$, a lo largo de 100 generaciones, y partiendo de una frecuencia inicial $p_0 = 0.5$. Si bien el proceso es discreto, interpolamos para poder visualizar las distintas trayectorias. Vemos que para algunas el alelo A_1 se fijó (llegan a 1, la barrera superior), para otras se fijó A_2 (llegan a 0, la barrera inferior) y para otras en la generación 100 aún no se ha fijado ninguno de los alelos.

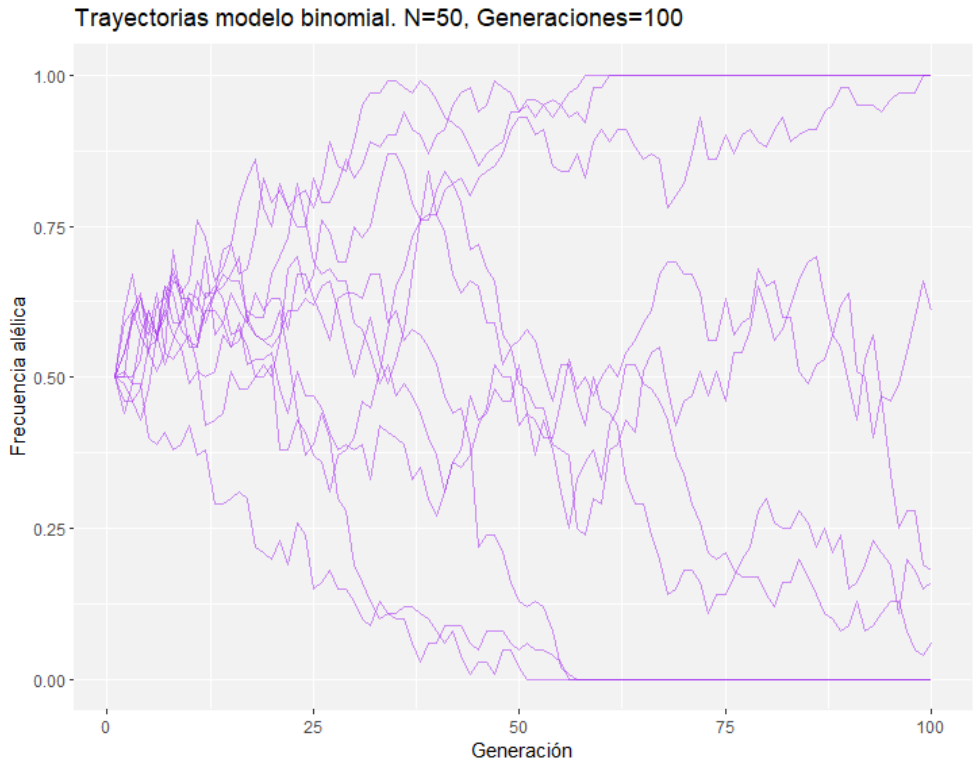

Figura 2.3: Simulación de 10 trayectorias del proceso de Wright-Fisher discreto, con frecuencia alélica inicial 0.5, tamaño poblacional $N = 50$, y 100 generaciones.

En resumen, la propiedad de martingala implica que en promedio la cantidad de alelos de tipo A_1 se mantendrá constante. Sin embargo, al ser una cadena de Markov con espacio de estados finito, eventualmente uno de los alelos se fijará. Nos interesa calcular la probabilidad de que un alelo

en particular se fije, y qué tan rápido lo hace. Para esto estudiamos la *probabilidad de fijación* y la *pérdida de heterocigosidad*.

Probabilidad de fijación.

Definimos el tiempo de fijación como el tiempo de parada

$$\tau = \text{mín}\{n : X_n = 0 \vee X_n = 2N\} \tag{2.5}$$

El siguiente teorema nos dice que la probabilidad de fijación del alelo A_1 es igual a su frecuencia inicial en la población [7].

Teorema 1. *En el modelo de Wright-Fisher la probabilidad de fijación de un alelo es su frecuencia inicial, es decir*

$$\mathbb{P}(X_\tau = 2N | X_0 = i) = \frac{i}{2N} \tag{2.6}$$

Demostración. Como el proceso es una martingala tenemos que

$$\mathbb{E}(X_n | X_0 = i) = i \tag{2.7}$$

Por otro lado, podemos descomponer $\mathbb{E}(X_n | X_0 = i)$ como

$$\begin{aligned} i = \mathbb{E}(X_n | X_0 = i) &= \mathbb{E}(X_n \mathbb{1}_{\{\tau \leq n\}} | X_0 = i) + \mathbb{E}(X_n \mathbb{1}_{\{\tau > n\}} | X_0 = i) \\ &= \mathbb{E}(X_\tau \mathbb{1}_{\{\tau \leq n\}} | X_0 = i) + \mathbb{E}(X_n \mathbb{1}_{\{\tau > n\}} | X_0 = i) \end{aligned}$$

pues para $\tau \leq n$ se tiene $X_n = X_\tau$.

Ahora bien, como la probabilidad de absorción del proceso es 1, tenemos que $\mathbb{P}(\tau < \infty) = 1$, y por tanto

$$\lim_{n \to \infty} X_n = X_\tau \tag{2.8}$$

Luego

$$\lim_{n \to \infty} \mathbb{E}(X_\tau \mathbb{1}_{\{\tau \leq n\}} | X_0 = i) = \mathbb{E}(X_\tau | X_0 = i)$$

Por otro lado, como $|X_n| \leq 2N$ y vale 2.8

$$\lim_{n \to \infty} \mathbb{E}(X_n \mathbb{1}_{\{\tau > n\}} | X_0 = i) = 0$$

y entonces

$$\mathbb{E}(X_\tau | X_0 = i) = i$$

Concluimos lo deseado pues

$$\begin{aligned} i = \mathbb{E}(X_\tau | X_0 = i) &= 0 \cdot \mathbb{P}(X_\tau = 0 | X_0 = i) + 2N \cdot \mathbb{P}(X_\tau = 2N | X_0 = i) \\ &= 2N \cdot \mathbb{P}(X_\tau = 2N | X_0 = i) \end{aligned}$$

□

Heterocigosidad.

La fijación de alguno de los alelos es un evento de probabilidad uno, y la probabilidad de que un alelo en particular se fije es su frecuencia inicial. Nos interesa además el tiempo esperado para que la fijación ocurra. Para esto, introducimos el concepto *heterocigosidad*: la probabilidad de que en tiempo n dos copias del locus dado (elegidas sin repetición) sean distintas

$$H_n^o = \frac{2X_n(2N - X_n)}{2N(2N-1)} \tag{2.9}$$

El siguiente teorema nos dice que la heterocigosidad disminuye en el tiempo, y el factor por el que lo hace depende del tamaño poblacional.

Teorema 2. *Sea H_n^o la heterocigosidad en tiempo n. En el modelo de Wright-Fisher vale:*

$$\mathbb{E}(H_n^o) = \left(1 - \frac{1}{2N}\right)^n \mathbb{E}(H_0^o) \tag{2.10}$$

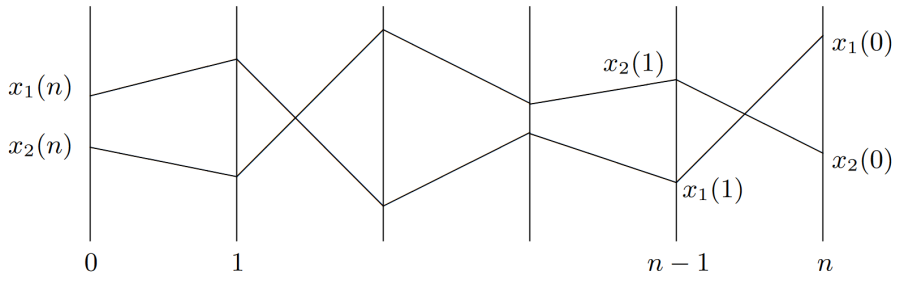

Figura 2.4: Dos genealogías. Imagen tomada de [7]

Demostración. Dada una población diploide de tamaño N, consideraremos individuos a los $2N$ alelos.

Sean $x_1(0)$ y $x_2(0)$ dos individuos en tiempo n. Cada individuo $x_i(0)$ es descendiente de $x_i(1)$ en tiempo $n-1$, que es descendiente de $x_i(2)$ en tiempo $n-2$, y así sucesivamente. La sucesión $x_i(m)$ describe el linaje (ancestros) de $x_i(0)$.

Como no hay mutación, si $x_1(m) = x_2(m)$, entonces $x_1(l) = x_2(l)$ para $m < l \leq n$.

Si $x_1(m) \neq x_2(m)$, la elección de alelos fue realizada de forma independiente, por lo que $x_1(m+1) \neq x_2(m+1)$ con probabilidad $1 - 1/2N$. Para que $x_1(n) \neq x_2(n)$, se deben elegir diferentes ancestros en todos los tiempos m tales que $1 \leq m \leq n$. Este evento tiene probabilidad $(1-1/2N)^n$.

Luego, $x_1(n)$ y $x_2(n)$ son dos individuos tomados al azar de la población en tiempo 0, por lo que la probabilidad de que sean distintos es H_0^o.

Por tanto

$$H_n^o = \left(1 - \frac{1}{2N}\right)^n H_0^o \tag{2.11}$$

Tomando esperanza llegamos a lo deseado. □

Por lo tanto, en el modelo de Wright-Fisher la heterocigosidad decae a razón de $1/2N$ por generación. La pérdida de heterocigosidad es una medida de deriva genética, y decimos que en el modelo de Wright-Fisher la deriva ocurre con tasa $1/2N$ por generación.

Observación: Notar que si elegimos los alelos con reposición tenemos

$$H_n = \frac{2X_n(2N - X_n)}{(2N)^2} = \frac{2N-1}{2N} H_n^o$$

y nuevamente vale

$$\mathbb{E}(H_n) = \left(1 - \frac{1}{2N}\right)^n \mathbb{E}(H_0) \tag{2.12}$$

Esta definición de heterocigosidad es la más utilizada, pero no nos permite realizar la demostración del teorema anterior, en la que miramos los linajes ancestrales. Este razonamiento es el que motiva la *teoría del coalescente*. En lugar de ocuparnos del destino de todos los alelos de la población, típicamente agrupados en clases según sus estados, consideraremos las propiedades de una muestra de n alelos tomados en el presente. Es decir, en vez de seguir el curso de la población hacia adelante, lo haremos hacia el pasado, partiendo del presente [37].

2.4. El coalescente

Cuando dos copias de un gen descienden de un ancestro común que les dio origen en alguna generación pasada, decimos que *coalescen* en esa generación [37]. Los eventos de coalescencia son simplemente eventos de replicación del ADN, y el interés en ellos se debe a su lugar en la historia de una muestra en particular.

La coalescencia de linajes se puede describir mediante un proceso matemático particular denominado el *coalescente*, presentado por el matemático británico John Kingman en 1982. Kingman probó que este modelo es el proceso ancestral límite para una amplia clase de modelos, incluyendo el de Wright-Fisher visto la sección anterior.

2.4.1. Tiempo discreto

La historia de una muestra de tamaño m comprende $m-1$ eventos de coalescencia. Cada evento de coalescencia disminuye el número de linajes ancestrales en uno. Es decir, el proceso toma una muestra de m linajes en el presente, y a través de una serie de pasos el número de linajes desciende de m a $m-1$, luego de $m-1$ a $m-2$, y así sucesivamente, hasta pasar de 2 a 1. Este último linaje, resultado del evento de coalescencia final, es llamado *ancestro en común más reciente* de la muestra (MRCA por sus siglas en inglés).

En cada evento de coalescencia dos linajes se fusionan en un linaje ancestral común. El resultado es un árbol bifurcado como el de la figura 2.5, que representa la historia de una muestra de tamaño cinco. La base del árbol representa las muestras en el presente, usualmente llamadas hojas del árbol. T_i se define como el tiempo en la historia de la muestra durante el cual existieron exactamente i linajes ancestrales.

Debido a la deriva genética, el coalescente es un proceso estocástico, que incluye tanto la estructura discreta de árbol, como los $m-1$ intervalos de tiempo de coalescencia.

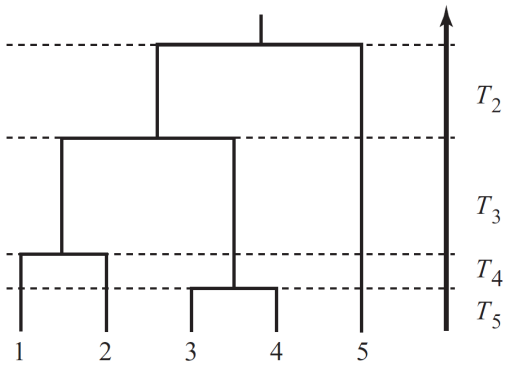

Figura 2.5: Una realización del coalescente para una muestra de tamaño 5.
Imagen tomada de [11]

Comenzaremos estudiando la distribución del tiempo de espera hasta el MRCA de dos alelos.

Coalescencia de una muestra de tamaño 2.

La probabilidad de que dos alelos en el presente tengan un ancestro en común en la generación anterior es $1/2N$, el primero elige el ancestro libremente, mientras que el segundo debe elegir el mismo ancestro que el primero. La probabilidad de que dos alelos tengan distintos ancestros en la generación anterior es entonces $1 - 1/2N$.

Dado que los muestreos en distintas generaciones son eventos independientes, la probabilidad de que dos alelos tengan un ancestro en común j generaciones hacia atrás es

$$\left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N} \tag{2.13}$$

En las primeras $j-1$ generaciones eligen distintos ancestros, y en la generación j eligen el mismo. Por tanto, el tiempo de coalescencia T_2 para que dos alelos encuentren el MRCA tiene distribución

$$\mathbb{P}(T_2 = j) = \left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N} \tag{2.14}$$

para $j = 1, 2, \cdots$, lo que implica que T_2 tiene distribución geométrica de parámetro $1/2N$.

La esperanza de T_2 es entonces $\mathbb{E}(T_2) = (1/2N)^{-1} = 2N$ generaciones. Es decir, el tiempo esperado hasta el MRCA de dos individuos es igual al número de alelos en la población.

Coalescencia de una muestra de tamaño m.

La probabilidad de que k ($\leq m$) alelos tengan k ancestros distintos en la generación anterior es

$$\begin{aligned}\frac{(2N-1)}{2N}\frac{(2N-2)}{2N}\cdots\frac{(2N-k+1)}{2N} = \prod_{i=1}^{k-1}\left(1-\frac{i}{2N}\right) &= 1 - \sum_{i=1}^{k-1}\frac{i}{2N} + O\left(\frac{1}{N^2}\right) \\ &= 1 - \binom{k}{2}\frac{1}{2N} + O\left(\frac{1}{N^2}\right)\end{aligned} \tag{2.15}$$

donde $O\left(\frac{1}{N^2}\right)$ son todos los términos que están divididos por N^2 o cualquier potencia de orden superior a N.

El razonamiento detrás de 2.15 es análogo al caso de dos alelos. El primer alelo elige el ancestro libremente entre las $2N$ opciones, el segundo tiene $2N-1$ opciones, el tercero $2N-2$, y así sucesivamente.

Si asumimos $m \ll N$, entonces $O\left(\frac{1}{N^2}\right)$ es despreciable y puede ser ignorado. Esta aproximación es equivalente a ignorar la probabilidad de que más de dos alelos tengan un ancestro en común en la misma generación.

Por tanto, la probabilidad de que no ocurra un evento de coalescencia es

$$1 - \binom{k}{2}\frac{1}{2N} \tag{2.16}$$

y la probabilidad de un evento de coalescencia en una generación dada es

$$\binom{k}{2}\frac{1}{2N} \tag{2.17}$$

En consecuencia, la probabilidad de que 2 de los k alelos tengan ancestro en común $T_k = j$ $(j = 1, 2, \cdots)$ generaciones atrás es

$$\mathbb{P}(T_k = j) \approx \left(1 - \binom{k}{2}\frac{1}{2N}\right)^{j-1}\binom{k}{2}\frac{1}{2N} \tag{2.18}$$

T_k tiene aproximadamente distribución geométrica de parámetro $\binom{k}{2}\frac{1}{2N}$.

Como todos los pares de alelos tienen la misma probabilidad de encontrar un ancestro en común, el par que encuentra el ancestro en común es elegido con igual probabilidad en los $\binom{k}{2}$ posibles pares. Y los tiempos $T_2, \cdots, T_m$ son independientes.

La precisión de la aproximación para N grande lleva a una formulación del coalescente con dos propiedades convenientes: un modelo que usa tiempo continuo, y que además es independiente del tamaño poblacional.

2.4.2. Tiempo continuo

En el modelo de Wright-Fisher el tiempo es medido en unidades discretas llamadas generaciones. Sin embargo, resulta conceptual y computacionalmente ventajoso considerar aproximaciones en tiempo continuo.

Una elección natural para el coalescente es escalar en tiempo continuo tal que una unidad de tiempo corresponda al tiempo de espera promedio para que dos alelos encuentren un ancestro en común [11]. En la subsección anterior mostramos que este es $2N$.

Usando esta transformación, el coalescente se vuelve independiente del tamaño poblacional. Esto enfatiza el hecho de que la estructura del proceso del coalescente es la misma para todas las poblaciones, siempre que el tamaño de la muestra m sea mucho menor al de la población N.

Sea $t = j/2N$, donde j es el tiempo medido en generaciones. Aquí $j = 2Nt$ transforma el tiempo continuo en discreto, llevándolo nuevamente a las generaciones. Si $j = 2Nt$ no es un entero, tomamos $j = \lfloor 2Nt \rfloor$ (mayor entero natural menor a j).

En la representación continua el tiempo de espera T_k^c para que k alelos tengan $k-1$ ancestros tiene distribución exponencial $T_k^c \sim \text{Exp}\left(\binom{k}{2}\right)$, es decir

$$\mathbb{P}(T_k^c \leq t) = 1 - e^{-\binom{k}{2}t} \tag{2.19}$$

Esto se debe a que cuando x es chico $(1-x) \approx e^{-x}$. Por tanto, si N es grande y sustituimos $j = 2Nt$ en 2.18 tenemos

$$\left(1 - \binom{k}{2}\frac{1}{2N}\right)^j = \left(1 - \binom{k}{2}\frac{1}{2N}\right)^{2Nt} \approx \left(e^{-\binom{k}{2}\frac{1}{2N}}\right)^{2Nt} = e^{-\binom{k}{2}t}$$

Debido a su distribución exponencial, la esperanza y la varianza de los tiempos de coalescencia son

$$\mathbb{E}(T_i) = \frac{1}{\binom{i}{2}} = \frac{2}{i(i-1)} \tag{2.20}$$

$$\text{Var}(T_i) = \left(\frac{1}{\binom{i}{2}}\right)^2 = \frac{4}{i^2(i-1)^2} \tag{2.21}$$

De la ecuación 2.20 concluimos que se espera que el tiempo de coalescencia más antiguo, en el que los últimos dos linajes coalescen al MRCA de la muestra, sea el más grande. En una muestra grande ocurrirán muchos eventos de coalescencia en un período corto de tiempo.

La simplicidad matemática del coalescente deriva del hecho de que los tiempos de coalescencia T_i son (1) independientes unos de otros y (2) independientes de la estructura de la genealogía. Como resultado podemos predecir varias cantidades, incluyendo el tiempo al ancestro en común más reciente de toda la muestra (T_{MRCA}), y la longitud total de todas las ramas en la genealogía (T_{tot}). Como T_i es el tiempo en la historia de la muestra durante el cual hubieron exactamente i linajes ancestrales tenemos

$$T_{\text{MRCA}} = \sum_{i=2}^{m} T_i \tag{2.22}$$

$$T_{\text{tot}} = \sum_{i=2}^{m} iT_i \tag{2.23}$$

La ecuación 2.22 es simplemente la suma de los $m-1$ tiempos de coalescencia, y la ecuación 2.23 es la suma de las longitudes de todas las ramas en la genealogía, divididdas en los intervalos de tiempos de coalescencia T_i.

Como T_{MRCA} y T_{tot} son funciones de variables aleatorias independientes con distribución exponencial, tenemos que

$$\begin{aligned}
\mathbb{E}(T_{\text{MRCA}}) = \sum_{i=2}^{m} \mathbb{E}(T_i) = \sum_{i=2}^{m} \frac{2}{i(i-1)} &= 2\sum_{i=2}^{m}\left(\frac{1}{i-1} - \frac{1}{i}\right) \\
&= 2\left(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \cdots + \frac{1}{m-1} - \frac{1}{m}\right) \\
&= 2\left(1 - \frac{1}{m}\right)
\end{aligned} \tag{2.24}$$

y

$$\mathbb{E}(T_{\text{tot}}) = \sum_{i=2}^{m} i\mathbb{E}(T_i) = \sum_{i=2}^{m} \frac{2i}{i(i-1)} = 2\sum_{i=1}^{m-1} \frac{1}{i} \tag{2.25}$$

Por otro lado

$$\text{Var}(T_{\text{MRCA}}) = \sum_{i=2}^{m} \text{Var}(T_i) = \sum_{i=2}^{m} \frac{4}{i^2(i-1)^2} \tag{2.26}$$

$$\text{Var}(T_{\text{tot}}) = \sum_{i=2}^{m} i^2\text{Var}(T_i) = \sum_{i=2}^{m} \frac{4i^2}{i^2(i-1)^2} = 4\sum_{i=1}^{m-1} \frac{1}{i^2} \tag{2.27}$$

Ahora veremos como utilizar la teoría desarrollada hasta el momento para derivar algunos resultados que permiten estudiar poblaciones con estructura, es decir, donde el apareamiento no es aleatorio.

2.5. Índice de fijación F_{ST}

La mayoría de las poblaciones naturales se desvían del supuesto de apareamiento aleatorio considerado hasta ahora. Por ejemplo, el apareamiento no podrá ser aleatorio en una especie que habite en un espacio con barreras geográficas establecidas, o que exceda la distancia que un individuo se puede trasladar a lo largo de su vida.

El apareamiento no aleatorio afecta la manera en la que los alelos se combinan para formar los genotipos, y altera las frecuencias genotípicas de una población. Esto puede tener consecuencias profundas en la dinámica evolutiva de una especie.

Existen diversas formas en las que una especie puede desviarse del apareamiento aleatorio. Dos de las más importantes son la *endogamia* (apareamiento preferencial entre individuos genéticamente relacionados) y la *subdivisión*. Ambas pueden ser analizadas al considerar una generalización del modelo de Hardy-Weinberg que incluye una nueva variable F [9].

Endogamia.

Consideremos una nueva variable F, que utilizaremos para describir la desviación de las frecuencias genotípicas esperadas en Hardy-Weinberg.

Genotipo:	A_1A_1	A_1A_2	A_2A_2
Frecuencia:	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$

Si $F = 0$, recuperamos las frecuencias de Hardy-Weinberg. Si $0 < F \leq 1$, hay un exceso de homocigotas comparado con lo esperado en Hardy-Weinberg. Si $F < 0$, hay un exceso de heterocigotas.

Estamos interesados en el caso $0 < F \leq 1$. Aquí F puede ser interpretado como la probabilidad de homocigosidad debido a "circunstancias especiales", que podrían ser, por ejemplo, preferencias de apareamiento entre individuos con fenotipos similares (que sean producto de genotipos

similares). En este caso se dice que existe endogamia en la población: el apareamiento entre individuos de ascendencia común es más probable que la reproducción entre individuos al azar. Esto conduce a un aumento en la proporción de homocigotas, y a una disminución en la proporción de heterocigotas en la población.

Para ver esto, consideremos un individuo tomado al azar en una población diploide. La probabilidad de que uno de sus alelos sea A_1 es p. La probabilidad de que su segundo alelo sea A_1, dado que el primero es A_1, es $F + (1 - F)p$. La probabilidad anterior es la suma de las probabilidades de dos eventos disjuntos: o bien el alelo es A_1 por homocigosidad debido a "circunstancias especiales"(que ocurre con probabilidad F), o bien no lo es (con probabilidad $1 - F$). En este último caso el segundo alelo es tomado al azar en la población, y la probabilidad de que sea A_1 es p. Como la elección de los alelos se realiza de forma independiente, la probabilidad de que un individuo sea A_1A_1 es $p[F + (1 - F)p]$.

Subdivisión.

Muchas especies ocupan extensas áreas geográficas, o tienen barreras efectivas para la migración, por lo que no se comportan como una única población con apareamiento aleatorio. En estos casos habrá diferenciación genética entre las subpoblaciones, lo que llevará a una desviación de las frecuencias esperadas bajo la ley de Hardy-Weinberg.

En el caso en el que hay apareamiento al azar dentro de cada subdivisión, las frecuencias genotípicas podrán describirse a través de una nueva interpretación de la variable F introducida anteriormente, llamada índice de fijación F_{ST} [9]. Valores altos de F_{ST} indican un grado de diferenciación entre las poblaciones.

Consideremos nuevamente un locus con dos alelos A_1, A_2 en una población (diploide) con d subpoblaciones. Sea p_i la frecuencia del alelo A_1 en la subpoblación P_i, y c_i el tamaño relativo de esta subpoblación en la población total. Tenemos $\sum_{i=1}^{d} c_i = 1$.

Sea $p = \sum_{i=1}^{d} c_i p_i$ la frecuencia promedio del alelo A_1 en la población total ponderada por los tamaños poblacionales, y $q = 1 - p$. Podemos escribir las frecuencias genotípicas como

Genotipo:	A_1A_1	A_1A_2	A_2A_2
Frecuencias en P_i	p_i^2	$2p_iq_i$	q_i^2
Frecuencias en población total	$\sum_{i=1}^{d} c_i p_i^2$	$\sum_{i=1}^{d} c_i 2p_iq_i$	$\sum_{i=1}^{d} c_i q_i^2$
Frecuencias en población total	$p^2(1 - F_{ST}) + pF_{ST}$	$2pq(1 - F_{ST})$	$q^2(1 - F_{ST}) + qF_{ST}$

Igualando las dos formas de escribir la frecuencia de heterocigotas en la población total

$$2pq(1 - F_{ST}) = \sum_{i=1}^{d} c_i 2p_iq_i$$

obtenemos

$$F_{ST} = \frac{2pq - \sum_{i=1}^{d} c_i 2p_iq_i}{2pq} \tag{2.28}$$

Proposición 1. *Sea $\bar{f}$ la probabilidad de que dos alelos tomados al azar con reposición de la población total sean iguales (o bien A_1, o bien A_2), y f_0 la probabilidad de que dos alelos tomados*

al azar con reposición de la misma subpoblación sean iguales. Entonces

$$F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}} \tag{2.29}$$

Demostración. Tenemos que

$$\bar{f} = p^2 + q^2$$
$$f_0 = \sum_{i=1}^{d} c_i (p_i^2 + q_i^2)$$

Luego, como

$$2pq = 1 - p^2 - q^2$$
$$\sum_{i=1}^{d} c_i 2p_i q_i = 1 - \sum_{i=1}^{d} c_i p_i^2 - \sum_{i=1}^{d} c_i q_i^2 = 1 - \sum_{i=1}^{d} c_i (p_i^2 + q_i^2)$$

podemos escribir F_{ST} como

$$F_{ST} = \frac{1 - p^2 - q^2 - 1 + \sum_{i=1}^{d} c_i (p_i^2 + q_i^2)}{2pq} = \frac{\sum_{i=1}^{d} c_i (p_i^2 + q_i^2) - p^2 - q^2}{1 - p^2 - q^2} = \frac{f_0 - \bar{f}}{1 - \bar{f}} \tag{2.30}$$

□

F_{ST} es entonces una medida de diferenciación entre la probabilidad de que dos alelos tomados al azar dentro de una subpoblación o en la población total sean del mismo tipo.

A modo de ejemplo, supongamos que tenemos una población con dos supoblaciones P_1, P_2 del mismo tamaño ($c_1 = c_2 = 1/2$), y que la frecuencia del alelo A_1 en P_1 es $p_1 = 1$, y en P_2 es $p_2 = 0$. Entonces la frecuencia total de A_1 es $p = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = 1/2 = q$. Luego

$$F_{ST} = \frac{\frac{1}{2}(1+1) - \frac{1}{4} - \frac{1}{4}}{1 - \frac{1}{4} - \frac{1}{4}} = 1$$

Este es el caso extremo de divergencia entre poblaciones.

La definición de F_{ST} en 2.28 nos permitirá escribirlo en función de otras cantidades utilizadas para medir la deriva genética: varianza de las frecuencias alélicas, heterocigosidad y tiempos esperados de coalescencia.

1) Varianza de las frecuencias alélicas

Proposición 2. *Sea* $\mathbf{p}$ *la variable aleatoria discreta que toma el valor* p_i *con probabilidad* c_i*. Entonces*

$$F_{ST} = \frac{\mathrm{Var}(\mathbf{p})}{p(1-p)} \tag{2.31}$$

Demostración.

$$\mathbb{E}(\mathbf{p}) = \sum_{i=1}^{d} c_i p_i = p$$
$$\mathbb{E}(\mathbf{p})^2 = \sum_{i=1}^{d} c_i p_i^2$$

por lo que

$$\mathrm{Var}(\mathbf{p}) = \mathbb{E}(\mathbf{p}^2) - \mathbb{E}(\mathbf{p})^2 = \sum_{i=1}^{d} c_i p_i^2 - p^2$$

Análogamente, si consideramos $\mathbf{q}$ la variable aleatoria que toma el valor q_i con probabilidad c_i tenemos

$$\mathrm{Var}(\mathbf{q}) = \mathbb{E}(\mathbf{q}^2) - \mathbb{E}(\mathbf{q})^2 = \sum_{i=1}^{d} c_i q_i^2 - q^2$$

Luego, si sustituimos estas ecuaciones en 2.30 tenemos

$$\begin{aligned} F_{ST} &= \frac{\sum_{i=1}^{d} c_i p_i^2 - p^2 + \sum_{i=1}^{d} c_i q_i^2 - q^2}{1 - p^2 - q^2} \\ &= \frac{\mathrm{Var}(\mathbf{p}) + \mathrm{Var}(\mathbf{q})}{2pq} \\ &= \frac{\mathrm{Var}(\mathbf{p})}{pq} \end{aligned}$$

pues $\mathbf{q} = 1 - \mathbf{p} \Rightarrow \mathrm{Var}(\mathbf{q}) = \mathrm{Var}(1 - \mathbf{p}) = \mathrm{Var}(\mathbf{p})$ □

Este resultado muestra que F_{ST} es siempre no negativo. Siempre que haya variación entre las subpoblaciones ($\mathrm{Var}(\mathbf{p}) > 0$), las frecuencias genotípicas de la población presentarán una deficiencia de heterocigotas, condición conocida como *efecto Wahlund*.

2) Heterocigosidad

Definimos la *heterocigosidad esperada* en la población como $H_{\mathrm{exp}} = 2pq$, y la *heterocigosidad observada* como $H_{\mathrm{obs}} = \sum_{i=1}^{d} c_i 2 p_i q_i$ (promedio de las heterocigosidades en las subpoblaciones, ponderadas por su tamaño).

Proposición 3. *Sean H_{exp} y H_{obs} las heterocigosidades esperada y observada respectivamente. Entonces*

$$F_{ST} = \frac{H_{exp} - H_{obs}}{H_{exp}} \tag{2.32}$$

Demostración. El resultado es directo por la ecuación 2.28. □

3) Tiempos esperados de coalescencia

Para escribir F_{ST} en función de los tiempos esperados de coalescencia primero debemos introducir el concepto de *identidad por descendencia*.

Hay tres formas fundamentales en las que dos alelos en un mismo locus pueden ser idénticos:

1. *Identidad por estado.* La definición de alelos idénticos por estado depende del contexto. Cuando hablamos de ADN, dos alelos son iguales por estado si no difieren en la secuencia de ADN. En caso de ser distintos por estado, la diferencia puede ser de un nucléotido o miles de nucleótidos.

2. *Identidad por origen.* Dos alelos son idénticos por origen si provienen del mismo locus y en el mismo cromosoma.

   Por ejemplo, dos alelos provenientes de individuos diferentes son siempre distintos por origen. En un individuo diploide, dos alelos provenientes del mismo locus pero en diferentes cromosomas son distintos por origen.

   El concepto de identidad por origen nos permite hablar de una muestra de n alelos distintos, sin implicar que los alelos son distintos por estado.

3. *Identidad por descendencia.* Dos alelos son idénticos por descendencia si comparten un ancestro.

   Estrictamente hablando, dos alelos del mismo locus nunca pueden ser distintos por descendencia, pues el origen de la vida es único y por ende todos los alelos tienen un ancestro en común. En la práctica, usualmente hablamos de un período corto de tiempo. Es decir, decimos que dos alelos son distintos por descendencia si no tienen un ancestro en común en, por ejemplo, las últimas 10 generaciones.

En 2.29 definimos F_{ST} en función de probabilidad de identidad por estado. Sin embargo, en la literatura se suele definir en términos de identidad por descendencia. Es decir, $F_{ST} = \frac{f_0 - \bar{f}}{1-\bar{f}}$, donde f_0 es la probabilidad de que dos alelos tomados de la misma población sean idénticos por descendencia, y $\bar{f}$ es la probabilidad de que dos alelos tomados de poblaciones distintas sean idénticos por descendencia.

La identidad por descendencia es igual a la identidad por estado si hacemos tender la tasa de mutación a cero [30]. Esto tiene sentido si consideramos períodos cortos de tiempo, pues la tasa de mutación es muy baja para algunas especies. En particular, la tasa de mutación en humanos es aproximadamente 10^{-8} por individuo por generación y cada generación tiene una duración aproximada de 25 años.

Para presentar la interpretación de F_{ST} en función de los tiempos esperados de coalescencia debemos introducir la mutación y hacerla tender a 0. La mutación es otro proceso aletorio, considerado un evento raro para muchas especies, por lo que se suele utilizar la distribución de Poisson para modelarlo. A los efectos del presente trabajo no será importante, sólo trabajaremos con una tasa de mutación μ constante.

Proposición 4. *Sea $\mathbb{E}(T_0)$ el tiempo esperado de coalescencia para dos alelos de una misma subpoblación y $\mathbb{E}(\bar{T})$ el tiempo esperado de coalescencia para dos alelos de la población general. Entonces*

$$F_{ST} = \frac{\mathbb{E}(\bar{T}) - \mathbb{E}(T_0)}{\mathbb{E}(\bar{T})} \tag{2.33}$$

Demostración. Sea $g(t)$ la probabilidad de que dos alelos no tengan un ancestro en común en la generación t en el pasado, o equivalentemente, la probabilidad de que dos alelos no coalescan en la generación t.

Sea $\mathbb{P}(T = t)$ la probabilidad de que el evento de coalescencia ocurra en la generación t. Entonces

$$\mathbb{P}(T = t) = g(t-1) - g(t) \tag{2.34}$$

Notar que como la población considerada es finita $\lim_{t\to\infty} g(t) = 0$, y $\sum_{t=1}^{\infty} \mathbb{P}(T = t) = g(0) = 1$ por ser una serie telescópica.

Por otro lado, el tiempo esperado de coalescencia es

$$\begin{aligned}\bar{t} = \sum_{t=1}^{\infty} t\mathbb{P}(T = t) &= \sum_{t=1}^{\infty} tg(t-1) - \sum_{t=1}^{\infty} tg(t) \\ &= \sum_{t=0}^{\infty} (t+1)g(t) - \sum_{t=1}^{\infty} tg(t) \\ &= \sum_{t=1}^{\infty} g(t)\end{aligned} \tag{2.35}$$

Sea μ la tasa de mutación. La probabilidad de identidad por estado de dos alelos es la probabilidad de que no haya ocurrido mutación luego del evento de coalescencia. Es decir, la probabilidad de que dos alelos sean idénticos es

$$f = \sum_{t=1}^{\infty} (1-\mu)^{2t} \mathbb{P}(T = t)$$

donde $\mathbb{P}(T = t)$ es la probabilidad de que dos alelos coalescan en la generación t, y $(1-\mu)^{2t}$ la probabilidad de que ninguno de los dos alelos haya mutado en las t generaciones siguientes al evento de coalescencia.

Si consideramos f como función de μ, entonces

$$\lim_{\mu\to 0} f = \sum_{t=1}^{\infty} \mathbb{P}(T = t) = 1$$

Sin embargo, el límite de F_{ST} cuando $\mu \to 0$ es no trivial, y lo calculamos aplicando la regla de l´Hôpital:

$$\begin{aligned}\lim_{\mu\to 0} F_{ST} &= \lim_{\mu\to 0} \frac{f_0 - \bar{f}}{1 - \bar{f}} \\ &= \lim_{\mu\to 0} \frac{-\sum_{t=1}^{\infty} 2t(1-\mu)^{2t-1}\mathbb{P}_0(T = t) - \left[-\sum_{t=1}^{\infty} 2t(1-\mu)^{2t-1}\mathbb{P}(T = t)\right]}{\sum_{t=1}^{\infty} 2t(1-\mu)^{2t-1}\mathbb{P}(T = t)} \\ &= \frac{\sum_{t=1}^{\infty} t\mathbb{P}(T = t) - \sum_{t=1}^{\infty} t\mathbb{P}_0(T = t)}{\sum_{t=1}^{\infty} t\mathbb{P}(T = t)} \\ &= \frac{\mathbb{E}(\bar{T}) - \mathbb{E}(T_0)}{\mathbb{E}(\bar{T})}\end{aligned}$$

□

El tiempo de coalescencia promedio entre alelos en diferentes subpoblaciones será mayor al esperado entre alelos de la misma subpoblación. Es decir, la subdivisión enlentece el proceso de pérdida de la diversidad genética al mantener a los individuos en poblaciones parcialmente aisladas.

La ecuación 2.33 es una aproximación de F_{ST} que nos permite expresarlo de forma tal que no depende de la tasa de mutación, suponiendo que la mutación es débil.

Por la ecuación 2.31 el índice de fijación F_{ST} mide la cantidad de variación genética que puede explicarse debido a la estructura de la población. En vistas de la ecuación 2.32, F_{ST} es la fracción de la diversidad total que no es consecuencia de la diversidad promedio dentro de las subpoblaciones, donde la diversidad se mide como la probabilidad de que dos alelos seleccionados al azar sean distintos (H_{exp}). Por otro lado, en la ecuación 2.33 F_{ST} se puede interpretar como una medida de cuánto más cerca están dos individuos de la misma subpoblación, en comparación con la población total.

Veamos el caso particular de una población con dos subpoblaciones.

F_{ST} para dos poblaciones.

Dado un locus con dos alelos A_1, A_2, supongamos que tenemos dos poblaciones P_1 y P_2 con frecuencias alélicas p_1 y p_2 para el alelo A_1, $q_i = 1 - p_i$. Supongamos además que las poblaciones tienen el mismo tamaño, es decir $c_1 = c_2 = 1/2$. Luego

$$\begin{aligned} F_{ST}(P_1, P_2) &= \frac{2\frac{(p_1+p_2)}{2}\frac{(q_1+q_2)}{2} - \frac{1}{2}(2p_1q_1 + 2p_2q_2)}{2\frac{(p_1+p_2)}{2}\frac{(q_1+q_2)}{2}} \\ &= \frac{\frac{1}{2}(p_1+p_2)(q_1+q_2) - p_1q_1 - p_2q_2}{\frac{1}{2}(p_1+p_2)(q_1+q_2)} \\ &= \frac{(p_1-p_2)^2}{p_1q_1 + p_1q_2 + p_2q_1 + p_2q_2} \end{aligned} \tag{2.36}$$

El denominador es la probabilidad de tomar dos individuos de la población total y que tengan distintos alelos. Los dos individuos pueden pertenecer a la misma población, o a distintas poblaciones.

Sin embargo, muchas veces se utiliza una definición alternativa de F_{ST}, en la que el denominador es la probabilidad de tomar individuos de distintas poblaciones [22]:

$$F_{ST}(P_1, P_2) = \frac{(p_1 - p_2)^2}{p_1q_2 + p_2q_1} \tag{2.37}$$

Esta definición se generaliza al caso de n poblaciones, y es posible derivar expresiones en función de distintas cantidades. En particular, una expresión que utilizaremos es la siguiente

$$F_{ST} = \frac{\mathbb{E}(T_B) - \mathbb{E}(T_W)}{\mathbb{E}(T_B)} \tag{2.38}$$

donde $\mathbb{E}(T_W)$ es el tiempo esperado de coalescencia para dos alelos de distintas subpoblaciones (between), y $\mathbb{E}(T_W)$ es el tiempo esperado de coalescencia para dos alelos den la misma subpoblación (within). Esta ecuación se diferencia de 2.33 en que en los individuos son tomados de distintas subpoblaciones, no de la población total.

En la práctica, ninguna de las cantidades utilizadas para definir F_{ST} es fácilmente medible, por lo que se propusieron varios estimadores. Uno de los más utilizados para estimar 2.38 cuando tenemos datos de secuencias de ADN es

$$F_{ST} = \frac{\pi_B - \pi_W}{\pi_B} \tag{2.39}$$

donde π_B y π_W representan el número promedio de diferencias por pares entre dos individuos muestreados en diferentes poblaciones y en la misma población respectivamente [32]. Lo definiremos formalmente en el siguiente capítulo, cuando estimemos los estadísticos F.

Capítulo 3

Estadísticos F

Los modelos demográficos bajo los cuales se abordan las preguntas realizadas en la introducción son llamados *árboles de poblaciones* y *grafos de mezcla*.

La filogenia o árbol es un modelo en el cual la relación entre las poblaciones se presenta como un árbol (figura 3.1 izquierda). En este, los nodos representan poblaciones, y las aristas la deriva genética, que se define como la variación en las frecuencias alélicas. En este sentido, la longitud de las aristas o ramas en la filogenia corresponde a la cantidad de deriva genética que ha ocurrido entre las poblaciones.

Por ejemplo, en la figura 3.1 izquierda, el camino azul se interpreta como la deriva genética que ha ocurrido entre las poblaciones P_1 y P_2, y el camino rojo que une R con P_0 (población ancestral a P_1 y P_2), se interpreta como la deriva compartida entre P_1 y P_2, ya que a lo largo de ese camino P_1 y P_2 fueron la misma población.

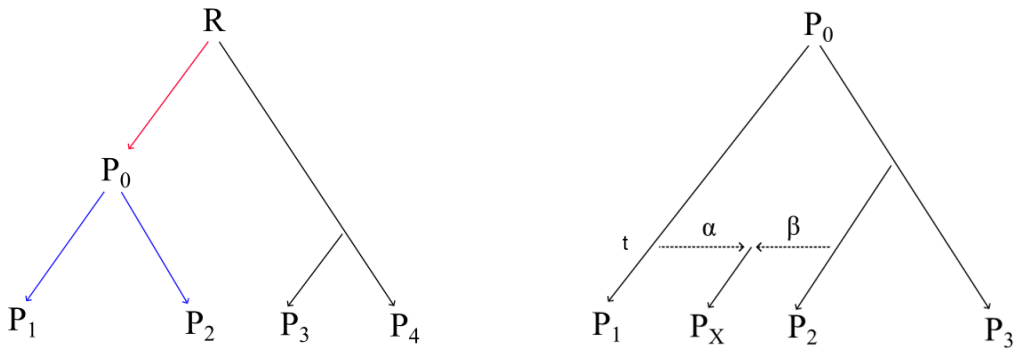

Figura 3.1: Árbol de poblaciones (izquierda) y grafo de mezcla (derecha).

Un modelo alternativo es el grafo de mezcla, que generaliza el árbol poblacional al permitir aristas que representen la fusión de poblaciones o el intercambio significativo de migrantes (figura 3.1 derecha). En los grafos de mezcla tenemos dos tipos de aristas: las que representan la deriva genética y las que representan el flujo génico.

En la figura 3.1 derecha, la población P_X es resultado de un evento de mezcla de las poblaciones P_1 y P_2 en tiempo t, con proporciones de mezcla α y $\beta = 1-\alpha$ respectivamente. Esto implica que

si tomamos un individuo al azar de la población P_X, con probabilidad α su ancestro en tiempo t pertenecerá a P_1, y con probabilidad β pertenecerá a P_2. Luego del evento de mezcla la deriva actúa sobre las tres poblaciones.

Como la deriva genética entre dos poblaciones se define como la variación en las frecuencias alélicas de las mismas, para estudiarla primero debemos modelar las frecuencias alélicas.

Frecuencias alélicas.

Consideremos un locus con dos alelos A_1, A_2, en poblaciones $P_1, P_2, \cdots$ de tamaño finito. Asumiremos que los individuos son haploides, y nos centraremos sólo en el efecto de la deriva genética. Ignoraremos los efectos de la mutación, selección y otras fuerzas evolutivas.

Llamemos p_i a la frecuencia del alelo A_1 en P_i. Tenemos dos formas de modelar la frecuencia alélica en una población variando en el tiempo.

En el modelo discreto tenemos una población de tamaño constante $2N$, en el que el tiempo es medido en generaciones. El número de copias de un alelo debe ser uno de los $2N + 1$ posibles valores: $0, 1, \cdots, 2N - 1, 2N$. Si hay k copias de un alelo, entonces su frecuencia en la generación j es $p_j = \frac{k}{2N}$. Es decir, la frecuencia alélica es

$$p_j = \frac{X_j}{2N}$$

donde X_j es la variable aleatoria que cuenta la cantidad de alelos de tipo A_1 en la generación j, que presentamos en la sección 2.3.1.

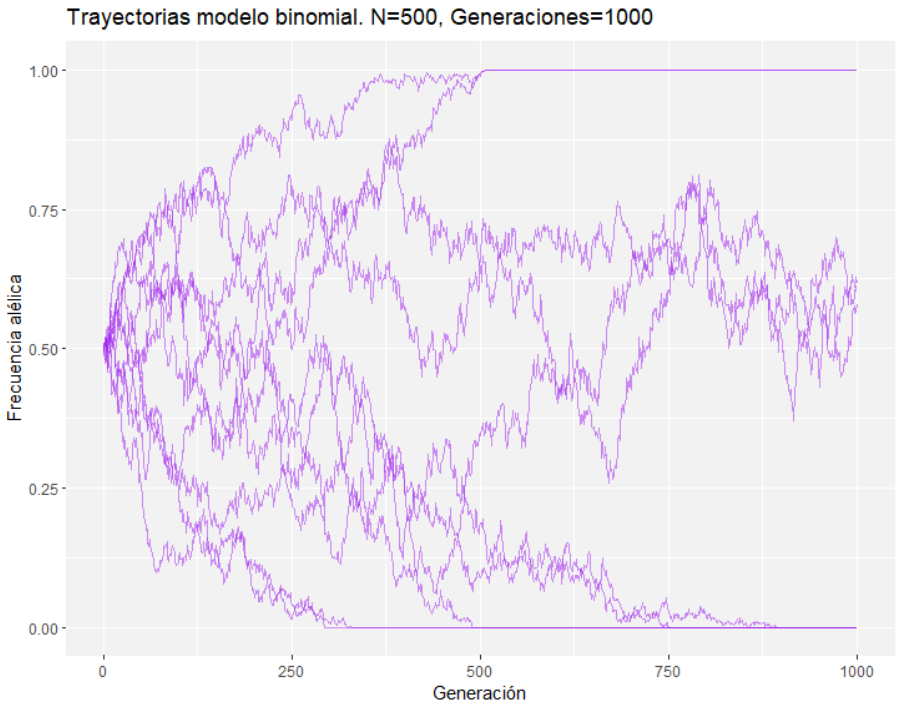

Figura 3.2: Simulación de 10 trayectorias de la frecuencia alélica bajo el proceso de Wright-Fisher discreto, con tamaño poblacional $N = 500$, 1000 generaciones y frecuencia inicial $p_0 = 0.5$.

En la figura 3.2 simulamos 10 trayectorias de este modelo, con frecuencia alélica inicial 0.5, en una población de tamaño $N = 500$ para 1.000 generaciones. Las trayectorias son discretas, pero al igual que en la figura 2.3 del capítulo anterior, interpolamos para poder visualizarlas.

Una alternativa es considerar un modelo en el que tanto el tiempo como la frecuencia alélica sean medidos de forma continua, es decir, $p_t \in [0, 1]$ y $t \in [0, \infty)$. Esto es logrado tomando límite $N \to \infty$, y escalando el tiempo en unidades de $2N$ generaciones como lo hicimos para el coalescente. Este proceso resulta ser la solución de la ecuación diferencial estocástica

$$dp_t = \sqrt{p_t(1 - p_t)}dW_t$$

donde W_t es un proceso de Wiener, y p_t es la frecuencia del alelo A_1 en la población. De esta forma obtenemos la *aproximación por difusión del modelo de Wright-Fisher* [6].

En la figura 3.3 simulamos 10 trayectorias de este modelo, con frecuencia alélica inicial 0.5, y tiempo final $T = 1$. Las trayectorias del proceso son continuas y no diferenciables en todo punto. Para visualizarlas simulamos el proceso en un conjunto discreto (partición del intervalo $[0, 1]$) e interpolamos entre los puntos.

Los modelos de difusión aproximan la dinámica de las frecuencias alélicas a través del tiempo en poblaciones de gran tamaño, y permiten computar cantidades de interés que son matemáticamente intratables en la mayoría de los modelos discretos.

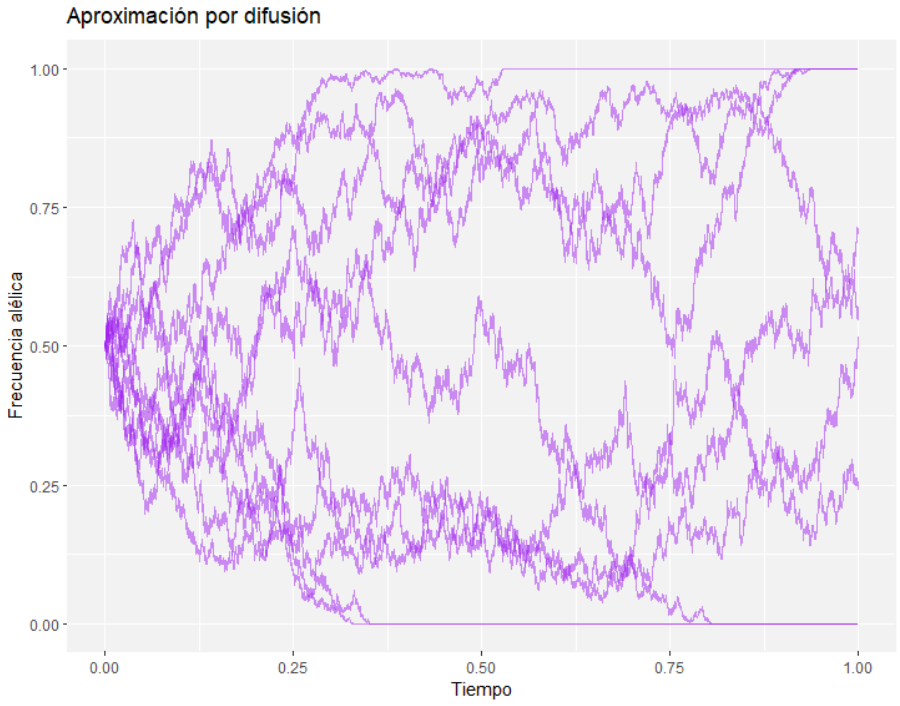

Figura 3.3: Simulación de 10 trayectorias del proceso de Wright-Fisher como difusión, con $t \in [0, 1]$ y frecuencia inicial $p_0 = 0.5$. Simulación realizada con el esquema de Euler.

En cualquiera de los dos casos el proceso $\{p_t\}$ es una martingala, con trayectorias discretas en el primero, y continuas en el segundo. A lo largo del capítulo utilizaremos dos propiedades

fundamentales.

1) Propiedad de martingala.

Para todos $s, t \geq 0$ tales que $t > s$

$$\mathbb{E}(p_t|p_s) = p_s \tag{3.1}$$

que implica

$$\mathbb{E}(p_t|p_0) = p_0 \tag{3.2}$$

$$\mathbb{E}(p_t) = \mathbb{E}(\mathbb{E}(p_t|p_0)) = \mathbb{E}(p_0) \tag{3.3}$$

Por ejemplo, si en la figura 3.1 izquierda p_0 es la frecuencia del alelo A_1 en la población P_0, y p_1, p_2 la frecuencia del alelo A_1 en P_1, P_2 respectivamente, entonces $\mathbb{E}(p_1|p_0) = \mathbb{E}(p_2|p_0) = p_0$. Análogamente, si r es la frecuencia del alelo A_1 en R, tenemos $\mathbb{E}(p_0|r) = \mathbb{E}(p_1|r) = \mathbb{E}(p_2|r) = r$, pues las tres poblaciones P_0, P_1, P_2 descienden de la población ancestral R.

2) Incrementos independientes.

Sean P_1 y P_2 poblaciones que se separaron de una población ancestral P_0 (figura 3.1 izquierda), y p_i la frecuencia del alelo A_1 en P_i. Fijado p_0, los incrementos $p_1 - p_0$ y $p_2 - p_0$ son variables aleatorias independientes, y entonces

$$\mathbb{E}[(p_1 - p_0)(p_2 - p_0)|p_0] = \mathbb{E}(p_1 - p_0|p_0)\mathbb{E}(p_2 - p_0|p_0) \tag{3.4}$$

Decimos que dos variables aleatorias X, Y son ortogonales si $\mathbb{E}(XY) = 0$. La propiedad 3.4, junto con 3.2 nos dice que $p_1 - p_0$ y $p_2 - p_0$ son variables aleatorias ortogonales:

$$\begin{aligned}\mathbb{E}[(p_1 - p_0)(p_2 - p_0)] &= \mathbb{E}(\mathbb{E}[(p_1 - p_0)(p_2 - p_0)|p_0]) \\ &= \mathbb{E}[\underbrace{\mathbb{E}(p_1 - p_0|p_0)}_{0}\underbrace{\mathbb{E}(p_2 - p_0|p_0)}_{0}] = 0\end{aligned} \tag{3.5}$$

3.1. Estadístico F_2

Definiremos F_2 como una forma de medir cuánta deriva genética ha ocurrido entre dos poblaciones.

Definición 1. *Dado un locus con dos alelos A_1 y A_2, sean p_1 y p_2 las frecuencias alélicas de A_1 en dos poblaciones P_1 y P_2 respectivamente. El estadístico F_2 se define como*

$$F_2(P_1, P_2) = F_2(p_1, p_2) = \mathbb{E}\left[(p_1 - p_2)^2\right] \tag{3.6}$$

Notar que F_2 es no negativo y simétrico por definición. Además, como estamos en el caso de un locus con dos alelos (A_1, A_2), la elección del alelo no altera el resultado del estadístico

$$F_2(P_1, P_2) = \mathbb{E}\left[(p_1 - p_2)^2\right] = \mathbb{E}\left[((1 - p_1) - (1 - p_2))^2\right]$$

Por otro lado, observar que la expresión cuadrática a la que le calculamos la esperanza coincide con el numerador de F_{ST} en 2.36 y 2.37. Esta es la definición original de F_2 dada en [28].

Antes de continuar, es necesario distinguir entre los estadísticos (cantidades calculadas a partir de los datos) y los parámetros subyacentes que son parte del modelo. Nuestra definición de F_2 (como esperanza respecto a un proceso histórico evolutivo) coincide con la última interpretación. En la práctica, lo que estimamos es $f_2(P_1, P_2) = (p_1 - p_2)^2$.

En las siguientes páginas veremos como se relaciona F_2 con distintas cantidades utilizadas para medir la deriva genética. La deriva puede ser cuantificada en términos de:

- Varianza en las frecuencias alélicas
- Heterocigosidad
- Probabilidad de identidad por descendencia
- Tiempos esperados de coalescencia

Mostraremos cómo se relaciona F_2 con estas cantidades en el caso de una población cambiando en el tiempo, y de dos poblaciones parcialmente aisladas.

1) Una población

Sea P_0 una población en tiempo t_0 y P_t la misma población medida en tiempo t. La siguiente proposición nos dice que $F_2(P_0, P_t)$ puede escribirse en función de las varianzas de las frecuencias alélicas.

Proposición 5.

$$F_2(P_0, P_t) = \mathrm{Var}(p_t - p_0) = \mathrm{Var}(p_t) - \mathrm{Var}(p_0) \tag{3.7}$$

Demostración. Por 3.2 y 3.3 se tiene

$$\begin{aligned}
\mathrm{Cov}(p_0, p_t) = \mathbb{E}(p_0 p_t) - \mathbb{E}(p_0)\mathbb{E}(p_t) &= \mathbb{E}(\mathbb{E}(p_0 p_t | p_0)) - \mathbb{E}(p_0)^2 \\
&= \mathbb{E}(p_0 \mathbb{E}(p_t | p_0)) - \mathbb{E}(p_0)^2 \\
&= \mathbb{E}(p_0^2) - \mathbb{E}(p_0)^2 \\
&= \mathrm{Var}(p_0)
\end{aligned}$$

Luego, como $\mathbb{E}(p_t - p_0) = 0$

$$\begin{aligned}
F_2(P_0, P_t) = \mathbb{E}\left[(p_t - p_0)^2\right] &= \mathrm{Var}(p_t - p_0) + [\mathbb{E}(p_t - p_0)]^2 \\
&= \mathrm{Var}(p_t - p_0) \\
&= \mathrm{Var}(p_t) + \mathrm{Var}(p_0) - 2\mathrm{COV}(p_0, p_t) \\
&= \mathrm{Var}(p_t) + \mathrm{Var}(p_0) - 2\mathrm{Var}(p_0) \\
&= \mathrm{Var}(p_t) - \mathrm{Var}(p_0)
\end{aligned}$$

□

Ahora introduciremos F_2 en función de la heterocigosidad.

Proposición 6. *Sean $H_0 = 2p_0(1 - p_0)$ la heterocigosidad en P_0 y $H_t = 2p_t(1 - p_t)$ la heterocigosidad en P_t. Entonces*

$$F_2(P_0, P_t) = \frac{\mathbb{E}(H_0) - \mathbb{E}(H_t)}{2} \tag{3.8}$$

Demostración.

$$\begin{aligned}\mathbb{E}(H_t) &= \mathbb{E}(2p_t(1-p_t)) \\ &= \mathbb{E}\left[2(p_0+p_t-p_0)(1-p_0-(p_t-p_0))\right] \\ &= \underbrace{\mathbb{E}(2p_0(1-p_0))}_{\mathbb{E}(H_0)} - 2(\mathbb{E}[p_0(p_t-p_0)] - \mathbb{E}[(p_t-p_0)(1-p_0)] + \underbrace{\mathbb{E}[(p_t-p_0)^2]}_{F_2(P_0,P_t)})\end{aligned}$$

Ahora bien, usando la fórmula de la esperanza total junto con 3.2 y 3.3 tenemos

$$\mathbb{E}(p_0(p_t-p_0)) = \mathbb{E}(\mathbb{E}(p_0(p_t-p_0)|p_0)) = \mathbb{E}(p_0\mathbb{E}((p_t-p_0)|p_0)) = 0$$
$$\mathbb{E}((p_t-p_0)(1-p_0)) = \mathbb{E}(p_t-p_0) - \mathbb{E}((p_t-p_0)p_0) = 0$$

Luego

$$\mathbb{E}(H_t) = \mathbb{E}(H_0) - 2F_2(P_0,P_t)$$

□

La proposición anterior nos permite escribir F_2 en términos de identidad por descendencia.

Proposición 7. *Sea $H_0 = 2p_0(1-p_0)$ la heterocigosidad en P_0 y f la probabilidad de que el ancestro en común de dos individuos de P_t esté a menos de t generaciones. Entonces:*

$$F_2(P_0,P_t) = \frac{1}{2}f\mathbb{E}(H_0) \tag{3.9}$$

Demostración. Si $2N$ es el tamaño de la población (haploide), tenemos que

$$1-f = \left(1-\frac{1}{2N}\right)^t$$

Por otro lado, usando la ecuación 2.10 demostrada en la sección 2.3 tenemos

$$\mathbb{E}(H_t) = \left(1-\frac{1}{2N}\right)^t \mathbb{E}(H_0) = (1-f)\mathbb{E}(H_0)$$

Luego, por la ecuación 3.8 tenemos que

$$F_2(P_0,P_t) = \frac{1}{2}\left(\mathbb{E}(H_0) - \mathbb{E}(H_t)\right) = \frac{1}{2}\left(\mathbb{E}(H_0) - (1-f)\mathbb{E}(H_0)\right) = \frac{1}{2}f\mathbb{E}(H_0)$$

□

Notar que f es chico cuando es baja la probabilidad de que el ancestro en común más reciente (MRCA) de dos individuos de P_t esté a menos de t generaciones, es decir, cuando no ha ocurrido tanta deriva entre P_0 y P_t. Por tanto, intuitivamente tiene sentido que F_2 aumente en función de f.

La ecuación 3.9 puede ser también interpretada en términos de probabilidad de identidad por descendencia: f es la probabilidad de que el MRCA de dos individuos esté a menos de t generaciones, que es igual a la probabilidad de que dos individuos sean idénticos por descendencia en P_t dado que en tiempo t_0 tienen un ancestro en común.

A través de estas ecuaciones conectamos F_2 con otras medidas de deriva genética

$$\mathbb{E}(H_t) = \mathbb{E}(H_0) - 2F_2(P_0, P_t) \tag{3.10}$$
$$= \mathbb{E}(H_0) - 2(\mathrm{Var}(p_t) - \mathrm{Var}(p_0)) \tag{3.11}$$
$$= \mathbb{E}(H_0)(1 - f) \tag{3.12}$$

2) Dos poblaciones.

Las ecuaciones 3.11 y 3.12 que describen la pérdida de heterocigosidad fueron establecidas por Wright en 1931. En poblaciones estructuradas existen relaciones muy similares cuando comparamos el número esperado de heterocigotas de la población total (H_{exp}) con el número de heterocigotas presente debido a diferencias en las frecuencias alélicas entre poblaciones (H_{obs}) [36].

Proposición 8. *Dado un locus con dos alelos A_1, A_2 en una población formada por dos subpoblaciones P_1 y P_2 del mismo tamaño, en las que las frecuencias alélicas de A_1 son p_1 y p_2 respectivamente, la proporción de heterocigotas se reduce de la siguiente forma:*

$$H_{obs} = H_{exp} - \frac{(p_1 - p_2)^2}{2} \tag{3.13}$$

Demostración. Sea P la población total y P_1, P_2 las subpoblaciones. La heterocigosidad observada (H_{obs}) es la suma de las heterocigosidades observadas en las subpoblaciones, ponderadas por los tamaños poblacionales. Como las poblaciones tienen el mismo tamaño, tenemos

$$\begin{aligned} H_{\mathrm{obs}} &= \frac{1}{2}(2p_1(1-p_1) + 2p_2(1-p_2)) \\ &= p_1 - p_1^2 + p_2 - p_2^2 \end{aligned}$$

Por otro lado, la heterocigosidad esperada en la población total es

$$H_{\mathrm{exp}} = 1 - (p^2 + (1-p)^2) = 2p(1-p)$$

donde p es la frecuencia del alelo A_1 en P y $p^2 + (1-p)^2$ la probabilidad de que dos individuos tomados al azar en P tengan el mismo alelo (homocigosidad).

Dado que P_1 y P_2 tienen el mismo tamaño, la frecuencia del alelo A_1 en P es $\frac{p_1+p_2}{2}$. Por tanto

$$\begin{aligned} H_{\mathrm{exp}} - \frac{(p_1-p_2)^2}{2} &= 1 - \left(\frac{p_1+p_2}{2}\right)^2 - \left(1 - \frac{p_1+p_2}{2}\right)^2 - \frac{(p_1-p_2)^2}{2} \\ &= 1 - \left(\frac{p_1+p_2}{2}\right)^2 - 1 + 2\left(\frac{p_1+p_2}{2}\right) - \left(\frac{p_1+p_2}{2}\right)^2 - \frac{(p_1-p_2)^2}{2} \\ &= p_1 + p_2 - \frac{(p_1+p_2)^2}{2} - \frac{(p_1-p_2)^2}{2} \\ &= p_1 + p_2 - p_1^2 - p_2^2 \\ &= H_{\mathrm{obs}} \end{aligned}$$

lo que concluye la demostración. □

Usaremos la proposición anterior para escribir $F_2(P_1, P_2)$ en función de la heterocigosidad, de la varianza de frecuencias alélicas y del índice de fijación F_{ST}.

Teorema 3. *Sean P_1 y P_2 dos poblaciones que descienden de una población ancestral P_0, y p_1, p_2, p_0 las frecuencias alélicas correspondientes. Entonces*

$$F_2(P_1, P_2) = 2[\mathbb{E}(H_{exp}) - \mathbb{E}(H_{obs})] \tag{3.14}$$
$$= \mathrm{Var}(p_1 - p_2) \tag{3.15}$$
$$= 2F_{ST}\mathbb{E}(H_{exp}) \tag{3.16}$$

Demostración. La primera igualdad se deduce de la ecuación 3.13 tomando esperanza y despejando

$$\mathbb{E}(H_{\text{obs}}) = \mathbb{E}(H_{\text{exp}}) - \mathbb{E}\left[\frac{(p_1 - p_2)^2}{2}\right] = \mathbb{E}(H_{\text{exp}}) - \frac{F_2(P_1, P_2)}{2}$$

Por otro lado, por 3.3 se tiene $\mathbb{E}(p_1) = \mathbb{E}(p_0) = \mathbb{E}(p_2) \Rightarrow \mathbb{E}(p_1 - p_2) = 0$, y entonces

$$\mathrm{Var}(p_1 - p_2) = \mathbb{E}\left[(p_1 - p_2)^2\right] - \underbrace{\mathbb{E}(p_1 - p_2)^2}_{0} = F_2(P_1, P_2)$$

Finalmente, recordar que F_{ST} puede escribirse como

$$F_{ST} = \frac{H_{\text{exp}} - H_{\text{obs}}}{H_{\text{exp}}}$$

Luego, si despejamos H_{exp} en la ecuación anterior, tomamos esperanza de ambos lados y utilizamos 3.14 tenemos

$$F_{ST}\mathbb{E}(H_{\text{exp}}) = \mathbb{E}(H_{\text{exp}}) - \mathbb{E}(H_{\text{obs}}) = \frac{F_2(P_1, P_2)}{2}$$

por lo que vale la tercera igualdad. □

Notar que a diferencia de F_{ST}, F_2 depende de la magnitud (en valor absoluto) de las frecuencias alélicas. En particular, como $H_{\text{exp}} = 2p(1-p)$ se maximiza en $p = 1/2$, En la práctica, agregar sitios sin variación contraerá el valor de los estadísticos a 0. Por esto es importante que todos los estadísticos sean calculados con el mismo conjunto de datos.

Hasta aquí vimos la relación entre F_2 y distintas medidas de deriva genética. En la siguiente sección veremos su representación en el árbol y grafo de mezcla, así como su interpretación en función de tiempos esperados de coalescencia.

3.1.1. Interpretación en el árbol coalescente

Los estadísticos F tienen una interpretación visual clara en el árbol de poblaciones. En este, las aristas que unen dos poblaciones representan la deriva genética que ha ocurrido entre ellas, es decir, la diferencia en las frecuencias alélicas. Como $F_2(P_1, P_2) = \mathbb{E}[(p_1 - p_2)^2] = \mathrm{Var}(p_1 - p_2)$, interpretamos $F_2(P_1, P_2)$ como la longitud del camino entre las poblaciones P_1 y P_2.

Si tenemos dos poblaciones P_1 y P_2 que se separaron de una población ancestral P_0, este camino está compuesto por dos ramas: la que une P_1 con P_0, y la que une P_0 con P_2 (figura 3.4). La longitud del camino es la suma de las longitudes de estos dos caminos. Esto se conoce como *propiedad aditiva* de F_2.

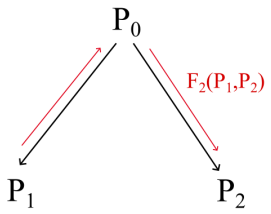

Figura 3.4: Árbol de dos poblaciones P_1, P_2 que descienden de P_0.

Teorema 4. *Sean P_1 y P_2 dos poblaciones que se separaron de una población ancestral P_0. Entonces*

$$F_2(P_1, P_2) = F_2(P_0, P_1) + F_2(P_0, P_2) \tag{3.17}$$

Demostración.

$$\begin{aligned} F_2(P_1, P_2) &= \mathbb{E}[(p_1 - p_2)^2] \\ &= \mathbb{E}[(p_1 - p_0 - (p_2 - p_0))^2] \\ &= \mathbb{E}[(p_1 - p_0)^2] - 2\mathbb{E}[(p_1 - p_0)(p_2 - p_0)] + \mathbb{E}[(p_2 - p_0)^2] \\ &= F_2(P_0, P_1) + F_2(P_0, P_2) \end{aligned}$$

donde en la tercera igualdad utilizamos la propiedad de incrementos independientes 3.5. □

En una filogenia de poblaciones existe un único camino de P_1 a P_2. Sin embargo, en un escenario de mezcla la deriva puede tomar caminos alternativos, y debemos considerar los árboles correspondientes a los caminos posibles, y la probabilidad de tomarlos [22].

Para esto, introduciremos una manera intuitiva de pensar F_2. En la definición, consideremos los términos que se multiplican como dos caminos entre P_1 y P_2:

$$F_2(P_1, P_2) = \mathbb{E}[\underbrace{(p_1 - p_2)}_{\text{camino 1}}\underbrace{(p_1 - p_2)}_{\text{camino 2}}] = \mathrm{Var}(p_1 - p_2) = \mathrm{Cov}(p_1 - p_2, p_1 - p_2) \tag{3.18}$$

F_2 se interpreta como la covarianza de dos posibles caminos de P_1 a P_2, que corresponde a la superposición de estos dos caminos, o en otras palabras, a la deriva genética compartida.

En un árbol de poblaciones existe un único camino de P_1 a P_2, por lo que los caminos que conectan a las poblaciones siempre coinciden, y su superposición es el propio camino (figura 3.4). Esto no será así para el grafo de mezcla. En árboles, la interpretación como superposición de caminos tendrá más sentido en el caso de los estadísticos F_3 y F_4, pero el razonamiento será el mismo.

Ahora bien, supongamos que estamos en el caso de la figura 3.5, en el que tenemos las poblaciones P_1, P_2 en el presente, donde P_2 es resultado de la deriva luego de un evento de mezcla de dos poblaciones ancestrales, con proporciones de mezcla α y $\beta = 1 - \alpha$. La población que contribuye α a la mezcla está formada por ancestros de la población P_1.

Tenemos dos caminos posibles de P_1 a P_2, con probabilidad α se tomará uno, y con probabilidad β el otro. Llamemos a estos caminos c_α y c_β. F_2 se interpreta como la superposición de dos caminos en el grafo, lo que nos lleva a considerar tres posibles combinaciones de caminos

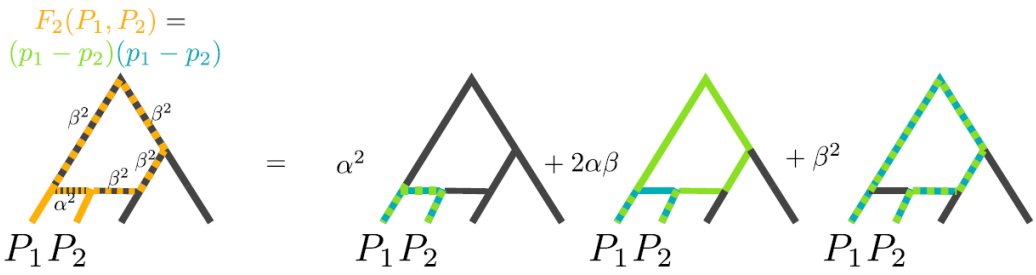

Figura 3.5: Interpretación de F_2 como caminos en el grafo. Imagen tomada de [23]

- $\{c_\alpha, c_\alpha\}$
- $\{c_\alpha, c_\beta\}$
- $\{c_\beta, c_\beta\}$

Además, la elección de los caminos se hace de forma independiente, por lo que las probabilidades se multiplican. Luego, F_2 será un promedio de la superposición de caminos en las topologías, ponderadas por sus probabilidades (figura 3.5):

$$F_2(P_1, P_2) = \alpha^2(c_\alpha \cap c_\alpha) + 2\alpha\beta(c_\alpha \cap c_\beta) + \beta^2(c_\beta \cap c_\beta)$$

Un inconveniente asociado a este marco de conteo de todos los posibles caminos a través del grafo es que escala cuadráticamente con el número de eventos de mezcla, lo que dificulta el cálculo cuando este es grande. Más importante aún, esta interpretación se restringe a subpoblaciones panmícticas, y no se puede utilizar cuando el modelo poblacional no se puede representar como promedio ponderado de árboles.

Es por esto que en *"Admixture, Population Structure, and F-Statistics"* el autor propone redefinir F_2 utilizando la teoría del coalescente, presentada en el capítulo 2 de este trabajo.

En lugar de frecuencias alélicas en un grafo de mezcla fijo, la teoría del coalescente pone foco en los ancestros de una muestra de individuos trazando su historia hasta el ancestro en común más reciente (MRCA). El árbol resultante es llamado *árbol coalescente*. Estos árboles varían entre los loci, y usualmente tendrán topologías distintas a la de la filogenia de poblaciones, pero son muy informativos en relación a la historia poblacional. Además, el valor esperado de los tiempos de coalescencia y las longitudes de las ramas se pueden calcular fácilmente bajo distintos modelos demográficos neutrales.

En la sección 3.1 vimos que F_{ST} puede ser interpretado en términos de tiempos esperados de coalescencia:

$$F_{ST} = \frac{\mathbb{E}(T_B) - \mathbb{E}(T_W)}{\mathbb{E}(T_B)}$$

donde $\mathbb{E}(T_B)$ y $\mathbb{E}(T_W)$ son los tiempos de coalescencia esperados de dos individuos muestreados en poblaciones distintas (between), y en una única población (within) respectivamente. Dada la relación entre F_2 y F_{ST} (ecuación 3.16), una expresión análoga existe para $F_2(P_1, P_2)$. Para explicitarla haremos lo siguiente. En primer lugar, consideraremos F_2 para dos muestras de tamaño 1, es decir, dos individuos. En función de esto expresaremos F_2 para tamaños arbitrarios

de muestras. Finalmente obtendremos una expresión independiente del tamaño muestral haciendo tender a infinito el tamaño de la muestra.

Consideremos un locus con dos alelos A_1, A_2. En una muestra de tamaño 2, sea I_i la indicatriz de que el individuo i tenga el alelo A_1. Es decir, I_i es la frecuencia del alelo A_1 en la población P_i que contiene solo al individuo i. Luego

$$(I_i - I_j)^2 = \begin{cases} 0 & \text{si } I_i = I_j \\ 1 & \text{si } I_i \neq I_j \end{cases} \tag{3.19}$$

es otra indicatriz. Entonces

$$F_2(I_i, I_j) = \mathbb{E}((I_i - I_j)^2) = 0 \cdot \mathbb{P}(I_1 = I_2) + 1 \cdot \mathbb{P}(I_i \neq I_j) \tag{3.20}$$

$$= \mathbb{P}(\{\text{mutación en el camino de } i \text{ a } j\}) \tag{3.21}$$

donde $\mathbb{P}(\{\text{mutación en el camino de } i \text{ a } j\})$ es la probabilidad de que haya una mutación en el camino de i a j en el árbol genealógico.

Luego, en lugar de considerar un único individuo I_i, consideremos una muestra $\{I_{1,1}, I_{1,2}, \cdots, I_{1,n_1}\}$ de n_1 individuos de P_1. La frecuencia alélica de la muestra es

$$\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} I_{1,i}$$

Entonces

$$\begin{aligned} F_2(\hat{p}_1, I_2) &= F_2\left(\left[\frac{1}{n_1}\sum_{i=1}^{n_1} I_{1,i}\right], I_2\right) \\ &= \mathbb{E}\left[\left(\left[\frac{1}{n_1}\sum_{i=1}^{n_1} I_{1,i}\right] - I_2\right)^2\right] \\ &= \mathbb{E}\left[\frac{1}{n_1^2}\sum_{i=1}^{n_1} I_{1,i}^2 + \frac{2}{n_1^2}\sum_{i<j} I_{1,i}I_{1,j} - \frac{2}{n_1}\sum_{i=1}^{n_1} I_{1,i}I_2 + I_2^2\right] \end{aligned}$$

Usando $\frac{1}{n_1^2} = \frac{1}{n_1} - \frac{n_1-1}{n_1^2}$ y linealidad de la esperanza:

$$\begin{aligned} F_2(\hat{p}_1, I_2) &= \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n_1} I_{1,i}^2 - \frac{2}{n_1}\sum_{i=1}^{n_1} I_{1,i}I_2 + \frac{n_1}{n_1} I_2^2\right] + \mathbb{E}\left[\frac{2}{n_1^2}\sum_{i<j} I_{1,i}I_{1,j} - \frac{n_1-1}{n_1^2}\sum_{i=1}^{n_1} I_{1,i}^2\right] \\ &= \mathbb{E}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}(I_{1,i}^2 - 2I_{1,i}I_2 + I_2^2)\right] + \mathbb{E}\left[\frac{2}{n_1^2}\sum_{i<j} I_{1,i}I_{1,j} - \frac{n_1-1}{n_1^2}\sum_{i=1}^{n_1} I_{1,i}^2\right] \end{aligned} \tag{3.22}$$

Ahora bien, por un lado tenemos $F_2(I_{1,i}, I_2) = \mathbb{E}[(I_{1,i} - I_2)^2] = \mathbb{E}(I_{1,i}^2 - 2I_{1,i}I_2 + I_2^2)$. Por lo que el primer término de la suma es $\frac{1}{n_1}\sum_{i=1}^{n_1} F_2(I_{1,i}, I_2)$.

Por otro lado

$$\begin{aligned}
\sum_{i<j} F_2(I_{1,i}, I_{1,j}) &= \sum_{i<j} \mathbb{E}(I_{1,i}^2 - 2I_{1,i}I_{1,j} + I_{1,j}^2) \\
&= \sum_{i=1}^{n_1-1} \sum_{j=i+1}^{n_1} \mathbb{E}(I_{1,i}^2 - 2I_{1,i}I_{1,j} + I_{1,j}^2) \\
&= \mathbb{E}\left[\sum_{i=1}^{n_1-1} \sum_{j=i+1}^{n_1} I_{1,i}^2 - 2\sum_{i=1}^{n_1-1} \sum_{j=i+1}^{n_1} I_{1,i}I_{1,j} + \sum_{i=1}^{n_1-1} \sum_{j=i+1}^{n_1} I_{1,j}^2 \right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n_1-1} (n_1 - i) I_{1,i}^2 + \sum_{i=2}^{n_1} (i-1) I_{1,i}^2 - 2\sum_{i<j} I_{1,i}I_{1,j} \right] \\
&= \mathbb{E}\left[(n_1 - 1) \sum_{i=1}^{n_1} I_{1,i}^2 - 2\sum_{i<j} I_{1,i}I_{1,j} \right]
\end{aligned}$$

Por lo que el segundo término de 3.22 es $-\frac{1}{n_1^2}\sum_{i<j} F_2(I_{1,i}, I_{1,j})$ y vale

$$F_2(\hat{p}_1, I_2) = \frac{1}{n_1}\sum_{i=1}^{n_1} F_2(I_{1,i}, I_2) - \frac{1}{n_1^2}\sum_{i<j} F_2(I_{1,i}, I_{1,j}) \tag{3.23}$$

donde la segunda suma es sobre todos los pares en P_1.

Luego, como $F_2(\hat{p}_1, \hat{p}_2) = F_2(\hat{p}_2, \hat{p}_1)$ podemos intercambiar los términos para obtener la misma expresión para una muestra $\{I_{2,1}, I_{2,2}, \cdots, I_{2,n_2}\}$ de una segunda población P_2:

$$F_2(I_1, \hat{p}_2) = \frac{1}{n_2}\sum_{j=1}^{n_2} F_2(I_1, I_{2,j}) - \frac{1}{n_2^2}\sum_{j<k} F_2(I_{2,j}, I_{2,k}) \tag{3.24}$$

Si en 3.23 sustituimos I_2 por $\hat{p}_2$ obtenemos

$$\begin{aligned}
F_2(\hat{p}_1, \hat{p}_2) &= \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\frac{1}{n_2}\sum_{j=1}^{n_2} F_2(I_{1,i}, I_{2,j}) - \frac{1}{n_2^2}\sum_{j<k} F_2(I_{2,j}, I_{2,k}) \right] - \frac{1}{n_1^2}\sum_{i<j} F_2(I_{1,i}, I_{1,j}) \\
&= \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} F_2(I_{1,i}, I_{2,j}) - \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{n_2^2}\sum_{j<k} F_2(I_{2,j}, I_{2,k}) - \frac{1}{n_1^2}\sum_{i<j} F_2(I_{1,i}, I_{1,j}) \\
&= \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} F_2(I_{1,i}, I_{2,j}) - \frac{1}{n_1^2}\sum_{i<j} F_2(I_{1,i}, I_{1,j}) - \frac{1}{n_2^2}\sum_{j<k} F_2(I_{2,j}, I_{2,k})
\end{aligned} \tag{3.25}$$

Por tanto, podemos escribir F_2 entre dos poblaciones como el promedio de las diferencias entre individuos de distintas poblaciones, menos algunos términos de diferencias dentro de cada muestra.

La ecuación 3.25 es general, no asume nada respecto al lugar de las muestras en un árbol. En el contexto del coalescente, es útil suponer lo siguiente:

- $F_2(I_{1,x_1}, I_{2,y_1}) = F_2(I_{1,x_2}, I_{2,y_2})$ para todo par de muestras (x_1, x_2) en P_1, (y_1, y_2) en P_2.
- Todas las muestras corresponden a las hojas del árbol, por lo que podemos estimar la longitud de las ramas en términos de tiempo a un ancestro en común T_{ij}.
- La tasa de mutación es constante en cada rama, y vale $\theta/2$.

Estas tres suposiciones implican que $F_2(I_{i,k}, I_{j,l}) = \theta\mathbb{E}(T_{ij})$ para todo par de individuos de P_i, P_j [23]. Es decir, $F_2(I_{i,k}, I_{j,l})$ es el largo del camino que une a los dos individuos en el árbol genealógico, multiplicado por la probabilidad de mutación. Como la probabilidad de mutación en una rama es $\theta/2$, la probabilidad de mutación en el camino es θ.

Esto reduce 3.25 a

$$F_2(\hat{p}_1, \hat{p}_2) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{i=1}^{n_2}\theta\mathbb{E}(T_{12}) - \frac{1}{n_1^2}\sum_{i<j}\theta\mathbb{E}(T_{11}) - \frac{1}{n_2^2}\sum_{i<j}\theta\mathbb{E}(T_{22}) \tag{3.26}$$

Luego

$$\begin{aligned}\frac{1}{n_1^2}\sum_{i<j}\mathbb{E}(T_{11}) = \frac{1}{n_1^2}\sum_{i=1}^{n_1-1}\sum_{j=i+1}^{n_1}\mathbb{E}(T_{11}) &= \frac{1}{n_1^2}\sum_{i=1}^{n_1-1}(n_1 - i)\mathbb{E}(T_{11}) \\ &= \frac{1}{n_1^2}\left[(n_1-1)n_1 - \sum_{i=1}^{n_1-1} i\right]\mathbb{E}(T_{11}) \\ &= \frac{n_1(n_1-1)}{2n_1^2}\mathbb{E}(T_{11}) \\ &= \frac{1}{2}\left(1 - \frac{1}{n_1}\right)\mathbb{E}(T_{11})\end{aligned}$$

donde usamos $\sum_{i=1}^{n_1-1} i = \frac{(n_1-1)n_1}{2}$.

Análogamente tenemos que $\frac{1}{n_2^2}\sum_{i<j}\mathbb{E}(T_{22}) = \frac{1}{2}\left(1 - \frac{1}{n_2}\right)\mathbb{E}(T_{22})$.

Luego, sustituyendo en 3.26 tenemos

$$F_2(\hat{p}_1, \hat{p}_2) = \theta \times \left(\mathbb{E}(T_{12}) - \frac{1}{2}\left(1 - \frac{1}{n_1}\right)\mathbb{E}(T_{11}) - \frac{1}{2}\left(1 - \frac{1}{n_2}\right)\mathbb{E}(T_{22})\right) \tag{3.27}$$

Por otro lado, haciendo tender el número de individuos n_1 y n_2 a infinito obtenemos

$$F_2(P_1, P_2) = \lim_{n_1, n_2 \to \infty} F_2(\hat{p}_1, \hat{p}_2) = \theta\left(\mathbb{E}(T_{12}) - \frac{\mathbb{E}(T_{11}) + \mathbb{E}(T_{22})}{2}\right) \tag{3.28}$$

A diferencia del F_{ST}, el parámetro de mutación θ no se cancela. θ puede ser considerado como una constante de proporcionalidad y no cambiará las propiedades teóricas de los estadísticos F. Sin embargo, influirá en las propiedades estadísticas, ya que a mayor θ, mayor cantidad de mutaciones.

3.1.2. Estimador de F_2

Supongamos que tenemos muestras (secuencias) de dos poblaciones P_1 y P_2. Queremos estimar el valor de F_2 como estadístico calculado a partir de los datos. Es decir, queremos encontrar un estimador de $f_2(P_1, P_2) = (p_1 - p_2)^2$, donde p_1 y p_2 son las frecuencias alélicas reales (desconocidas) del alelo A_1 en las poblaciones P_1 y P_2.

En la sección anterior escribimos F_2 en función de tiempos esperados de coalescencia entre pares de poblaciones. Un estimador de f_2 puede ser derivado usando el *número promedio de diferencias entre pares de secuencias* π_{ij} como estimador de θT_{ij} [32]. Por la ecuación 3.28 este estimador es

$$\hat{f}_2(P_1, P_2) = \pi_{12} - \frac{\pi_{11} + \pi_{22}}{2} \tag{3.29}$$

Veamos cómo calcularlo a partir de un conjunto de secuencias. Ordenamos las muestras de forma tal que las primeras n_1 pertenecen a P_1, y las siguientes n_2 muestras a P_2.

La definición general (considerando varios loci) del promedio de diferencias entre pares de secuencias de las poblaciones P_1 y P_2 es

$$\pi_{12} = \frac{1}{n_1 n_2} \sum_{r=1}^{n_1} \sum_{s=n_1+1}^{n_1+n_2} k_{rs} \tag{3.30}$$

donde k_{rs} es el número de diferencias entre la secuencia r y la secuencia s.

Por otro lado, el promedio de diferencias entre pares de secuencias pertenecientes a la misma población es

$$\pi_{11} = \frac{1}{\binom{n_1}{2}} \sum_{r=1}^{n_1-1} \sum_{s=r+1}^{n_1} k_{rs} \tag{3.31}$$

$$\pi_{22} = \frac{1}{\binom{n_2}{2}} \sum_{r=1}^{n_2-1} \sum_{s=r+1}^{n_2} k_{rs} \tag{3.32}$$

La siguiente proposición nos dice que podemos escribir $\hat{f}_2(P_1, P_2)$ en función de las frecuencias alélicas muestrales $\hat{p}_i$ y los tamaños de la muestras n_i.

Proposición 9. *Sean P_1, P_2 dos poblaciones para las que tenemos muestras de tamaño n_1 y n_2 respectivamente, con frecuencias alélicas muestrales $\hat{p}_1, \hat{p}_2$ para un locus dado. Entonces*

$$\hat{f}_2(P_1, P_2) = (\hat{p}_1 - \hat{p}_2)^2 - \frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1} \tag{3.33}$$

Demostración. Como estamos en el caso de un único locus, dos muestras r, s pueden ser:

- del mismo tipo, es decir, ambas A_1 o ambas A_2, y entonces en 3.30 tenemos $k_{rs} = 0$, o bien
- distintas, y entonces $k_{rs} = 1$.

Sea j_1 el número de individuos con alelo A_1 en P_1, y j_2 el número de individuos con alelo A_1 en P_2. Agrupando los individuos en 3.30 tenemos que

$$\pi_{12} = \frac{1}{n_1 n_2} \sum_{r=1}^{n_1} \sum_{s=n_1+1}^{n_1+n_2} k_{rs} = \frac{j_1(1-j_2) + j_2(1-j_1)}{n_1 n_2} = \hat{p}_1(1-\hat{p}_2) + \hat{p}_2(1-\hat{p}_1)$$

Por otro lado

$$\pi_{11} = \frac{1}{\binom{n_1}{2}} \sum_{r=1}^{n_1-1} \sum_{s=r+1}^{n_1} k_{rs} = \frac{2j_1(1-j_1)}{n_1(n_1-1)} = 2\hat{p}_1(1-\hat{p}_1)\frac{n_1}{n_1-1}$$

donde en la última igualdad multiplicamos y dividimos por n_1 para obtener las frecuencias muestrales. Análogamente tenemos

$$\pi_{22} = 2\hat{p}_2(1-\hat{p}_2)\frac{n_2}{n_2-1}$$

Sustituyendo lo anterior en 3.29 y usando $\frac{n_1}{n_1-1} = 1 + \frac{1}{n_1-1}$ tenemos

$$\begin{aligned}
\hat{F}_2(P_1, P_2) &= \pi_{12} - \pi_{11}/2 - \pi_{22}/2 \\
&= [\hat{p}_1(1-\hat{p}_2) + \hat{p}_2(1-\hat{p}_1)] - \hat{p}_1(1-\hat{p}_1)\frac{n_1}{n_1-1} - \hat{p}_2(1-\hat{p}_2)\frac{n_2}{n_2-1} \\
&= \hat{p}_1\left(1 - 1 - \frac{1}{n_1-1}\right) + \hat{p}_2\left(1 - 1 - \frac{1}{n_2-1}\right) - 2\hat{p}_1\hat{p}_2 \\
&+ \hat{p}_1^2\left(1 + \frac{1}{n_1-1}\right) + \hat{p}_2^2\left(1 + \frac{1}{n_2-1}\right) \\
&= (\hat{p}_1 - \hat{p}_2)^2 - \frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}
\end{aligned}$$

□

Proposición 10. *El estimador*

$$\hat{f}_2(P_1, P_2) = (\hat{p}_1 - \hat{p}_2)^2 - \frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}$$

de $f_2(P_1, P_2) = (p_1 - p_2)^2$ *es insesgado.*

Demostración. Para probar que el estimador 3.33 es insesgado, calculemos $\mathbb{E}[(\hat{p}_1 - \hat{p}_2)^2]$ y veamos qué términos debemos agregar para que sea insesgado.

$$\mathbb{E}[(\hat{p}_1 - \hat{p}_2)^2] = \mathrm{Var}(\hat{p}_1 - \hat{p}_2) + [\mathbb{E}(\hat{p}_1 - \hat{p}_2)]^2 = \mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2) + (p_1 - p_2)^2$$

Lo anterior muestra que $(\hat{p}_1 - \hat{p}_2)^2$ no es un estimador insesgado de $(p_1 - p_2)^2$.

Luego, si $\hat{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$, por independencia de los x_{1i} tenemos

$$\mathrm{Var}(\hat{p}_1) = \mathrm{Var}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}\right) = \frac{1}{n_1^2}\sum_{i=1}^{n_1} \mathrm{Var}(x_{1i})$$

Como x_{1i} es el genotipo del individuo i (haploide), x_{1i} es una variable aleatoria binaria, que toma el valor 1 si el individuo es A_1, y 0 si es A_2. Luego

$$\mathbb{E}(x_{1i}) = 0 \cdot \mathbb{P}(x_{1i} = 0) + 1 \cdot \mathbb{P}(x_{1i} = 1) = p_1$$
$$\mathbb{E}(x_{1i}^2) = 0 \cdot \mathbb{P}(x_{1i}^2 = 0) + 1 \cdot \mathbb{P}(x_{1i}^2 = 1) = p_1$$

y entonces vale

$$\begin{aligned}\mathrm{Var}(\hat{p}_1) = \frac{1}{n_1^2}\sum_{i=1}^{n_1}\mathrm{Var}(x_{1i}) &= \frac{1}{n_1^2}\sum_{i=1}^{n_1}\mathrm{E}(x_{1i}) - [\mathrm{E}(x_{1i})]^2 \\ &= \frac{1}{n_1^2}\sum_{i=1}^{n_1} p_1(1-p_1) \\ &= \frac{p_1(1-p_1)}{n_1}\end{aligned}$$

Ahora bien, sea $h_1 = p_1(1-p_1)$.

Afirmación: El estimador

$$\hat{h}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}(n_1 - \sum_{i=1}^{n_1} x_{1i})}{n_1(n_1-1)}$$

de h_1 es insesgado.

En efecto:

$$\begin{aligned}\mathbb{E}(\hat{h}_1) = \mathbb{E}\left(\frac{\sum_{i=1}^{n_1} x_{1i}(n_1 - \sum_{i=1}^{n_1} x_{1i})}{n_1(n_1-1)}\right) &= \frac{n_1\sum_{i=1}^{n_1}\mathbb{E}(x_{1i}) - \mathbb{E}[(\sum_{i=1}^{n_1} x_{1i})^2]}{n_1(n_1-1)} \\ &= \frac{n_1\sum_{i=1}^{n_1}\mathbb{E}(x_{1i}) - \mathrm{Var}(\sum_{i=1}^{n_1} x_{1i}) - [\mathbb{E}(\sum_{i=1}^{n_1} x_{1i})]^2}{n_1(n_1-1)} \\ &= \frac{n_1^2 p_1 - n_1 p_1(1-p_1) - n_1^2 p_1^2}{n_1(n_1-1)} \\ &= \frac{(n_1-1)p_1(1-p_1)}{n_1-1} = p_1(1-p_1)\end{aligned}$$

Por otro lado tenemos que

$$\frac{\hat{h}_1}{n_1} = \frac{\sum_{i=1}^{n_1} x_{1i}(n_1 - \sum_{i=1}^{n_1} x_{1i})}{n_1^2(n_1-1)} = \frac{1}{n_1-1}\left(\frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}\right)\left(\frac{n_1 - \sum_{i=1}^{n_1} x_{1i}}{n_1}\right) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1}$$

Y por tanto

$$\begin{aligned}\mathbb{E}\left[(\hat{p}_1-\hat{p}_2)^2 - \frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}\right] &= \mathbb{E}\left[(\hat{p}_1-\hat{p}_2)^2 - \frac{\hat{h}_1}{n_1} - \frac{\hat{h}_2}{n_2}\right] \\ &= [(p_1-p_2)^2 + \mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2)] - \frac{\mathbb{E}(\hat{h}_1)}{n_1} - \frac{\mathbb{E}(\hat{h}_2)}{n_2} \\ &= \left[(p_1-p_2)^2 + \frac{p_1(1-p_1)}{n_1} + \frac{p_1(1-p_1)}{n_1}\right] \\ &- \frac{p_1(1-p_1)}{n_1} - \frac{p_2(1-p_2)}{n_2} \\ &= (p_1-p_2)^2\end{aligned}$$

□

3.1.3. Condicionando a la topología del árbol

Una característica importante de la ecuación 3.28 es que solo depende de los tiempos esperados de coalescencia entre pares de individuos. Por tanto, el comportamiento de F_2 puede ser caracterizado al considerar cuatro individuos, tomando dos al azar de cada población. Esto es todo lo que se necesita para estudiar la distribución conjunta de T_{12}, T_{11} y T_{22}, y por ende de F_2. Luego, por linealidad de la esperanza, podemos tomar muestras de mayor tamaño sumando los valores esperados sobre todos los posibles cuartetos.

Hay dos posibles topologías para un árbol sin raíz de una muestra de tamaño cuatro:

- Topología concordante $\mathcal{T}_c$, en la que linajes de la misma población coalescen antes (figura 3.6 izquierda).
- Topología discordante $\mathcal{T}_d$, en la que linajes de poblaciones distintas coalescen antes (figura 3.6 derecha).

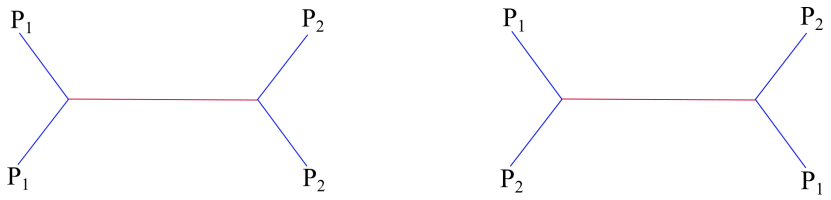

Figura 3.6: Topología concordante (izquierda) y topología discordante (derecha) en árbol sin raíz.

Si condicionamos a la topología del árbol tenemos

$$\begin{aligned} F_2(P_1, P_2) &= \mathbb{E}(F_2(P_1, P_2)|\mathcal{T}) \\ &= \mathbb{P}(\mathcal{T}_c)\mathbb{E}(F_2(P_1, P_2)|\mathcal{T}_c) + \mathbb{P}(\mathcal{T}_d)\mathbb{E}(F_2(P_1, P_2)|\mathcal{T}_d) \end{aligned}$$

La figura 3.7B corresponde a la representación gráfica de $\mathbb{E}(F_2(P_1, P_2)|\mathcal{T}_c)$ y la figura 3.7C a $\mathbb{E}(F_2(P_1, P_2)|\mathcal{T}_d)$. En este caso ignoramos el parámetro de mutación θ.

T_{12} es la longitud del camino de un individuo cualquiera de P_1 a un individuo cualquiera de P_2, y T_{11}, T_{22} corresponden a la longitud del camino entre dos individuos de P_1 y dos de P_2 respectivamente.

En el árbol tenemos dos tipos de ramas. Las ramas *externas* son aquellas que unen las hojas (poblaciones en el presente) con otros nodos. Las ramas *internas* son aquellas que unen únicamente nodos internos (poblaciones ancestrales). Veremos que las ramas externas se cancelan en la expresión de F_2 en función de tiempos esperados de coalescencia, por lo que una manera interesante de representar F_2 es en términos de ramas internas sobre todas las posibles genealogías.

Proposición 11. *Sea $\mathcal{B}_c$ la longitud promedio de la rama interna de $\mathcal{T}_c$ y $\mathcal{B}_d$ la longitud promedio de la rama interna en $\mathcal{T}_d$, no condicionadas. Entonces F_2 puede ser escrito en términos de estas longitudes como*

$$F_2(P_1, P_2) = \theta(2\mathcal{B}_c - \mathcal{B}_d) \tag{3.34}$$

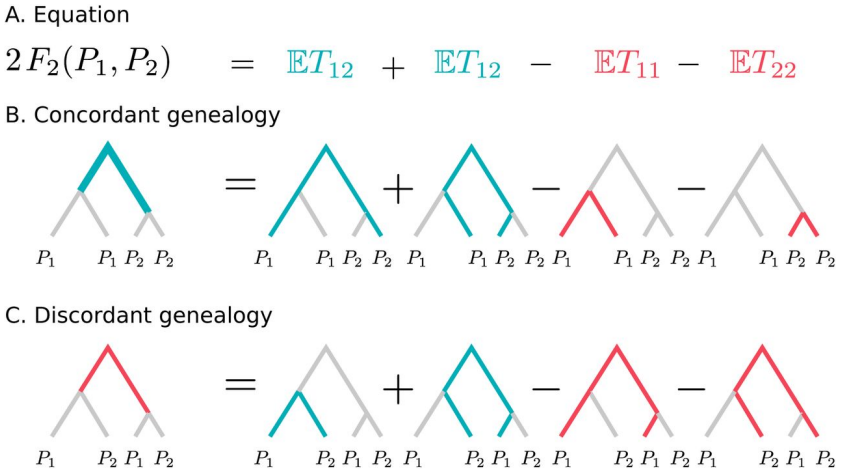

Figura 3.7: Representación de F_2 en el árbol.
Imagen tomada de [23].

Demostración. Daremos una idea de la demostración.

En $\mathcal{T}_c$ los individuos de la misma población coalescen antes. Es decir, en el árbol sin raíz tenemos dos nodos internos: el ancestro en común de los dos individuos de P_1 y el ancestro en común de los dos individuos de P_2. La rama interna es la que une estos dos ancestros.

Como podemos ver la figura 3.7B, para $\mathcal{T}_c$ la rama interna está siempre incluida en T_{12}, y nunca en T_{11} o T_{22}. Por otro lado, las ramas externas son incluidas con probabilidad $\frac{1}{2}$ en T_{12} en cualquier camino del árbol. T_{11} y T_{22} consisten sólo en ramas externas. Por tanto, las longitudes de las ramas externas se cancelan. Esto implica que $\mathbb{E}(F_2(P_1, P_2)|\mathcal{T}_c)$ puede escribirse en función de las ramas internas.

Por otro lado, en $\mathcal{T}_d$ los individuos de distintas poblaciones coalescen antes. Tenemos dos nodos internos: el ancestro en común entre un individuo de P_1 y uno de P_2, y el ancestro en común entre otro individuo de P_1 y otro de P_2. La rama interna es la que los une.

Como podemos ver la figura 3.7C, para $\mathcal{T}_d$ la rama interna siempre está incluida en T_{11} y T_{22} y solo la mitad del tiempo en T_{12}. Por tanto, las ramas internas contribuyen negativamente a F_2, pero sólo con la mitad de la magnitud que $\mathcal{T}_c$. Al igual que en $\mathcal{T}_c$, T_{11} y T_{22} contienen exactamente dos ramas externas, que se cancelan con las ramas externas de T_{12}.

Luego, si $\mathcal{B}_c$ es la longitud promedio de la rama interna de $\mathcal{T}_c$ (no condicionada) y $\mathcal{B}_d$ la longitud promedio de la rama interna en $\mathcal{T}_d$ (no condicionada), tenemos $F_2(P_1, P_2) = \theta(2\mathcal{B}_c - \mathcal{B}_d)$. □

A modo de ejemplo, supongamos que tenemos una población con dos supoblaciones P_1, P_2. Dados cuatro individuos $x_{11}, x_{12} \in P_1$ y $x_{21}, x_{22} \in P_2$, tenemos tres árboles posibles, es decir, tres formas de agrupar a los individuos en las hojas:

- $\{x_{11}, x_{12}\}$ y $\{x_{21}, x_{22}\}$ (concordante)
- $\{x_{11}, x_{21}\}$ y $\{x_{12}, x_{22}\}$ (discordante)
- $\{x_{11}, x_{22}\}$ y $\{x_{12}, x_{21}\}$ (discordante)

donde la longitud de las ramas es independiente de la topología. Si en la población total no hay estructura, los tres árboles tienen la misma probabilidad, y entonces $\mathcal{T}_d$ es dos veces más probable que $\mathcal{T}_c$. Por ende, $\mathcal{B}_d = 2\mathcal{B}_c$ y F_2 será cero, como es esperado para poblaciones sin estructura.

3.1.4. Prueba de arborescencia

En las secciones anteriores interpretamos $F_2(P_1, P_2)$ como una medida de disimilitud entre dos poblaciones P_1 y P_2.

Muchas aplicaciones consideran decenas de poblaciones simultáneamente, con el objetivo de inferir entre cuáles ha ocurrido mezcla. En este contexto la estrategia consiste en calcular el estadístico $F_2(P_i, P_j)$ para cada par, combinarlos en una matriz de disimilitud, y preguntarnos si esa matriz es consistente con un árbol. Por tanto, una *prueba de mezcla* se puede considerar como una *prueba de arborescencia* (consistencia con un árbol).

Ahora bien, para que una matriz de disimilitud sea consistente con un árbol, se tienen que satisfacer dos propiedades:

- La longitud de todas las ramas debe ser positiva.

  Esto no es estrictamente necesario para los árboles filogenéticos, algunos algoritmos pueden devolver longitudes de ramas negativas. Sin embargo, en nuestro caso las ramas tienen una interpretación de deriva genética, y la deriva genética negativa no tiene un sentido biológico, por lo que longitudes negativas deberían ser interpretadas como una violación a los supuestos del modelo.

- Para todo conjunto de cuatro poblaciones P_i, P_j, P_k y P_l se tiene que cumplir que

  $$F_2(P_i, P_j) + F_2(P_k, P_l) \leq \max\{F_2(P_i, P_k) + F_2(P_j, P_l), F_2(P_i, P_l) + F_2(P_j, P_k)\} \quad (3.35)$$

  Esto es, si comparamos sumas de pares de distancias, dos de esas sumas serán iguales, y no menores a la tercera.

  Esta propiedad es usualmente llamada *condición de los cuatro puntos*, o *teorema fundamental de la filogenética* [2]. Intuitivamente, se puede entender notando que en un árbol de cuatro hojas dos de las sumas incluirán a la rama interna, mientras que la tercera no, y por eso será más corta. Es por eso que la condición de los cuatro puntos puede traducirse como: *cualquier árbol de cuatro hojas tiene a lo sumo una rama interna.*

Ahora bien, resulta que las pruebas de mezcla basadas en los estadísticos F_3 y F_4 que veremos en las siguientes secciones pueden interpretarse como tests de estas propiedades: el test F_3 puede ser interpretado como una prueba de positividad de una rama, y el F_4 como una prueba de la condición de los cuatro puntos. Por tanto, el funcionamiento de estos dos estadísticos de prueba se puede interpretar en términos de propiedades fundamentales de los árboles filogenéticos.

Como la consistencia con un árbol implica que todo subconjunto de poblaciones es también consistente con un árbol, el rechazo de arborescencia para sub-árboles de tamaño 3 (para F_3) o 4 (para F_4) es suficiente para rechazar arborescencia para todo el conjunto [28]. Además, las pruebas en estos subconjuntos identifican a las poblaciones involucradas en los posibles eventos de mezcla.

3.2. Estadístico F_3

En la sección anterior vimos distintas interpretaciones de F_2, como longitudes de ramas en árboles, como una superposición de caminos o en términos de ramas internas. Además, lo definimos en función de distintas cantidades que miden la deriva genética, y en particular en función de los tiempos esperados de coalescencia.

En esta sección daremos resultados análogos para el estadístico F_3 que definiremos a continuación, y nos centraremos en sus dos aplicaciones más usuales: *estadístico F_3 del grupo externo* y *prueba de mezcla*.

Definición 2. *Dado un locus con dos alelos A_1 y A_2, sean p_1, p_2 y p_X las frecuencias alélicas de A_1 en las poblaciones P_1, P_2 y P_X respectivamente. El estadístico F_3 se define como*

$$F_3(P_X; P_1, P_2) = F_3(p_X; p_1, p_2) = \mathbb{E}[(p_X - p_1)(p_X - p_2)] \tag{3.36}$$

Notar que P_1 y P_2 son intercambiables, pero P_X no lo es.

La siguiente proposición nos dice que F_3 puede escribirse en función de F_2.

Proposición 12.

$$F_3(P_X; P_1, P_2) = \frac{1}{2}\left(F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2)\right) \tag{3.37}$$

Demostración. Esto surge fácilmente de la identidad

$$(p_1 - p_2)^2 = ((p_X - p_2) - (p_X - p_1))^2 = (p_X - p_2)^2 + (p_X - p_1)^2 - 2(p_X - p_2)(p_X - p_1)$$

□

La ecuación 3.37 es conocida como el *producto Gromov* en filogenética, y vale para cualquier métrica en un árbol.

La proposición anterior nos permite escribir F_3 en función de las distintas medidas de deriva genética. Por ejemplo, combinándola con la ecuación 3.28 tenemos una expresión de F_3 en función de los tiempos esperados de coalescencia:

Proposición 13.

$$F_3(P_X; P_1, P_2) = \frac{\theta}{2}\left(\mathbb{E}(T_{1X}) + \mathbb{E}(T_{2X}) - \mathbb{E}(T_{12}) - \mathbb{E}(T_{XX})\right) \tag{3.38}$$

Demostración.

$$\begin{aligned} F_3(P_X; P_1, P_2) &= \frac{\theta}{2}\left(\mathbb{E}(T_{1X}) - \frac{\mathbb{E}(T_{11}) + \mathbb{E}(T_{XX})}{2}\right) + \frac{\theta}{2}\left(\mathbb{E}(T_{2X}) - \frac{\mathbb{E}(T_{22}) + \mathbb{E}(T_{XX})}{2}\right) \\ &\quad - \frac{\theta}{2}\left(\mathbb{E}(T_{12}) - \frac{\mathbb{E}(T_{11}) + \mathbb{E}(T_{22})}{2}\right) \\ &= \frac{\theta}{2}\left(\mathbb{E}(T_{1X}) + \mathbb{E}(T_{2X}) - \mathbb{E}(T_{12}) - \mathbb{E}(T_{XX})\right) \end{aligned}$$

□

De manera análoga, la expresión en términos de varianzas se obtiene combinando las ecuaciones 3.15 y 3.37.

Proposición 14.

$$F_3(P_X; P_1, P_2) = \mathrm{Var}(p_X) + \mathrm{Cov}(p_1, p_2) - \mathrm{Cov}(p_1, p_X) - \mathrm{Cov}(p_2, p_X) \tag{3.39}$$

Demostración.

$$\begin{aligned} F_3(P_X; P_1, P_2) &= \frac{1}{2}[\mathrm{Var}(p_X - p_1) + \mathrm{Var}(p_X - p_2) - \mathrm{Var}(p_1 - p_2)] \\ &= \frac{1}{2}[\mathrm{Var}(p_X) + \mathrm{Var}(p_1) - 2\mathrm{Cov}(p_1, p_X) + \mathrm{Var}(p_X) + \mathrm{Var}(p_2) - 2\mathrm{Cov}(p_2, p_X) \\ &\quad - \mathrm{Var}(p_1) - \mathrm{Var}(p_2) + 2\mathrm{Cov}(p_1, p_2)] \\ &= \mathrm{Var}(p_X) + \mathrm{Cov}(p_1, p_2) - \mathrm{Cov}(p_1, p_X) - \mathrm{Cov}(p_2, p_X) \end{aligned}$$

□

Ahora veamos las aplicaciones de F_3.

3.2.1. Prueba de mezcla

F_3 fue originalmente motivado y utilizado como un test de mezcla [28]. La hipótesis nula es que los datos fueron generados por un árbol, y por tanto F_3 debe ser no negativo, como veremos a continuación. En caso de que sea negativo, se toma la hipótesis alternativa a favor de un grafo de mezcla.

Teorema 5. *Sean P_1, P_2, P_3 poblaciones consistentes con un árbol. Entonces se cumple que $F_3(P_1; P_2, P_3), F_3(P_2; P_1, P_3), F_3(P_3; P_1, P_2) \geq 0$.*

Demostración. Sean P_1, P_2, P_3 poblaciones distintas que descienden de una población ancestral R (figura 3.8). En este caso tenemos tres posibles árboles:

1. $\{\{P_1, P_2\}, P_3\}$ (figura 3.8 superior izquierda)
2. $\{\{P_1, P_3\}, P_2\}$ (figura 3.8 superior derecha)
3. $\{\{P_2, P_3\}, P_1\}$ (figura 3.8 inferior)

Probaremos que $F_3(P_1; P_2, P_3) \geq 0$. Notar que para este estadístico P_2, P_3 son intercambiables, por lo que es suficiente probarlo para los árboles 1 y 3.

Supongamos que estamos en el caso 1. Es decir, P_1 y P_2 descienden de una población ancestral X, y a su vez X y P_3 descienden de una población ancestral R. Sean x y r las frecuencias alélicas en las poblaciones X y R respectivamente. Entonces

$$\begin{aligned} F_3(P_1; P_2, P_3) &= \mathbb{E}[(p_1 - p_2)(p_1 - p_3)] \\ &= \mathbb{E}[(p_1 - x - (p_2 - x))(p_1 - x - (p_3 - x))] \\ &= \mathbb{E}[(p_1 - x)^2] - \underbrace{\mathbb{E}[(p_1 - x)(p_3 - x)]}_{A} - \underbrace{\mathbb{E}[(p_2 - x)(p_1 - x)]}_{0} + \underbrace{\mathbb{E}[(p_2 - x)(p_3 - x)]}_{B} \end{aligned}$$

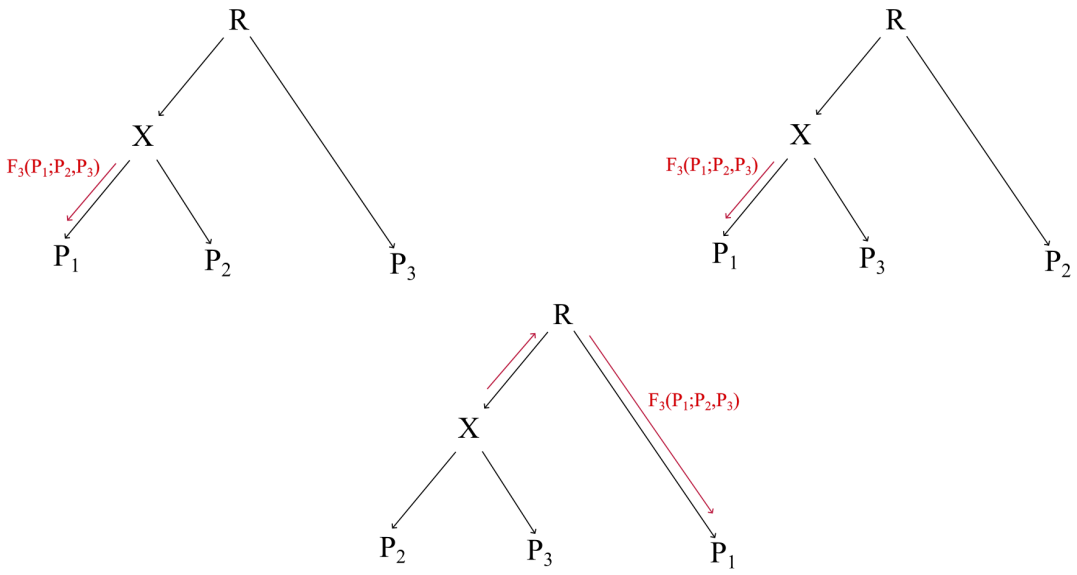

Figura 3.8: Posibles árboles de tres poblaciones.

donde $\mathbb{E}[(p_2 - x)(p_1 - x)] = 0$ por la propiedad 3.4 de independencia de los incrementos. Además

$$\begin{aligned} B = \mathbb{E}[(p_2 - x)(p_3 - x)] &= \mathbb{E}[(p_2 - x)(p_3 - r - (x - r)] \\ &= \mathbb{E}[(x - p_2)(x - r)] - \mathbb{E}[(x - p_2)(r - p_3)] \end{aligned}$$

La propiedad 3.4 implica $\mathbb{E}[(x - p_2)(r - p_3)] = 0$, y como $\mathbb{E}(p_2|x) = x$, y $\mathbb{E}(x|r) = \mathbb{E}(p_2|r) = r$, tenemos

$$\begin{aligned} \mathbb{E}[(x - p_2)(x - r)] &= \mathbb{E}(\mathbb{E}[(x^2 - xr - p_2x - p_2r)|x]) \\ &= \mathbb{E}[x^2 - x\mathbb{E}(p_2|x) - \mathbb{E}(xr - p_2r|x)] \\ &= \mathbb{E}[\mathbb{E}(p_2r - xr|r)] \\ &= \mathbb{E}[r\mathbb{E}(p_2|r) - r\mathbb{E}(x|r)] \\ &= 0 \end{aligned} \tag{3.40}$$

donde en la línea 1 condicionamos a x, y en la línea 3 a r.

Luego

$$B = \mathbb{E}[(p_2 - x)(p_3 - x)] = \mathbb{E}[(x - p_2)(x - r)] - \mathbb{E}[(x - p_2)(r - p_3)] = 0 \tag{3.41}$$

Análogamente se tiene $\mathbb{E}[(p_1 - x)(x - p_3)] = 0$.

Por tanto

$$F_3(P_1; P_2, P_3) = \mathbb{E}[(p_1 - x)^2] \geq 0$$

Ahora veamos el caso 3 (figura 3.8 inferior). Tenemos

$$\begin{aligned} F_3(P_1;P_2,P_3) &= \mathbb{E}[(p_1-p_2)(p_1-p_3)] \\ &= \mathbb{E}[(p_1-x-(p_2-x))(p_1-x-(p_3-x))] \\ &= \mathbb{E}[(p_1-x)^2] - \mathbb{E}[(p_1-x)(p_3-x)] - \mathbb{E}[(p_2-x)(p_1-x)] + \underbrace{\mathbb{E}[(p_2-x)(p_3-x)]}_{0} \end{aligned}$$

Por otro lado, utilizando los mismos argumentos que en 3.41 obtenemos que

$$\mathbb{E}[(p_1-x)(p_3-x)] = \mathbb{E}[(p_2-x)(p_1-x)] = 0$$

y nuevamente

$$F_3(P_1;P_2,P_3) = \mathbb{E}[(p_1-x)^2] \geq 0 \tag{3.42}$$

De forma análoga obtenemos que $F_3(P_2;P_1,P_3), F_3(P_3;P_1,P_2) \geq 0$. □

Del teorema y la demostración anterior podemos deducir varios resultados.

1) Interpretación de F_3 como longitud de una rama externa.

La ecuación 3.42 nos muestra la representación de F_3 como una rama externa en un árbol. En todos los casos tenemos que $F_3(P_1;P_2,P_3) = \mathbb{E}[(p_1-x)^2]$, es decir, $F_3(P_1;P_2,P_3)$ es la longitud de la rama externa que une P_1 con la población X.

En la figura 3.8 (superiores izquierda y derecha) P_1 es más cercana a una de las poblaciones P_i, y $F_3(P_1;P_2,P_3)$ será la longitud de la rama externa que une P_1 con X (ancestro de P_1 y P_i).

En el caso de la figura 3.8 inferior, las poblaciones P_2 y P_3 son más cercanas, y $F_3(P_1;P_2,P_3)$ será la longitud de la rama externa que une P_1 con X (ancestro de P_2 y P_3).

2) Interpretación de F_3 como superposición de caminos.

Recordando la interpretación como superposición de caminos realizada para F_2 tenemos

$$\begin{aligned} F_3(P_1;P_2,P_3) = \mathbb{E}[\overbrace{(p_1-p_2)}^{\text{camino 1}}\overbrace{(p_1-p_3)}^{\text{camino 2}}] &= \mathbb{E}[(p_1-p_2-\mathbb{E}(p_1-p_2))(p_1-p_3-\mathbb{E}(p_1-p_3))] \\ &= \mathrm{Cov}(p_1-p_2, p_1-p_3) \end{aligned}$$

Es decir, $F_3(P_1;P_2,P_3)$ se interpreta como la superposición del camino de P_1 a P_2 y de P_1 a P_3, que coincide con la interpretación anterior como longitud de la rama externa.

3) F_2 es una métrica en el árbol de poblaciones.

Corolario 1. *F_2 es una métrica en el árbol poblacional.*

Demostración. F_2 es simétrico y no negativo por definición. Además, por la aditividad de F_2 vale

$$F_2(P_1,P_2) = F_2(P_0,P_1) + F_2(P_0,P_2) = 0 \iff F_2(P_0,P_1) = 0 = F_2(P_0,P_2)$$

Y esto vale si y solamente si no ha habido deriva de P_0 a P_1 y de P_0 a P_2. Es decir, si en el árbol los nodos coinciden $P_0 = P_1 = P_2$.

Queda por probar que se cumple la desigualdad triangular. Sean P_1, P_2, P_X poblaciones en el árbol. Despejando en la definición 3.37 obtenemos que

$$F_2(P_1, P_2) = F_2(P_X, P_1) + F_2(P_X, P_2) - 2F_3(P_X; P_1, P_2) \tag{3.43}$$

Y entonces F_2 es una métrica en el árbol si y sólo si $F_3(P_X; P_1, P_2) \geq 0$, pues en ese caso cumple la desigualdad triangular:

$$F_2(P_1, P_2) = F_2(P_X, P_1) + F_2(P_X, P_2) - 2F_3(P_X; P_1, P_2) \leq F_2(P_X, P_1) + F_2(P_X, P_2)$$

Por el teorema 5 vale lo deseado.

□

La negatividad de uno de los estadísticos F_3 sugiere la existencia de un escenario de mezcla de poblaciones. En el modelo de mezcla más simple (figura 3.9), una población ancestral R se divide en P_1' y P_2' en tiempo t_r. En tiempo t_1 las poblaciones se mezclan para formar P_X', tal que con probabilidad α los individuos en P_X' pertenecen a P_1', y con probabilidad $\beta = 1 - \alpha$ pertenecen a P_2'. Las poblaciones P_1, P_2, P_X en el presente son resultado de la deriva luego del evento de mezcla.

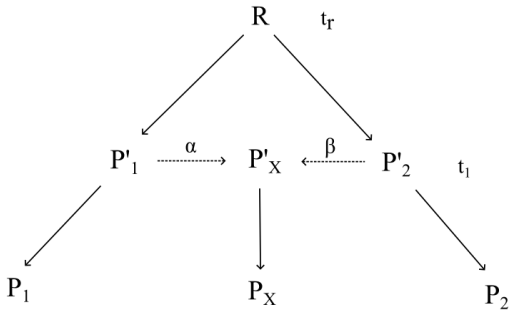

Figura 3.9: Grafo de mezcla.

En este caso tenemos el siguiente resultado.

Teorema 6. *$F_3(P_X; P_1, P_2)$ es negativo si se cumple*

$$\frac{t_1}{t_r} < 2\alpha(1 - \alpha) \tag{3.44}$$

Observación: Cuando decimos que dos individuos coalescen antes de t_1, estamos mirando hacia el pasado. Es decir, nos referimos a que el ancestro en común es posterior a t_1 en el tiempo.

Demostración. Asumiremos que todas las poblaciones tienen tamaño 1. Esto implica que el tiempo esperado de coalescencia de dos individuos que pertenecen a la misma población es 1 (sección 2.4.1). Por 3.38 tenemos

$$F_3(P_X; P_1, P_2) \propto \mathbb{E}(T_{1X}) + \mathbb{E}(T_{2X}) - \mathbb{E}(T_{12}) - \mathbb{E}(T_{XX})$$

donde $\mathbb{E}(T_{ij})$ es el tiempo esperado de coalescencia de un individuo tomado de P_i y un individuo tomado de P_j. Para calcularlo debemos considerar los posibles caminos que pueden tomar los dos individuos en el grafo, ponderando por la probabilidad de tomarlos.

La figura 3.16 representa el grafo bajo las hipótesis del teorema. Partiendo de P_X, un individuo puede tomar el camino de la izquierda con probabilidad α, o el de la derecha con probabilidad β. En el primer caso tenemos que el individuo tiene ancestro en P_1' en tiempo t_1, y en el segundo caso tiene ancestro en P_2'.

Calculemos $\mathbb{E}(T_{1X})$. Para esto, tomamos al azar un individuo de P_1 y uno de P_X. Con probabilidad α el individuo de P_X toma el camino de la izquierda. En este caso, los individuos pueden coalescer en tiempo t_1, o luego. Si la coalescencia no se da en P_1', el tiempo esperado de coalescencia es 1 por hipótiesis (pues ya se encuentran en la misma población).

Por otro lado, el individuo de P_X toma el camino de la derecha con probabilidad β. En este caso, los individuos pueden coalescer en tiempo t_r, o luego. Si la coalescencia no se da en R, el tiempo esperado de coalescencia es 1 por hipótesis (pues ya se encuentran en la misma población). Por tanto

$$\mathbb{E}(T_{1X}) = \alpha(t_1 + 1) + \beta(t_r + 1) = \alpha t_1 + \beta t_r + 1$$

Análogamente tenemos

$$\mathbb{E}(T_{2X}) = \alpha t_r + \beta t_1 + 1$$

Para calcular $\mathbb{E}(T_{12})$ notar que un individuo de P_1 y uno de P_2 coalescen en tiempo t_r, o después, y el tiempo esperado de coalescencia una vez que se encuentran en R es 1. En este caso no hay caminos alternativos. Luego

$$\mathbb{E}(T_{12}) = t_r + 1$$

Por último, calculemos $\mathbb{E}(T_{XX})$. Dos individuos de P_X pueden seguir el mismo camino o caminos distintos en el grafo. Partiendo de P_X, con probabilidad α^2 siguen ambos el camino de la izquierda. El tiempo esperado de coalescencia de estos individuos es 1 por hipótesis. Análogamente, con probabilidad β^2 los dos individuos siguen el camino de la derecha. Ahora bien, los individuos toman caminos distintos en el grafo con probabilidad $2\alpha\beta$, pero estos individuos no pueden coalescer antes de t_r: el ancestro en común más reciente de dos individuos que toman caminos distintos debe pertenecer a R (tiempo t_r) o a una población anterior a R en el tiempo. Si la coalescencia no se da una vez que se encuentran en R, el tiempo esperado de coalescencia es 1. Luego

$$\mathbb{E}(T_{XX}) = \alpha^2 \cdot 1 + \beta^2 \cdot 1 + 2\alpha\beta(t_r + 1)$$

Entonces

$$\begin{aligned} F_3(P_X; P_1, P_2) &\propto (\alpha t_1 + \beta t_r + 1) + (\alpha t_r + \beta t_1 + 1) - t_r - 1 - \alpha^2 - \beta^2 - 2\alpha\beta(t_r + 1) \\ &= (\alpha + \beta)t_1 + (\alpha + \beta)t_r - t_r + 1 - (\alpha + \beta)^2 - 2\alpha\beta t_r \\ &= t_1 - 2\alpha(1 - \alpha)t_r \end{aligned}$$

Luego $F_3(P_X; P_1, P_2)$ es negativo si

$$\frac{t_1}{t_r} < 2\alpha(1 - \alpha)$$

donde la tasa de mutación θ no influye pues es positiva.

□

Por tanto, F_3 es efectivo para detectar mezcla de poblaciones si:

- La proporción de mezcla α es cercana a 50 %. Esto se debe a que $\alpha(1-\alpha)$ se maximiza en $\alpha = \frac{1}{2}$, o sea, cuando P'_X es mezcla de P'_1, P'_2 en iguales proporciones.
- El ratio entre el tiempo del contacto secundario (t_1) y el tiempo de la división original (t_r) es chico. Es decir, el evento de mezcla es reciente relativo al tiempo de la primera división.

Condicionando a la topología del árbol.

Una condición más general para la negatividad de F_3 se obtiene considerando las ramas internas de las posibles topologías de los árboles, como lo hicimos para F_2 en la sección 3.2.3.

Como la ecuación 3.38 incluye a $\mathbb{E}(T_{XX}), \mathbb{E}(T_{1X}), \mathbb{E}(T_{2X})$ y $\mathbb{E}(T_{12})$, para estudiar la distribución conjunta necesitamos dos individuos de P_X, uno de P_1, y uno de P_2. El caso minimal nuevamente contiene cuatro individuos.

Además, P_1 y P_2 son intercambiables en el estadístico $F_3(P_X; P_1, P_2)$, por lo que tenemos dos posibles topologías para el árbol sin raíz de cuatro individuos:

- Topología concordante $\mathcal{T}_c$, en la que los individuos de P_X coalescen antes (figura 3.10 izquierda).
- Topología discordante $\mathcal{T}_d$, en la que los individuos de P_X coalescen antes con los individuos de P_1 y P_2 (figura 3.10 derecha).

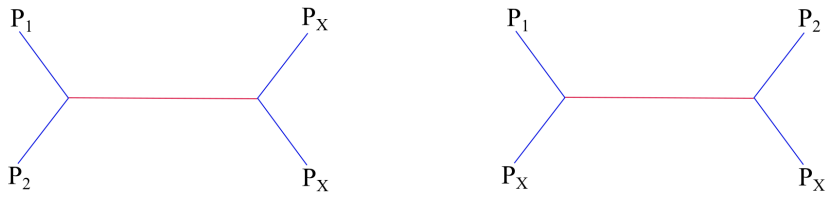

Figura 3.10: Topología concordante (izquierda) y topología discordante (derecha) en árbol sin raíz.

Si condicionamos a la topología del árbol tenemos

$$\begin{aligned} F_3(P_X; P_1, P_2) &= \mathbb{E}(F_3(P_X; P_1, P_2)|\mathcal{T}) \\ &= \mathbb{P}(\mathcal{T}_c)\mathbb{E}(F_3(P_X; P_1, P_2)|\mathcal{T}_c) + \mathbb{P}(\mathcal{T}_d)\mathbb{E}(F_3(P_X; P_1, P_2)|\mathcal{T}_d) \end{aligned}$$

La figura 3.11B corresponde a la representación gráfica de $\mathbb{E}(F_3(P_X; P_1, P_2)|\mathcal{T}_c)$ y la figura 3.11C a $\mathbb{E}(F_3(P_X; P_1, P_2)|\mathcal{T}_d)$.

T_{12} es la longitud del camino de un individuo cualquiera de P_1 a un individuo cualquiera de P_2. De la misma forma, T_{iX} con $i \in \{1, 2\}$ es la longitud del camino de un individuo cualquiera de P_i a un individuo cualquiera de P_X, y T_{XX} es la longitud del camino entre dos individuos de P_X. En la figura 3.11 los términos y caminos azules corresponden a contribuciones positivas al estadístico, mientras que los términos y caminos rojos corresponden a contribuciones negativas. Las hojas representan individuos tomados al azar en las poblaciones.

Recordemos que las ramas externas son aquellas que unen las hojas con otros nodos y las internas unen únicamente nodos internos. Al igual que para F_2, veremos que las ramas externas se cancelan, por lo que F_3 puede ser escrito en función de las ramas internas en las topologías.

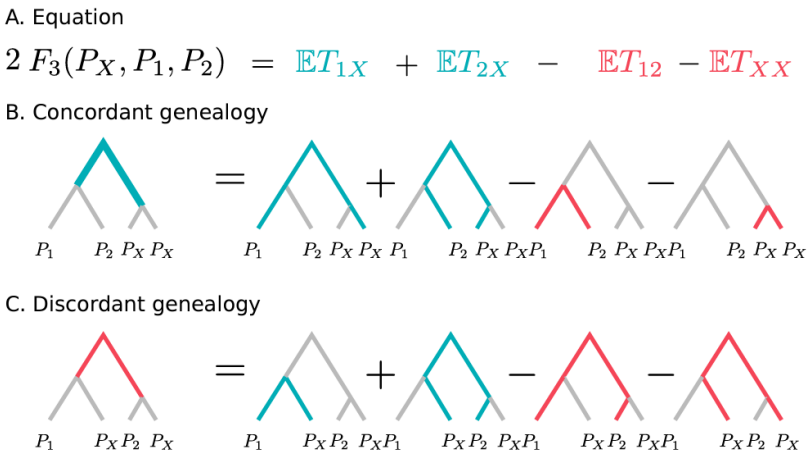

Figura 3.11: Representación de F_3 en el árbol.
Imagen tomada de [23].

Proposición 15. *Sea $\mathcal{B}_c$ la longitud promedio de la rama interna de $\mathcal{T}_c$ y $\mathcal{B}_d$ la longitud promedio de la rama interna en $\mathcal{T}_d$, no condicionadas. Entonces F_2 puede ser escrito en términos de estas longitudes como*

$$F_3(P_X; P_1, P_2) = \theta(2\mathcal{B}_c - \mathcal{B}_d) \tag{3.45}$$

La demostración es idéntica que para F_2. La idea es que en la topología concordante (figura 3.11 B) la contribución de la rama interna es positiva y con peso 2, mientras que en la topología discordante (figura 3.11 C) la contribución es negativa y con peso 1, y las ramas externas se cancelan en ambas topologías.

Al igual que para F_2, si todos los individuos pertenecen a una población (no hay estructura), $\mathcal{T}_d$ es dos veces más probable que $\mathcal{T}_c$, y todas las ramas medirán lo mismo, por lo que $F_3(P_X; P_1, P_2)$ será 0.

Corolario 2. $F_3(P_X; P_1, P_2) < 0$ *si*

$$2\mathcal{B}_c < \mathcal{B}_d \tag{3.46}$$

Es decir, para que F_3 sea negativo $\mathcal{B}_d$ deberá ser más de dos veces más larga que $\mathcal{B}_c$.

3.2.2. Grupo externo

Una aplicación simple de F_3 son las estadísticas F_3 de grupo externo [27].

El objetivo es encontrar, para una población desconocida P_U, la población más cercana perteneciente a un conjunto de k poblaciones $\{P_i : i = 1, \cdots, k\}$. Para todo i calculamos $F_3(P_O; P_U, P_i)$, donde P_O es una población externa (llamada grupo externo), que se asume que diverge ampliamente de P_U y del resto de las poblaciones en el conjunto. Esto mide la deriva genética compartida (o el camino compartido) entre P_U y las poblaciones del conjunto considerado. Usando 3.37

$$F_3(P_O; P_U, P_i) = \frac{1}{2}\left(F_2(P_O, P_U) + F_2(P_O, P_i) - F_2(P_U, P_i)\right)$$

por lo que valores altos de F_3 implican cercanía entre las poblaciones P_U y P_i ($F_2(P_U; P_i)$ pequeño en relación a los otros dos términos). Es decir, $F_3(P_O; P_U, P_i) < F_3(P_O; P_U, P_j)$ implicaría que P_U es más cercana a P_j que a P_i.

Podemos visualizar esto en la figura 3.12, en la que P_U está más cerca de P_1 que de P_4 en el árbol, y por tanto $F_3(P_O; P_U, P_1)$ es mayor a $F_3(P_O; P_U, P_4)$. En otras palabras, la deriva compartida es representada como el camino compartido por las poblaciones P_U y P_1 hasta P_O, y tiene mayor longitud.

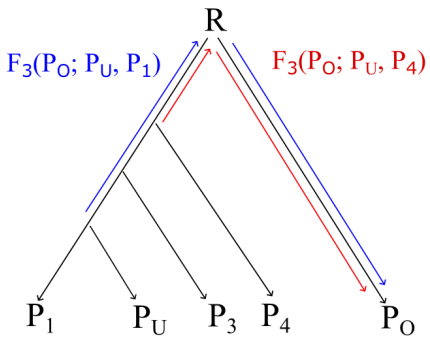

Figura 3.12: Representación de $F_3(P_O; P_U, P_1)$ (camino rojo) y $F_3(P_O; P_U, P_4)$ (camino azul)

Usando la ecuación 3.38, el estadístico del grupo externo puede ser escrito como

$$F_3(P_O; P_U, P_i) \propto \mathbb{E}(T_{UO}) + \mathbb{E}(T_{iO}) - \mathbb{E}(T_{Ui}) - \mathbb{E}(T_{OO}) \tag{3.47}$$

donde $\mathbb{E}(T_{UO})$ y $\mathbb{E}(T_{OO})$ no dependen de P_i. Además, si P_O es realmente un grupo externo, entonces todos los $\mathbb{E}(T_{iO})$ deberían valer lo mismo, pues pares de individuos del conjunto de poblaciones y de la población externa pueden coalescer solo una vez que están en la población ancestral. Por tanto, solo el término $\mathbb{E}(T_{Ui})$ variará entre las distintas poblaciones del conjunto, lo que sugiere que usar $F_3(P_O; P_U, P_i)$ es equivalente a utilizar el número de diferencias por pares π_{iU}, que utilizamos para estimar $\mathbb{E}(T_{Ui})$. De esta manera, en el conjunto $\{P_i : i = 1, \cdots, k\}$ la población más cercana será aquella que presente menos diferencias con P_U.

3.3. Estadístico F_4

Definición 3. *Dado un locus con dos alelos A_1 y A_2, sean p_1, p_2, p_3 y p_4 las frecuencias alélicas de A_1 en las poblaciones P_1, P_2, P_3 y P_4 respectivamente. El estadístico F_4 se define como*

$$F_4(P_1, P_2; P_3, P_4) = F_4(p_1, p_2; p_3, p_4) = \mathbb{E}[(p_1 - p_2)(p_3 - p_4)] \tag{3.48}$$

Los estadísticos F_3 y F_2 pueden definirse como casos especiales de F_4:

$$F_4(P_1, P_2; P_1, P_4) = \mathbb{E}[(p_1 - p_2)(p_1 - p_4)] = F_3(P_1; P_2, P_4)$$

$$F_4(P_1, P_2; P_1, P_2) = \mathbb{E}[(p_1 - p_2)(p_1 - p_2)] = F_2(P_1, P_2)$$

La siguiente proposición nos dice que F_4 se puede escribir como una combinación lineal de F_3, y también como una combinación lineal de F_2.

Proposición 16.

$$F_4(P_1, P_2; P_3, P_4) = F_3(P_1; P_2, P_4) - F_3(P_1; P_2, P_3) \tag{3.49}$$
$$= \frac{1}{2}(F_2(P_1, P_4) + F_2(P_2, P_3) - F_2(P_1, P_3) - F_2(P_2, P_4)) \tag{3.50}$$

Demostración. Escribiendo $p_3 - p_4 = p_1 - p_4 - (p_1 - p_3)$ tenemos

$$\begin{aligned}(p_1 - p_2)(p_3 - p_4) &= (p_1 - p_2)(p_1 - p_4 - (p_1 - p_3)) \\ &= (p_1 - p_2)(p_1 - p_4) - (p_1 - p_2)(p_1 - p_3)\end{aligned}$$

Luego, al tomar esperanza y usar 3.37

$$\begin{aligned}F_4(P_1, P_2; P_3, P_4) &= F_3(P_1; P_2, P_4) - F_3(P_1; P_2, P_3) \\ &= \frac{1}{2}(F_2(P_1, P_2) + F_2(P_1, P_4) - F_2(P_2, P_4) \\ &- F_2(P_1, P_2) - F_2(P_1, P_3) + F_2(P_2, P_3)) \\ &= \frac{1}{2}(F_2(P_1, P_4) + F_2(P_2, P_3) - F_2(P_1, P_3) - F_2(P_2, P_4))\end{aligned}$$

□

La proposición anterior nos permite escribir F_4 en función de tiempos esperados de coalescencia.

Proposición 17.

$$F_4(P_1, P_2; P_3, P_4) = \frac{\theta}{2}\big(\mathbb{E}(T_{14}) + \mathbb{E}(T_{23}) - \mathbb{E}(T_{13}) - \mathbb{E}(T_{24})\big) \tag{3.51}$$

Demostración. Combinando las ecuaciones 3.38 y 3.49 tenemos

$$\begin{aligned}F_4(P_1, P_2; P_3, P_4) &= \frac{\theta}{2}\big(\mathbb{E}(T_{12}) + \mathbb{E}(T_{14}) - \mathbb{E}(T_{24}) - \mathbb{E}(T_{11}) - \mathbb{E}(T_{12}) - \mathbb{E}(T_{13}) + \mathbb{E}(T_{23}) + \mathbb{E}(T_{11})\big) \\ &= \frac{\theta}{2}\big(\mathbb{E}(T_{14}) + \mathbb{E}(T_{23}) - \mathbb{E}(T_{13}) - \mathbb{E}(T_{24})\big)\end{aligned}$$

□

3.3.1. Dos interpretaciones de F_4: longitud de rama y test

Hay 4! = 24 formas de colocar las poblaciones en la definición 3.48 de $F_4(P_1, P_2; P_3, P_4)$. Sin embargo, cuatro posibles permutaciones de argumentos darán el mismo resultado:

$$F_4(P_1, P_2; P_3, P_4) = F_4(P_2, P_1; P_4, P_3) = F_4(P_3, P_4; P_1, P_2) = F_4(P_4, P_3; P_2, P_1) \tag{3.52}$$

Entonces sólo seis valores serán distintos. Además, en valor absoluto habrá solo tres distintos, pues

$$F_4(P_1, P_2; P_3, P_4) = -F_4(P_2, P_1; P_3, P_4) \tag{3.53}$$

De estos tres valores, uno podrá ser escrito a partir de los otros dos, dejando solamente dos independientes:

$$\begin{aligned} F_4(P_1,P_2;P_3,P_4)+F_4(P_1,P_3;P_4,P_2) &= F_3(P_1;P_2,P_4)-F_3(P_1;P_2,P_3) \\ &\quad + F_3(P_1;P_3,P_2)-F_3(P_1;P_3,P_4) \\ &= F_3(P_1;P_2,P_4)-F_3(P_1;P_3,P_4) \\ &= F_4(P_1,P_4;P_3,P_2) \end{aligned} \tag{3.54}$$

Como para un árbol fijo hay dos índices F_4 independientes, tenemos dos interpretaciones distintas para los índices F_4.

Es engorroso que la interpretación de F_4 dependa del orden de los argumentos. Para dejar clara la intención, en lugar de cambiar los argumentos para las dos interpretaciones, se introducen los supraíndices (T) (test) y (B) (branch length - longitud de rama):

$$F_4^{(T)}(P_1,P_2;P_3,P_4) = F_4(P_1,P_2;P_3,P_4) \tag{3.55}$$

$$F_4^{(B)}(P_1,P_2;P_3,P_4) = F_4(P_1,P_3;P_2,P_4) \tag{3.56}$$

Para justificar la notación, veamos la interpretación de F_4 en ambos casos.

Al igual que para F_2 y F_3, podemos interpretar F_4 como una superposición de caminos en el árbol.

$$\begin{aligned} F_4(P_1,P_2;P_3,P_4) = \mathbb{E}[\overbrace{(p_1-p_2)}^{\text{camino 1}}\overbrace{(p_3-p_4)}^{\text{camino 2}}] &= \mathbb{E}[(p_1-p_2-\mathbb{E}(p_1-p_2))(p_3-p_4-\mathbb{E}(p_3-p_4))] \\ &= \mathrm{Cov}(p_1-p_2,p_3-p_4) \end{aligned}$$

Es decir, $F_4(P_1,P_2;P_3,P_4)$ se interpreta como la superposición del camino de P_1 a P_2 y de P_3 a P_4.

Ahora bien, consideremos el árbol de la figura 3.13. En este caso $F_4(P_1,P_2;P_3,P_4)=0$ pues los caminos azules (camino de P_1 a P_2 y de P_3 a P_4) no se superponen. Esta es la interpretación de F_4 como test.

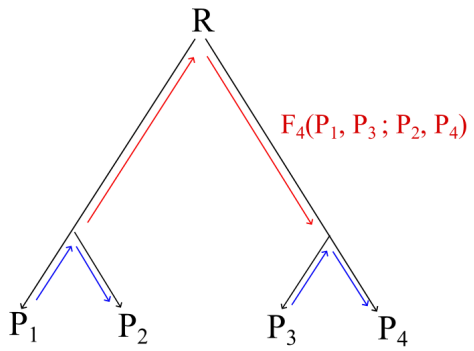

Figura 3.13: Interpretación de F_4 como test y logitud de rama

Por otro lado, $F_4(P_1,P_3;P_2,P_4)$ mide la superposición de los caminos de P_1 a P_3 y de P_2 a P_4, que es la rama interna en la figura 3.13 (camino rojo), y por tanto será positivo. Esta es la interpretación de F_4 como longitud de rama.

3.3.2. Condición de los cuatro puntos

Como adelantamos en la sección 3.1.4, F_4 está estrechamente relacionado con la condición de los cuatro puntos. El siguiente teorema fue probado por Buneman en 1971 [1, 2].

Teorema 7. *Un grafo es un árbol si y solo si es conexo, no contiene triángulos y la distancia gráfica asociada d satisface la condición de los cuatro puntos:*

$$d(x,y) + d(z,t) \leq \text{máx}\{d(x,z) + d(y,t), d(x,t) + d(y,z)\} \tag{3.57}$$

Demostración. Ver [1, 2] □

Observación: La condición de los cuatro puntos es más fuerte que la desigualdad triangular. En efecto, si en 3.57 tomamos $z = t$:

$$d(x,y) + \underbrace{d(z,z)}_{0} \leq \text{máx}\{d(x,z) + \underbrace{d(y,z)}_{\geq 0}, d(x,z) + \underbrace{d(z,z)}_{0}\} = d(x,z) + d(y,z)$$

En la sección 3.2.2 probamos que F_2 es una distancia en un árbol (corolario 1). El teorema anterior implica que para toda permutación de las poblaciones se satisface la condición de los cuatro puntos:

$$F_2(P_1,P_2) + F_2(P_3,P_4) \leq \text{máx}\{F_2(P_1,P_3) + F_2(P_2,P_4), F_2(P_1,P_4) + F_2(P_2,P_3)\} \tag{3.58}$$

Es decir, dos de las sumas deben ser iguales y no menores a la tercera. A continuación veremos que la condición de los cuatro puntos implica que al menos uno de los valores F_4 es cero, y los otros serán iguales en valor absoluto.

Teorema 8. *Sean P_1, P_2, P_3 y P_4 poblaciones consistentes con un árbol. Entonces para alguna permutación de los índices se cumple*

$$F_4(P_i,P_j;P_k,P_l) = 0 \tag{3.59}$$

$$F_4(P_i,P_k;P_j,P_l) = C \tag{3.60}$$

$$F_4(P_i,P_l;P_k,P_j) = -C \tag{3.61}$$

Demostración. Sabemos que en el árbol F_2 es una distancia y por tanto vale la condición de los cuatro puntos. Sin pérdida de generalidad asumamos que

$$F_2(P_1,P_2) + F_2(P_3,P_4) \leq F_2(P_1,P_3) + F_2(P_2,P_4)$$
$$F_2(P_1,P_3) + F_2(P_2,P_4) = F_2(P_1,P_4) + F_2(P_2,P_3)$$

Sustituyendo en las tres posibles ecuaciones de F_4:

$$F_4(P_1,P_2;P_3,P_4) = \frac{1}{2}(F_2(P_1,P_4) + F_2(P_2,P_3) - F_2(P_1,P_3) - F_2(P_2,P_4)) = 0$$

$$F_4(P_1,P_3;P_2,P_4) = \frac{1}{2}(F_2(P_1,P_4) + F_2(P_3,P_2) - F_2(P_1,P_2) - F_2(P_3,P_4)) = C$$

$$\begin{aligned} F_4(P_1, P_4; P_3, P_2) &= \frac{1}{2}(F_2(P_1, P_2) + F_2(P_4, P_3) - F_2(P_1, P_3) - F_2(P_4, P_2)) \\ &= \frac{1}{2}(F_2(P_1, P_2) + F_2(P_3, P_4) - (F_2(P_1, P_4) + F_2(P_3, P_2))) \\ &= -C \end{aligned}$$

□

Notar que el recíproco no es cierto. Si

$$\begin{aligned} F_2(P_1, P_2) + F_2(P_3, P_4) &> F_2(P_1, P_3) + F_2(P_2, P_4) \\ F_2(P_1, P_3) + F_2(P_2, P_4) &= F_2(P_1, P_4) + F_2(P_2, P_3) \end{aligned}$$

la condición de los cuatro puntos no se cumple, y sin embargo $F_4(P_1, P_2; P_3, P_4) = 0$, y los otros dos valores seguirán teniendo la misma magnitud.

En resumen, si los datos son consistentes con un árbol, uno de los estadísticos ($F_4^{(T)}$) será 0 y los otros dos serán iguales en valor absoluto. El estadístico positivo ($F_4^{(B)}$) será la longitud de la rama interna.

3.3.3. Árboles

La interpretación de F_4 en árboles dependerá de si estamos considerando la definición como estadístico de prueba $F_4^{(T)}$, o como longitudes de ramas $F_4^{(B)}$.

Consideremos $F_4^{(T)} = F_4(P_1, P_2; P_4, P_3) \propto \mathbb{E}(T_{13}) + \mathbb{E}(T_{24}) - \mathbb{E}(T_{14}) - \mathbb{E}(T_{23})$. Para estudiar la distribución conjunta necesitamos un individuo tomado al azar de cada población.

Sean P_1, P_2, P_3 y P_4 individuos tomados de las poblaciones correspondientes. Existen tres topologías posibles, es decir, tres formas de agrupar a los individuos en las hojas:

- $\{\{P_1, P_2\}, \{P_3, P_4\}\}$ correspondiente a la topología concordante $\mathcal{T}_c$ (figura 3.15B)
- $\{\{P_1, P_3\}, \{P_2, P_4\}\}$ correspondiente a la topología discordante $\mathcal{T}_{d_1}$ (figura 3.15C)
- $\{\{P_1, P_4\}, \{P_2, P_3\}\}$ correspondiente a la topología discordante $\mathcal{T}_{d_2}$ (figura 3.15D)

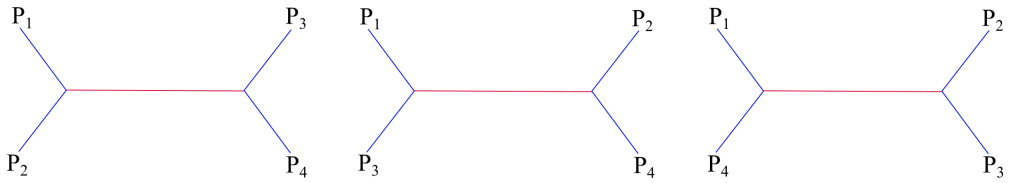

Figura 3.14: Topología concordante $\mathcal{T}_c$ (izquierda), discordante $\mathcal{T}_{d_1}$ (centro) y discordante $\mathcal{T}_{d_2}$ para un árbol sin raíz.

Si condicionamos a la topología del árbol tenemos

$$\begin{aligned} F_4(P_1, P_2; P_4, P_3) &= \mathbb{E}(F_4(P_1, P_2; P_4, P_3)|\mathcal{T}) \\ &= \mathbb{P}(\mathcal{T}_c)\mathbb{E}(F_4(P_1, P_2; P_4, P_3)|\mathcal{T}_c) + \mathbb{P}(\mathcal{T}_{d_1})\mathbb{E}(F_4(P_1, P_2; P_4, P_3)|\mathcal{T}_{d_1}) \\ &+ \mathbb{P}(\mathcal{T}_{d_2})\mathbb{E}(F_4(P_1, P_2; P_4, P_3)|\mathcal{T}_{d_2}) \end{aligned}$$

La figura 3.15B corresponde a la representación gráfica de $\mathbb{E}(F_4(P_1, P_2; P_4, P_3)|\mathcal{T}_c)$, la figura 3.15C a $\mathbb{E}(F_4(P_1, P_2; P_4, P_3)|\mathcal{T}_{d_1})$ y la figura 3.15D a $\mathbb{E}(F_4(P_1, P_2; P_4, P_3)|\mathcal{T}_{d_2})$.

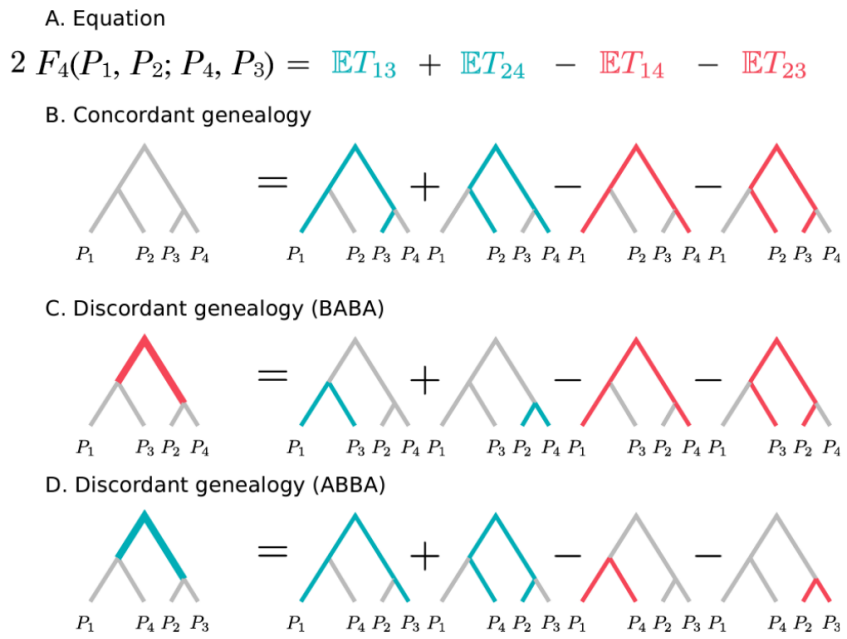

Figura 3.15: Representación de F_4 en el árbol.
Imagen tomada de [23].

En la figura 3.15 los términos y caminos azules corresponden a contribuciones positivas al estadístico, mientras que los términos y caminos rojos corresponden a contribuciones negativas. Siguiendo el mismo razonamiento que para F_2 y F_3, las longitudes de las ramas externas se cancelan para las tres topologías, por lo que $F_4^{(T)}$ puede escribirse en función de las ramas internas. En el caso de la topología concordante las ramas internas también se cancelan, por lo que su contribución al estadístico será 0. Las dos topologías discordantes contribuyen con peso -1 y +1 respectivamente.

Proposición 18. *Sea $\mathcal{B}_{d_1}$ es la longitud promedio de la rama interna en $\mathcal{T}_{d_1}$ y $\mathcal{B}_{d_2}$ la longitud promedio de la rama interna en $\mathcal{T}_{d_2}$. Entonces $F_4^{(T)}$ puede ser escrito en términos de estas longitudes como*

$$F_4^{(T)} = \theta(\mathcal{B}_{d_2} - \mathcal{B}_{d_1}) \tag{3.62}$$

Consideremos ahora $F_4^{(B)} = F_4(P_1, P_3; P_2, P_4) \propto \mathbb{E}(T_{14}) + \mathbb{E}(T_{23}) - \mathbb{E}(T_{12}) - \mathbb{E}(T_{34})$. Con los mismos argumentos que antes se puede demostrar que en las tres topologías las ramas externas se cancelan. En el caso de la topología concordante, la contribución de la rama interna es positiva y con peso 1. Por otro lado, la contribución de la topología discordante $\mathcal{T}_{d_1}$ será 0, y de la topología discordante $\mathcal{T}_{d_2}$ será -1. Luego

$$F_4^{(B)} = \theta(\mathcal{B}_c - \mathcal{B}_d) \tag{3.63}$$

3.3.4. Prueba de rango

Una de las aplicaciones de F_4 utiliza su interpretación como longitud de una rama. Específicamente, el rango de la matriz que contiene los estadísticos F_4 calculados en todo conjunto de cuatro poblaciones es utilizado para obtener una cota inferior del número de eventos de mezcla requeridos para explicar el conjunto de datos.

Supongamos que tenemos muestras de n poblaciones. Si calculamos los índices $F_4(P_i, P_j; P_k, P_l)$ para todo conjunto de cuatro poblaciones obtenemos una matriz de tamaño $\binom{n}{2} \times \binom{n}{2}$.

$$\begin{array}{c} \\ (P_i, P_j) \end{array} \begin{array}{c} \begin{array}{c} (P_k, P_l) \\ \vdots \end{array} \\ \begin{pmatrix} \cdots & F_4(P_i, P_j; P_k, P_l) & \cdots \\ & \vdots & \end{pmatrix} \end{array}$$

Ahora bien, en un árbol sin raíz, el número de ramas internas es a lo sumo $n-3$. Como F_4 es una suma de longitudes de ramas internas, los índices F_4 deberían ser sumas de a lo sumo $n-3$ ramas, o $n-3$ componentes independientes. Esto implica que si los datos son consistentes con un árbol, el rango de la matriz es a lo sumo $n-3$.

Sin embargo, los eventos de mezcla pueden aumentar el rango de la matriz al agregar ramas internas. Por tanto, si el rango de la matriz es r, el número de eventos de mezcla es al menos $r-n+3$.

3.3.5. Proporción de mezcla

Otra aplicación de F_4 está relacionada con la estimación de las proporciones de mezcla.

Supongamos que estamos en el caso de la figura 3.16, donde P_X es mezcla de P_1 y P_2 en proporciones α y $\beta = 1-\alpha$ respectivamente. P_O y P_K son poblaciones de referencia que no contribuyen a P_X. Además, P_K es más cercana a P_1 que a P_2, y P_O es un grupo externo (lejano a todas las poblaciones). Queremos estimar α.

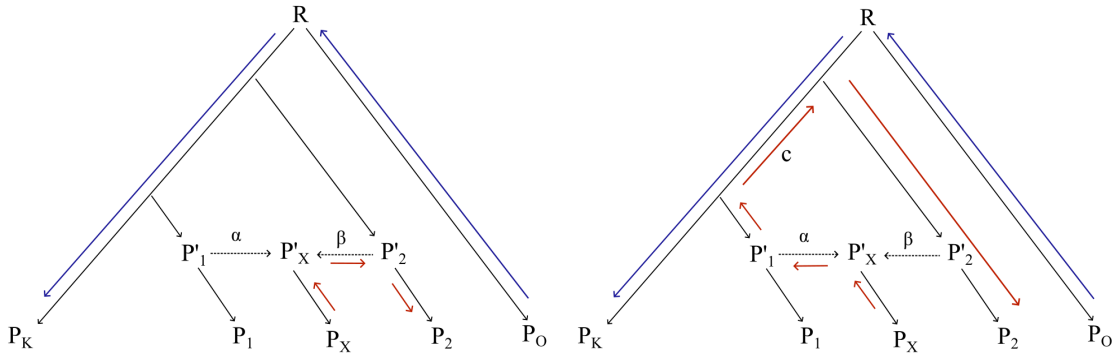

Figura 3.16: Grafo de mezcla.

Como P_X es mezcla de P_1 y P_2, tenemos dos caminos de P_X a P_2 (camino rojo en ambas figuras). Como P_K es más cercana a P_1 que a P_2, uno de ellos (izquierda) no se superpone con el camino

azul que va de P_O a P_K, por lo que su contribución a $F_4(P_O, P_K; P_X, P_2)$ será 0. Por tanto

$$F_4(P_O, P_K; P_X, P_2) = \alpha c$$

donde c es la superposición de los caminos de P_X a P_2 y de P_O a P_K en la figura derecha. Además, c es la superposición de los caminos de P_1 a P_2 y de P_O a P_K, es decir, $c = F_4(P_O, P_K; P_1, P_2)$. Luego

$$\alpha = \frac{F_4(P_O, P_K; P_X, P_2)}{F_4(P_O, P_K; P_1, P_2)} \tag{3.64}$$

Notar que α está bien definido incluso cuando P_X no es mezcla de P_1 y P_2, por lo que 3.64 tiene sentido si estamos seguros de que P_X es mezcla de dichas poblaciones. Esto lo podemos inferir, por ejemplo, usando el estadístico F_3 como prueba de mezcla.

Capítulo 4

Estadísticos F y PCA

El análisis de componentes principales (PCA) y los estadísticos F son dos de las herramientas más utilizadas para estudiar la variación genética. En las siguientes secciones derivaremos conexiones explícitas entre ellos.

4.1. Estadísticos F en el espacio S-dimensional

En el capítulo anterior vimos la definición de F_2, F_3 y F_4 como el valor esperado del producto de las diferencias entre frecuencias alélicas de poblaciones, para un único locus bialélico. En la práctica, como el promedio es un estimador de la esperanza, los estadísticos son calculados a partir de miles de loci.

Supongamos que tenemos un conjunto de poblaciones para el que tenemos datos de frecuencias de polimorfismos de un solo nucleótido (SNPs) para S loci bialélicos. Definimos x_{il} como la frecuencia de un alelo de referencia (arbitrario) para el SNP l, en la población i.

$$F_2(X_1, X_2) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})^2 \tag{4.1}$$

$$F_3(X_1; X_2, X_3) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{1l} - x_{3l}) \tag{4.2}$$

$$F_4(X_1, X_2; X_3, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{3l} - x_{4l}) \tag{4.3}$$

Las coordenadas del vector $X_i = (x_{i1}, \cdots, x_{iS})$ corresponden a las frecuencias alélicas en la población i. Como X_i será el único vector de datos para la población i, no haremos distinción entre la población y el vector que usamos para representarla. Por tanto, los estadísticos F pueden ser pensados como productos internos:

$$F_2(X_1, X_2) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})^2 = \frac{1}{S}\langle X_1 - X_2\ ,\ X_1 - X_2\rangle = \frac{1}{S}\|X_1 - X_2\|^2 \tag{4.4}$$

$$F_3(X_1; X_2, X_3) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{1l} - x_{3l}) = \frac{1}{S}\langle X_1 - X_2\ ,\ X_1 - X_3\rangle \tag{4.5}$$

$$F_4(X_1, X_2; X_3, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{3l} - x_{4l}) = \frac{1}{S}\langle X_1 - X_2\ ,\ X_3 - X_4\rangle \tag{4.6}$$

donde $\|\cdot\|$ denota la norma euclídea, y $\langle\cdot\ ,\ \cdot\rangle$ el producto interno usual. La normalización por el número de SNPs S se realiza en todos los cálculos y por ende la omitiremos.

Un inconveniente asociado a esta interpretación geométrica es que cada SNP agrega una dimensión, y los espacios de alta dimensión son difíciles de visualizar, interpretar y analizar. Además, usualmente tenemos $n \ll S$, es decir, menos poblaciones (observaciones) que SNPs (variables), lo que resulta en un aumento de la complejidad.

Afortunadamente se ha observado que la estructura poblacional suele tener baja dimensión. Es decir, si bien el espacio de frecuencias alélicas tiene dimensión S, los datos pertenecen a una variedad de dimensión $k \ll S$, resultado de la correlación existente entre los mismos.

En este sentido, una de las técnicas más utilizadas es el análisis de componentes principales (PCA), un método de reducción de la dimensionalidad mediante el cual se obtiene una representación de los datos en un espacio de baja dimensión, minimizando la pérdida de información. Esta reducción se consigue transformando las variables originales en un nuevo conjunto de variables, las componentes principales, no correlacionadas, y ordenadas de forma tal que las primeras retienen la mayor parte de la variación presente en las variables originales.

A continuación presentaremos de forma general PCA para el caso de datos genéticos. Más detalles se pueden encontrar en el Apéndice C y en [12].

Análisis de componentes principales para datos genéticos.

Supongamos que tenemos datos de frecuencias alélicas de n poblaciones $\{X_1, \cdots, X_n\}$. Sea $\mathbf{X}$ la matriz cuya entrada x_{il} es la frecuencia del alelo de referencia para el locus l, en la población i.

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1S} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nS} \end{bmatrix} \in \mathbb{R}^{n\times S}$$

Centramos la matriz $\mathbf{X}$, obteniendo una matriz $\mathbf{Y}$ tal que

$$y_{il} = x_{il} - \mu_l$$

donde μ_l es el promedio de las frecuencias alélicas del alelo de referencia para el locus l (promedio por columnas de la matriz $\mathbf{X}$).

Dado $k \ll S$, el objetivo es encontrar un conjunto ortonormal de vectores $\{v_1, \cdots, v_k\} \in \mathbb{R}^S$ que expliquen *la mayor parte de la variación en los datos*. Es decir, para $i = 1, \cdots, n$ aproximamos X_i por una combinación lineal $z_{i1}v_1 + \cdots + z_{ik}v_k$ con $z_{i1}, \cdots, z_{ik} \in \mathbb{R}$, y el elegimos los v_i que optimicen la calidad de esta aproximación en el conjunto de datos [12]. Los vectores óptimos $v_1, \cdots, v_k$ son las primeras k *componentes principales*: v_1 es la dirección de máxima varianza en el conjunto de datos, v_2 es la dirección de máxima varianza que es a su vez ortogonal a v_1, y así sucesivamente.

Si bien existen varios algoritmos para realizar PCA, el más utilizado se basa en la *descomposición de valores singulares* (SVD) de la matriz $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{4.7}$$

donde $\mathbf{U} \in \mathbb{R}^{n\times n}$ y $\mathbf{V} \in \mathbb{R}^{S\times S}$ son matrices ortogonales, y $\mathbf{\Sigma} \in \mathbb{R}^{n\times S}$ es una matriz formada por los valores singulares de $\mathbf{X}$ en su diagonal principal, ordenados de mayor a menor. Si $n < S$

$$\mathbf{Y} = \begin{pmatrix} | & & | \\ u_1 & \dots & u_n \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ 0 & & \sigma_n & \dots 0 \end{pmatrix} \begin{pmatrix} — & v_1 & — \\ & \vdots & \\ — & v_S & — \end{pmatrix}$$

donde

- La matriz $\mathbf{Y}\mathbf{Y}^T$ es simétrica y semidefinida positiva, por lo que los valores propios son todos positivos o nulos. Si el rango de $\mathbf{Y}$ es r, entonces $\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$.
- $\sigma_i = \sqrt{\lambda_i}$ es el i-ésimo valor singular de $\mathbf{Y}$.
- La columna i-ésima de $\mathbf{U}$ es el vector propio unitario u_i de la matriz $\mathbf{Y}\mathbf{Y}^T$, asociado al valor propio λ_i.
- El conjunto $\{u_1, \cdots, u_n\}$ es ortonormal, y $\{u_1, \cdots, u_r\}$ es una base ortonormal del subespacio fundamental Col($\mathbf{Y}$).
- La i-ésima fila de la matriz $\mathbf{V}^T$ es el vector unitario v_i de la matriz $\mathbf{Y}^T\mathbf{Y}$, asociado al vector propio λ_i, recordando que los valores propios de $\mathbf{Y}^T\mathbf{Y}$ y $\mathbf{Y}\mathbf{Y}^T$ coinciden.

Notar que si el rango de $\mathbf{Y}$ es r, $\lambda_i = 0$ para todo $i > r$, por lo que podemos escribir $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ con $U \in \mathbb{R}^{n\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$ y $\mathbb{V} \in \mathbb{R}^{r\times S}$ y obtenemos una representación similar, en la que dejamos fuera los valores singulares σ_i nulos y sus vectores propios u_i (columnas de $\mathbf{U}$), v_i (filas de $\mathbf{V}^T$) asociados. Esta es la representación que utilizaremos de aquí en adelante.

Intuitivamente, SVD expresa la transformación lineal representada por $\mathbf{Y}$ como una rotación ($\mathbf{V}^T$), seguida por un escalamiento ($\mathbf{\Sigma}$), seguido por otra rotación ($\mathbf{U}$).

La SVD produce una descomposición de $\mathbf{Y}$ como una suma de r matrices de rango 1

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sigma_1 \underbrace{u_1}_{n\times 1} \overbrace{v_1^T}^{1\times S} + \cdots + \sigma_n \underbrace{u_n v_n^T}_{n\times S} \tag{4.8}$$

que es equivalente a $\mathbf{Y}v_i = \sigma_i u_i$.

Se puede demostrar que $\sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T$ es la aproximación de rango k de $\mathbf{Y}$ que minimiza la distancia entre el conjunto de datos original y la proyección en el subespacio afin de dimensión k generado por $\{u_1, \cdots, u_k\}$ [12]. Si los datos están altamente correlacionados, esperamos que muchos de los valores σ_i sean pequeños, y por tanto pueden ser ignorados.

PCA nos permite aproximar puntos en el espacio de frecuencias alélicas (de dimensión alta), a través de un subespacio de dimensión k de los datos. Si consideramos todos los componentes principales, $k = n - 1$, los datos son simplemente rotados.

Existen varias formas de realizar PCA, dependiendo del tipo de estandarización de los SNPs, del tratamiento los datos faltantes o de la utilización de individuos o poblaciones como unidades

de análisis. La versión de PCA presentada en [24] y estudiada aquí es aquella que maximiza las similitudes con los estadísticos F, pero no es la versión de PCA más utilizada a en el contexto de análisis de datos genómicos. En particular, asumiremos que PCA se realiza en las frecuencias alélicas estimadas de las poblaciones, sin escalar, mientras que en la mayoría de las aplicaciones PCA se realiza con las frecuencias alélicas a nivel individuos, escaladas por la estimación de la desviación estándar de cada SNP.

En el contexto de PCA la matriz de componentes principales $\mathbf{P} = \mathbf{U\Sigma}$ (*scores*) tiene tamaño $n \times n$ contiene información sobre la estructura poblacional, es decir, de cómo están relacionadas unas muestras con otras. Las filas de $\mathbf{L} = \mathbf{V}^T$ (*loadings*) forman una base ortonormal de tamaño $n \times S$, y tienen información de cómo se relacionan los SNPs.

Alternativamente, las componentes principales pueden ser obtenidas de la descomposición en valores propios de la matriz de covarianza $\mathbf{YY}^T = \mathbf{PLL}^T\mathbf{P}^T = \mathbf{PP}^T$.

Al hacer PCA rotamos nuestros datos para revelar los ejes con mayor varianza. Sin embargo, el producto interno es invariante por rotación, y los estadísticos F pueden ser pensados como productos internos. Esto implica que podemos calcular F_2 en nuestros datos $\mathbf{X}$, en nuestros datos centrados $\mathbf{Y}$, o en cualquier base ortonormal, por ejemplo, en la de los componentes principales $\mathbf{P}$.

Teorema 9.

$$F_2(X_i, X_j) = F_2(Y_i, Y_j) = F_2(P_i, P_j). \tag{4.9}$$

Demostración. La primera igualdad es inmediata

$$F_2(X_i, X_j) = \sum_{l=1}^{S}(x_{il} - x_{jl})^2 = \sum_{l=1}^{S}((x_{il} - \mu_l) - (x_{jl} - \mu_l))^2 = F_2(Y_i, Y_j).$$

Por otro lado, utilizando $\mathbf{Y} = \mathbf{PL}$ tenemos $y_{il} = L_{1l}P_{i1} + \cdots + L_{nl}P_{in}$, entonces

$$\begin{aligned}
F_2(Y_i, Y_j) &= \sum_{l=1}^{S}(y_{il} - y_{jl})^2 \\
&= \sum_{l=1}^{S}\left(\sum_{k=1}^{n} L_{kl}P_{ik} - \sum_{k=1}^{n} L_{kl}P_{jk}\right)^2 \\
&= \sum_{l=1}^{S}\left(\sum_{k=1}^{n} L_{kl}(P_{ik} - P_{jk})\right)^2 \\
&= \sum_{l=1}^{S}\left(\sum_{k=1}^{n} L_{kl}^2(P_{ik} - P_{jk})^2 + 2\sum_{k\neq k'} L_{kl}L_{k'l}(P_{ik} - P_{jk})(P_{ik'} - P_{jk'})\right) \\
&= \sum_{k=1}^{n}\underbrace{\left(\sum_{l=1}^{S} L_{kl}^2\right)}_{1}(P_{ik} - P_{jk})^2 + 2\sum_{k\neq k'}\underbrace{\left(\sum_{l=1}^{S} L_{kl}L_{k'l}\right)}_{0}(P_{ik} - P_{jk})(P_{ik'} - P_{jk'}) \\
&= \sum_{k=1}^{n}(P_{ik} - P_{jk})^2 \\
&= F_2(P_i, P_j).
\end{aligned}$$

En la segunda línea utilizamos $\mathbf{Y} = \mathbf{PL}$. En la línea cinco utilizamos que $\mathbf{L}$ es ortogonal (i.e. $\mathbf{L}^T\mathbf{L} = \mathbf{I}$), $\mathbf{L}\mathbf{L}^T = \mathbf{I}$ y entonces $\sum_{l=1}^{S} L_{kl}^2 = 1$ y $\sum_{l=1}^{S} L_{kl}L_{k'l} = 0$ con $k \neq k'$. □

Lo anterior vale también para F_3 y F_4 pues se escriben como combinaciones lineales de F_2.

En la mayoría de las aplicaciones no se utilizan todas las componentes principales, sino que consideramos las primeras K, que explican la mayor parte de la variación genética entre las poblaciones.

$$F_2(P_i, P_j) = \underbrace{\sum_{k=1}^{K}(p_{ik} - p_{jk})^2}_{\hat{F}_2^{(K)}(P_i,P_j)} + \underbrace{\sum_{k=K+1}^{n}(p_{ik} - p_{jk})^2}_{\epsilon^{(K)}(P_i,P_j)} \tag{4.10}$$

$\hat{F}_2^{(K)}$ es la aproximación de F_2 considerando las primeras K componentes principales, y $\epsilon^{(K)}$ el error de aproximación. Omitiremos el supraíndice cuando el número exacto de componentes principales no sea relevante.

Si sumamos los cuadrados de los errores sobre todos los pares de poblaciones obtenemos

$$\sum_{i,j} \epsilon^{(K)}(P_i, P_j)^2 = \sum_{i,j}(F_2^{(K)}(P_i, P_j) - \hat{F}_2^{(K)}(P_i, P_j))^2 = \|\mathbf{F}_2 - \hat{\mathbf{F}}_2\|_F^2 \tag{4.11}$$

donde la norma Frobenius de una matriz $\mathbf{A}$ se define como

$$\|A\|_F = \sqrt{(\mathrm{tr}\mathbf{A}^T\mathbf{A})}$$

4.2. Estadísticos F y proyecciones.

El producto interno está fuertemente relacionado con la proyección ortogonal

$$proj_b a = \frac{\langle a\ ,\ b\rangle}{\|b\|^2} b \tag{4.12}$$

que es un vector colineal a b, cuya norma mide cuánto apunta el vector a en la dirección de b:

$$\langle a\ ,\ b\rangle = \|a\| \cdot \|b\| \cdot \cos\theta \tag{4.13}$$

con θ el ángulo entre a y b.

La figura 4.1 superior corresponde a la representación de los estadísticos en árboles (la raíz es arbitraria) y la inferior a la representación en un plot bidimensional de PCA. Para facilitar la comprensión, suponemos que no hay error en la aproximación en 2 componentes principales.

En 4.1 (a) $F_2(X_1, X_4)$ representa la distancia euclidea al cuadrado entre X_1 y X_4, y tenemos:

$$proj_{X_1 - X_4}(X_1 - X_4) = \frac{F_2(X_1, X_4)}{\|X_1 - X_4\|^2}(X_1 - X_4) = X_1 - X_4 \tag{4.14}$$

En 4.1 (b), $F_3(X_1; X_3, X_4)$ corresponde a la rama externa de X_1 al nodo interno que une las tres poblaciones, y es proporcional a la proyección ortogonal de $X_1 - X_3$ en $X_1 - X_4$:

$$proj_{X_1 - X_4}(X_1 - X_3) = \frac{F_3(X_1; X_3, X_4)}{\|X_1 - X_4\|^2}(X_1 - X_4) \tag{4.15}$$

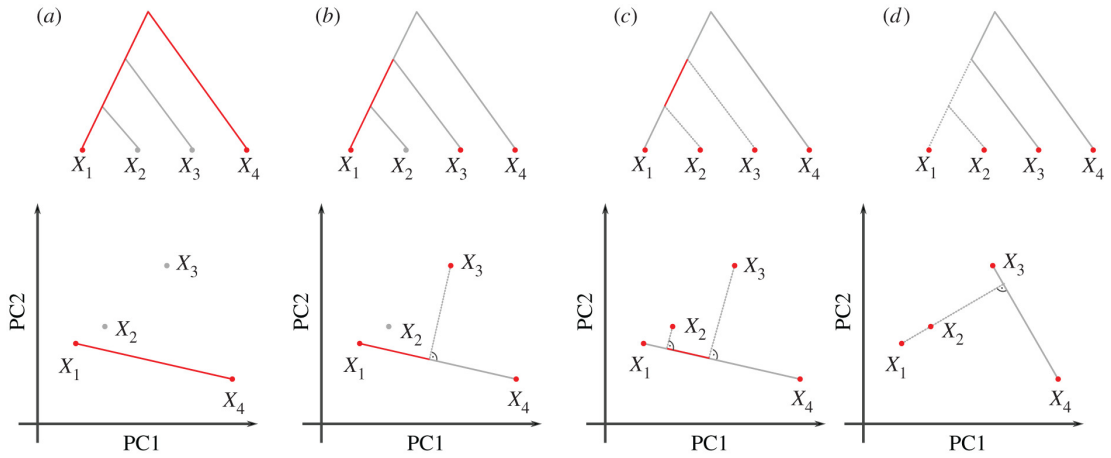

Figura 4.1: Representación de los estadísticos F en árboles y en gráficos de PCA de dos dimensiones. Imagen tomada de [24].

En 4.1 (c), $F_4(X_1, X_4; X_2, X_3)$ corresponde a la rama interna en el árbol, y es proporcional a la proyección ortogonal de $X_2 - X_3$ sobre $X_1 - X_4$:

$$proj_{X_1 - X_4}(X_2 - X_3) = \frac{F_4(X_2, X_3; X_1, X_4)}{\|X_1 - X_4\|^2}(X_1 - X_4) \tag{4.16}$$

Por último, en 4.1 (d), el camino de X_1 a X_2 no se superpone con el camino de X_3 a X_4. Por lo visto en el capítulo anterior $F_4(X_1, X_2; X_3, X_4) = 0$. Como consecuencia, $X_1 - X_2$ y $X_3 - X_4$ serán vectores ortogonales en el espacio:

$$proj_{X_3 - X_4}(X_1 - X_2) = \frac{F_4(X_1, X_2; X_3, X_4)}{\|X_3 - X_4\|^2}(X_3 - X_4) \tag{4.17}$$

Proyección de nuevas poblaciones.

Supongamos que tenemos datos de una nueva población, no incluida en la realización de PCA. Notar que

$$\mathbf{Y}\mathbf{L}^T = \mathbf{P}\mathbf{L}\mathbf{L}^T = \mathbf{P}$$

Entonces una nueva población (centrada) Y_{new} puede ser proyectada en un PCA existente simplemente multiplicando por $\mathbf{L}^T$ por derecha

$$P_{\text{proj}} = Y_{\text{new}}\mathbf{L}^T$$

La coordenada k-ésima de P_{proj} da las coordenadas de la nueva muestra en la componente principal número k.

En este caso hay un error de proyección

$$\|Y_{\text{new}} - P_{\text{proj}}\mathbf{L}\|^2 = F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}})$$

Sea X_{new} una población cualquiera, Y_{new} la población centrada, X_i una población incluida en el PCA, e Y_i la población X_i centrada. Entonces por 4.9 se tiene

$$\begin{aligned} F_2(X_i, X_{\text{new}}) &= F_2(Y_i, Y_{\text{new}}) \\ &= F_2(Y_i, P_{\text{proj}}\mathbf{L}) + F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}) \\ &= F_2(P_i, P_{\text{proj}}) + F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}) \\ &= \hat{F}_2(P_i, P_{\text{proj}}) + \epsilon(P_i, P_{\text{proj}}) + F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}) \end{aligned} \tag{4.18}$$

donde en la segunda línea usamos el teorema de Pitágoras, pues el error de proyección y la proyección son ortogonales. La mayor implicación de la ecuación 4.18 es que tanto la proyección como truncar en k componentes introducen error, y que $\hat{F}_2(P_i, P_{\text{proj}})$ será una buena aproximación de $F_2(X_i, X_{\text{new}})$ solo si ambos errores son pequeños.

Si la población Y_{new} tiene datos faltantes, se puede usar una proyección similar en la que removemos filas de Y_{new} y $\mathbf{L}$ para las que faltan datos en Y_{new}.

4.3. F_2 en el espacio PCA

El estadístico F_2 es una estimación del cuadrado de la diferencia entre las frecuencias alélicas de dos poblaciones. Vimos que en un árbol corresponde al camino entre dos poblaciones. En el espacio de frecuencias alélicas corresponde a la distancia euclídea al cuadrado, y por tanto refleja la intuición de que poblaciones cercanas estarán cerca en el espacio de frecuencias alélicas, y los estadísticos F_2 correspondientes serán pequeños.

Sin embargo, como F_2 puede escribirse como una suma de términos al cuadrado (por ende no negativos), al truncar la distancia en un gráfico de PCA será siempre una subestimación de la distancia F_2 total.

Como consecuencia, PCA podría proyectar cerca dos poblaciones con F_2 grande, lo que indicaría que esas componentes principales no son adecuadas para visualizar la relación entre ellas, y por tanto debería considerarse un número mayor de componentes principales.

En contraste, si dos poblaciones se encuentren lejos en las primeras componentes principales, está garantizado que tengan también distancia F_2 grande, ya que las componentes principales omitidas no pueden contribuir con distancias negativas.

4.4. Negatividad de F_3

En la sección 3.3.2 vimos que la negatividad de uno de los estadísticos F_3 sugiere la existencia de un escenario de mezcla de poblaciones. Una pregunta natural es: dadas dos poblaciones X_i, X_j, ¿podemos utilizar PCA para predecir para qué poblaciones X_x el estadístico $F_3(X_x; X_i, X_j)$ será negativo?

Consideremos el escenario de mezcla de la figura 4.2(a), en el que la población X_y es mezcla de las poblaciones X_2 y X_3, y la deriva genética cambia las frecuencias alélicas de la población X_y a la población X_x.

Para todo SNP i tenemos que

$$x_{yi} = \alpha x'_{2i} + (1 - \alpha)x'_{3i}$$

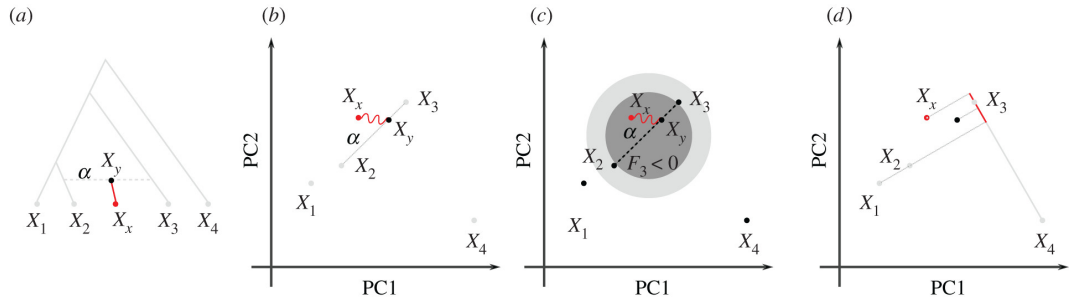

Figura 4.2: Representación de la mezcla de poblaciones en un gráfico de PCA de dos dimensiones. Imagen tomada de [24].

Es decir, las frecuencias alélicas de X_y son combinaciones lineales de las frecuencias alélicas de X_2' y X_3', poblaciones ancestrales a X_1 y X_2, que se mezclaron para formar X_y. Luego

$$X_y = \alpha X_2' + (1-\alpha)X_3'$$

Por lo que X_y caerá en el segmento que une las dos poblaciones (figura 4.2(b)), en una ubicación que se puede predecir con la proporción de mezcla α. La deriva genética cambiará las frecuencias hasta X_x, por lo que en general caerá en un punto distinto en el gráfico de PCA (figura 4.2(b)). Pero si, por ejemplo, la mezcla fuera reciente relativa al tiempo de la primera división entonces el error es pequeño

$$X_x \approx X_y = \alpha X_2' + (1-\alpha)X_3' \approx \alpha X_2 + (1-\alpha)X_3$$

Aquí estamos considerando todas las componentes principales. Si truncamos en k, agregamos el error de proyección. En particular, nuevamente para la representación gráfica estamos suponiendo que no hay error de proyección al considerar dos componentes principales.

Ahora bien, como adelantamos, dadas dos poblaciones X_1, X_2, nos interesa saber en qué región del espacio será negativo $F_3(X_x; X_1, X_2)$.

Teorema 10. *Sean X_1, X_2 poblaciones en $\mathbb{R}^S$. La región en la que F_3 es negativo es una n-bola de centro $(X_1 + X_2)/2$ y diámetro $\overline{X_1X_2}$.*

Demostración. Sin pérdida de generalidad asumamos que $X_1 = (r, 0, \cdots, 0)$ y $X_2 = (-r, 0, \cdots, 0)$. Sea $X_x = (x_1, \cdots, x_S)$. Como F_3 se puede escribir como sumas de F_2, usando (4.4) tenemos

$$\begin{aligned}
2F_3(X_x; X_1, X_2) &= \|X_x - X_1\|^2 + \|X_x - X_2\|^2 - \|X_1 - X_2\|^2 \\
&= \left[(x_1 - r)^2 + \sum_{i=2}^{n} x_i^2\right] + \left[(x_1 + r)^2 + \sum_{i=2}^{n} x_i^2\right] - (r + r)^2 \\
&= 2\left[x_1^2 + r^2 + \sum_{i=2}^{n} x_i^2\right] - 4r^2 \\
&= -2r^2 + 2\|X_x\|^2
\end{aligned}$$

Luego, $F_3(X_x; X_1, X_3) = -r^2 + \|X_x\|^2$ y $F_3(X_x; X_1, X_3) < 0$ si y solo si $\|X_x\|^2 < r^2$, es decir, si X_x pertenece a la bola de centro $\mathbf{0}$ y radio r. □

Consideremos nuevamente el caso de la figura 4.2(a). Al hacer PCA de dimensión 2, proyectamos la n-bola en un gráfico de dimensión 2. En general $\overline{X_2X_3}$ no estará alineada con las componentes principales, por lo que la bola parecerá más pequeña de lo que realmente es.

La forma más sencilla de visualizar esta geometría es proyectar $F_3(X_x; X_1, X_2)$ de un espacio de dos dimensiones en uno de dimensión uno (figura 4.3).

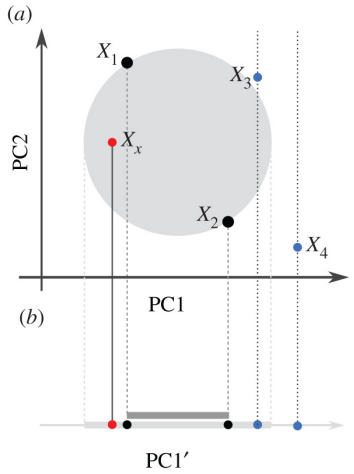

Figura 4.3: Reducción de la dimesionalidad y estadístico F_3 como prueba de mezcla.
Imagen tomada de [24]

En este ejemplo, la distancia entre X_1 y X_2 tiene una contribución sustancial de PC2, y entonces la región negativa (gris claro) es mayor a lo que podríamos predecir con una sola dimensión (línea gris oscuro).

Los puntos como X_x caerán dentro de la barra gris claro, pero no tienen porqué proyectarse en la barra gris oscuro. Sin embargo si $\hat{F}_3 \approx F_3$, la diferencia entre las áreas será pequeña.

Además, algunos puntos que están fuera de la región negativa en el espacio original (como X_3 en la figura 4.2) podrán ser proyectados dentro de la región negativa, pero tendrán F_3 positivo.

Sin embargo, la interpretación recíproca es más estricta: si una población está fuera del círculo en alguna proyección, el estadístico F_3 asociado será positivo. Este es el caso de X_4 en la figura 4.2.

Teorema 11. *Si una población X_x no pertenece a la n-bola de diámetro $\overline{X_1X_2}$ para alguna proyección bidimensional, entonces $F_3(X_x; X_1, X_2) > 0$.*

Demostración. Asumamos que el centro de la n-bola es $C = (X_1 + X_2)/2 = (c_1, c_2, \cdots, c_S)$ y $X_x = (x_1, x_2, \cdots, x_S)$. En la demostración del teorema 10 vimos que $F_3(X_x; X_1, X_2) = 0$ corresponde al borde de la n-bola de centro C y diámetro $\overline{X_1X_2}$. Si el radio de esta bola es r, entonces:

$$F_3(X_x; X_1, X_2) = \|X_x - C\|^2 - r^2 = \underbrace{(x_1 - c_1)^2 + (x_2 - c_2)^2}_{>r^2} + \underbrace{\sum_{i=3}^{S} (x_i - c_i)^2}_{\geq 0} - r^2 > 0$$

donde la condición $(x_1 - c_1)^2 + (x_2 - c_2)^2 > r^2$ se cumple siempre que X_x esté fuera del círculo obenido al proyectar la n-bola en las primeras dos dimensiones. Un argumento análogo vale para cualquier representación en dimensión baja. □

4.5. Estadísticos F_3 del grupo externo como proyecciones

En la sección 3.2.2 vimos una de las aplicaciones más comunes de F_3, como estadístico del grupo externo. Dada una muestra desconocida X_U, queremos encontrar la población más cercana en un panel de referencia $\{X_1, \cdots, X_l\}$. Esto se hace usando el estadístico $F_3(X_O; X_U, X_i)$, donde X_O es un grupo externo, es decir, una población lejana a X_U y X_i para todo $i = 1, \cdots, l$.

$F_3(X_O; X_U, X_i)$ representa la longitud de la rama entre X_O y el nodo en común entre las tres poblaciones en el estadístico (figura 4.4). Cuanto más cerca estén este nodo y X_U, mayor longitud tendrá esta rama, y más grande será F_3, lo que implica cercanía entre las poblaciones X_U y X_i.

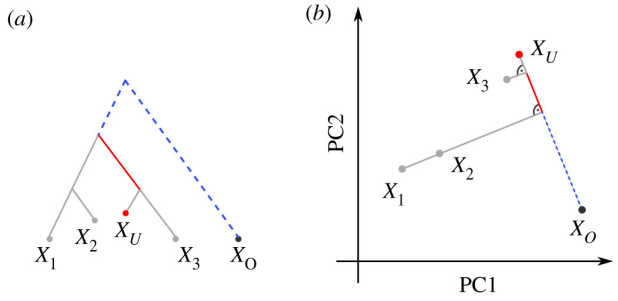

Figura 4.4: Interpretación del estadístico F_3 del grupo externo en un árbol (a) y en un gráfico de PCA (b). Imagen tomada de [24].

Para interpretar este estadístico en PCA, usamos la asociación de los estadísticos con las proyecciones (ecuación 4.12). En un gráfico de PCA podemos visualizar $F_3(X_O; X_U, X_i)$ como

$$\begin{aligned} proj_{X_U - X_O}(X_i - X_O) &= \frac{\langle X_U - X_O \ , \ X_i - X_O \rangle}{\|X_U - X_O\|^2}(X_U - X_O) \\ &= F_3(X_O; X_U, X_i)\frac{X_U - X_O}{F_2(X_O, X_U)} \end{aligned} \tag{4.19}$$

En (4.21) el único término que depende de X_i es $F_3(X_O; X_U, X_i)$. La fracción puede ser pensada como una constante de normalización, por lo que $F_3(X_O; X_U, X_i)$ es proporcional al largo del vector proyectado. Esto significa que el estadístico F_3 del grupo externo será más grande para aquella población X_i que se proyecte más lejos a lo largo del eje desde el grupo externo a la población desconocida (en figura 4.4 es X_3)

4.6. Estadísticos F_4 como ángulos

Una interpretación de F_4 en el gráfico de PCA es similar a la de F_3, como una proyección de un vector sobre otro, con la diferencia de que ahora los cuatro puntos pueden ser distintos.

Vimos que si el estadístico F_4 corresponde a una rama que no existe, entonces esperamos $F_4(X_1, X_2; X_3, X_4) = \langle X_1 - X_2, X_3 - X_4 \rangle = 0$. Esto implica que los vectores $X_1 - X_2$ y $X_3 - X_4$ son ortogonales, es decir, que X_1 y X_2 se proyectan en el mismo punto en $\overline{X_3 X_4}$ (figura 4.1 (d)).

En el caso de un grafo de mezcla, esto no es así. Las poblaciones X_y y X_x en la figura 4.2 no se proyectan al mismo punto que X_1 y X_2, lo que implica que, por ejemplo, $F_4(X_1, X_x; X_3, X_4) \neq 0$.

Como F_4 es una covarianza, su magnitud carece de interpretación. Por tanto, usualmente se utilizan coeficientes de correlación. Para F_4 podemos escribir

$$\begin{aligned} \mathrm{Cor}(X_1 - X_2, X_3 - X_4) = \frac{\mathrm{Cov}(X_1 - X_2, X_3 - X_4)}{\sqrt{\mathrm{Var}(X_1 - X_2)\mathrm{Var}(X_3 - X_4)}} &= \frac{F_4(X_1, X_2; X_3, X_4)}{\sqrt{F_2(X_1, X_2)F_2(X_3, X_4)}} \\ &= \frac{\langle X_1 - X_2 \ , \ X_3 - X_4 \rangle}{\|X_1 - X_2\| \|X_3 - X_4\|} \\ &= \cos(\phi) \end{aligned} \tag{4.20}$$

donde ϕ es el ángulo entre $X_1 - X_2$ y $X_3 - X_4$. De este modo, eventos de deriva independientes llevan a $\cos(\phi) = 0$ (ángulo es 90 grados), mientras que un ángulo cercano a cero ($\cos(\phi) \approx 1$) implica que la mayor parte de la deriva genética en esta rama es compartida.

Capítulo 5

Conclusiones

Los estadísticos F son una herramienta potente para describir la variación genética poblacional, en particular para el análisis de variación genética humana con un número grande de individuos con relaciones heterogéneas. En este trabajo vimos sus principales propiedades y aplicaciones. A continuación presentaré algunas interrogantes que podrían derivar en trabajos futuros.

El estadístico F_2 mide la deriva genética entre dos poblaciones, y se interpreta en el árbol como la longitud de la rama que las une. Vimos que se relaciona con distintas medidas de divergencia entre poblaciones, como la varianza en las frecuencias alélicas, la heterocigosidad y la probabilidad de identidad por descendencia. Una ventaja de F_2 frente al resto de las medidas es que es una distancia en el árbol, lo que nos permite probar algunos resultados como los relacionados con la condición de los cuatro puntos. Sería interesante estudiar qué otras propiedades o resultados de los estadísticos pueden derivarse de la noción de métrica en un árbol, y qué utilidad podrían tener en el contexto de análisis de datos genéticos.

Además, los estadísticos pueden escribirse en función de los tiempos esperados de coalescencia entre pares de individuos. Esto permitiría su interpretación en distintos modelos demográficos neutrales que incorporen migración: modelo de islas, stepping stone, stepping stone jerárquico, entre otros. En estos modelos, los tiempos esperados de coalescencia quedan definidos mediante sistemas de ecuaciones que dependen, por ejemplo, de la cantidad de islas y de las tasas de migración. Como los estadísticos están definidos en función de los tiempos de coalescencia, podrían calcularse bajo estos modelos, y su interpretación sería útil a la hora de hacer inferencia sobre la historia evolutiva de poblaciones que se adecuen a los mismos.

Por otro lado, la geometría de los estadísticos F conduce a una serie de interpretaciones simples de los estadísticos en los gráficos de PCA. Como PCA es usualmente realizado en etapas tempranas en el análisis de datos, esto puede ayudar a la generación de hipótesis que pueden ser evaluadas de forma más directa, típicamente usando un número bajo de poblaciones. Sin embargo, el PCA que presentamos aquí está basado en frecuencias alélicas estimadas en las poblaciones, mientras que en la mayoría de los estudios de variación genética humana se utilizan frecuencias alélicas de individuos. Por tanto, una pregunta natural sería cómo extendemos la interpretación de los estadísticos F al PCA realizado en individuos.

Por último, en este trabajo estudiamos los estadísticos básicamente bajo un modelo demográfico neutral. Sin embargo, sería de gran relevancia la incorporación de fuerzas evolutivas como la selección y la mutación, así como de otros fenómenos que contribuyen a la evolución. La inter-

pretación geométrica en un espacio vectorial nos permite pensar en las poblaciones y el flujo génico de manera continua, y en este contexto, las difusiones resultan una herramienta útil de modelado, pues se extienden de manera directa a procesos que incorporan, por ejemplo, selección y mutación. En este sentido, podría ser interesante el estudio y modelado mediante difusiones multidimensionales que incorporen estas fuerzas evolutivas, aprovechando la potencia del cálculo estocástico, pero pudiendo recuperar los estadísticos y las propiedades discretas fundamentales de los árboles filogenéticos, que son las que hacen posible la interpretación sencilla de los problemas biológicos subyacentes.

Apéndice A

Cadenas de Markov

Consideremos un conjunto **I** finito o numerable. Sea X_0, X_1, X_2, ... una sucesión de variables aleatorias que toman valores en **I**, definidas en un espacio de probabilidad $(\Omega, A, \mathbb{P})$. El conjunto **I** se denomina *espacio de estados*, y designaremos a sus elementos, los *estados* mediante las letras i, j, k, l.

Si X_0, X_1, X_2, ... una sucesión de variables aleatorias independientes, los sucesos:

$$A = \{X_0 = i_0, ..., X_n = i_n\},\ B = \{X_{n+1} = i_{n+1}, \cdots, X_{n+m} = i_{n+m}\}$$

son independientes para cualquier $n = 0, 1, 2...$ y $m = 1, 2, 3...$ y cualquier sucesión de estados $i_0, \cdots, i_{n+m}$.

La dependencia markoviana consiste en que la probabilidad del suceso B depende únicamente del valor que toma la variable aleatoria X_n, y no de los valores que toman las variables aleatorias $X_0, ..., X_{n-1}$. Es decir, si el índice de la sucesión representa el tiempo y n es el instante presente, entonces la probabilidad de un suceso en el futuro, que ocurre en los instantes $n+1, ..., n+m$, depende solamente del estado en que se encuentra la sucesión en el instante presente n, y no de los estados en los que se encontró en los instantes pasados $0, 1, \cdots, n-1$.

Definición 4. *Sea X_0, X_1, $X_2, \cdots$, una sucesión de variables aleatorias definidas en un espacio de probabilidad $(\Omega, A, \mathbf{P})$, que toman valores en un conjunto* **I** *finito o numerable.*

(a) Decimos que la sucesión X_0, X_1, $X_2, \cdots$, es una cadena de Markov si se verifica

$$\mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n, \cdots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} | X_n = i_n) \tag{A.1}$$

para todo $n = 1, 2, \cdots$ y cualquier sucesión de estados $i_0, ..., i_{n+1}$ en **I**, *siempre que $\mathbf{P}(X_n = i_n, \cdots, X_0 = i_0) > 0$. La identidad A.1 se llama propiedad de Markov.*

(b) Decimos que una cadena de Markov es homogénea en el tiempo si para todo par de estados i, j la probabilidad condicional $\mathbb{P}(X_{n+1} = j | X_n = i)$ no depende de n. Es decir, cuando se verifica

$$\mathbb{P}(X_1 = j | X_0 = i) = \mathbb{P}(X_2 = j | X_1 = i) = \cdots = \mathbb{P}(X_{n+1} = j | X_n = i) \tag{A.2}$$

En general, decimos cadena de Markov para referirnos a una cadena de Markov homogénea en el tiempo.

Matriz de transición.

Consideremos una cadena de Markov $X_0, X_1, X_2, \cdots$ con espacio de estados $\mathbf{I}$, y dos estados i, j. Definimos

$$p_{ij} = \mathbb{P}(X_1 = j | X_0 = i)$$
$$\pi_i = \mathbb{P}(X_0 = i)$$

La matriz $\mathcal{P} = (p_{ij})_{i,j \in \mathbf{I}}$ (posiblemente infinita) se denomina *matriz de transición*, y el vector $\pi = (\pi_i)_{i \in \mathbf{I}}$ *distribución inicial* de la cadena de Markov.

La matriz de transición verifica las siguientes propiedades:

(M1) $p_{ij} \geq 0$ para todo par de estados $i, j \in \mathbf{I}$.

(M2) $\sum_{j \in \mathbf{I}} p_{ij} = 1$ para todo estado $i \in \mathbf{I}$

La distribución inicial verifica las siguientes propiedades:

(D1) $\pi_i \geq 0$ para todo estado $i \in \mathbf{I}$

(D2) $\sum_{i \in \mathbf{I}} \pi_i = 1$

Matriz de transición de orden n.

Dados dos estados i, j, definimos

$$p_{ij}^n = \mathbb{P}(X_n = j | X_0 = i)$$
$$\pi_i^n = \mathbb{P}(X_n = i)$$

Llamamos *matriz de transición de orden n* a $\mathcal{P}^n = (p_{ij}^n)$, y *distribución de probabilidad en el instante n* al vector $\pi^n = (\pi_i^n)$.

Las probabilidades de transición verifican la *ecuación de Kolmogorv-Chapman*:

$$p_{ij}^{m+n} = \sum_{k \in \mathbf{I}} p_{ik}^m p_{kj}^n, \tag{A.3}$$

para todo par de índices m, n y todo par de estados $i, j \in \mathbf{I}$.

En notación matricial la ecuación A.3 es

$$\mathcal{P}^{m+n} = \mathcal{P}^m \times \mathcal{P}^n \tag{A.4}$$

En efecto, aplicando la fórmula de la probabilidad total y la propiedad de Markov A.1 se obtiene:

$$\begin{aligned} p_{ij}^{m+n} &= \mathbb{P}(X_{m+n} = j | X_0 = i) \\ &= \sum_{k \in \mathbb{I}} \mathbb{P}(X_{m+n} = j, X_m = k | X_0 = i) \\ &= \sum_{k \in \mathbf{I}} \mathbb{P}(X_{m+n} = j | X_m = k) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_{k \in \mathbf{I}} p_{ij}^m p_{kj}^n \end{aligned}$$

donde en la última igualdad utilizamos que $\mathbb{P}(X_{m+n} = j | X_m = k) = \mathbb{P}(X_n = j | X_0 = k)$

Análogamente, para todo par de índices m, n y todo estado i, la distribución de probabilidad en el instante n cumple:

$$\begin{aligned} \pi_j^{m+n} &= \mathbb{P}(X_{m+n} = j) \\ &= \sum_{k \in \mathbf{I}} \mathbb{P}(X_{m+n} = j, X_m = k) \\ &= \sum_{k \in \mathbf{I}} \mathbb{P}(X_{m+n} = j | X_m = k) \mathbb{P}(X_m = k) \\ &= \sum_{k \in \mathbf{I}} \pi_k^m p_{kj}^n \end{aligned} \tag{A.5}$$

En notación matricial:

$$\pi^{m+n} = \pi^m \times \mathcal{P}^n \tag{A.6}$$

En particular, de A.4 resulta

$$\mathcal{P}^n = \underbrace{\mathcal{P} \times \cdots \times \mathcal{P}}_{\text{n}}$$

Es decir, la matriz de transición de orden n es la potencia n-ésima de la matriz de transición $\mathcal{P}$.

La distribución de probabilidad en el instante n se obtiene mediante la fórmula

$$\pi^n = \pi \times \mathcal{P}^n \tag{A.7}$$

Proposición 19. *Sea $X_0, X_1, \cdots$, una cadena de Markov en un espacio de estados $\mathbf{I}$, una matriz de transición $\mathcal{P} = (p_{ij})$ y distribución inicial $\pi = (\pi_i)_{i \in \mathbf{I}}$. Entonces se cumple que*

$$\mathbb{P}(X_n = i_n, \cdots, X_0 = i_0) = p_{i_{n-1} i_n} \cdots p_{i_0 i_1} \pi_{i_0} \tag{A.8}$$

Demostración. Aplicando la definición de probabilidad condicional tenemos que

$$\begin{aligned} &\mathbb{P}(X_n = i_n, \cdots, X_0 = i_0) = \\ &\qquad \mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) \mathbb{P}(X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) \end{aligned}$$

Ahora bien, por A.1

$$\mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) = \mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1}) = p_{i_{n-1} i_n}$$

Luego

$$\mathbb{P}(X_n = i_n, \cdots, X_0 = i_0) = p_{i_{n-1} i_n} \mathbb{P}(X_{n-1} = i_{n-1}, \cdots, X_0 = i_0)$$

Aplicando $n-1$ veces lo anterior concluimos que

$$\begin{aligned} \mathbb{P}(X_n = i_n, \cdots, X_0 = i_0) &= p_{i_{n-1} i_n} \cdots p_{i_0 i_1} \mathbb{P}(X_0 = i_0) \\ &= p_{i_{n-1} i_n} \cdots p_{i_0 i_1} \pi_{i_0} \end{aligned}$$

□

La ecuación A.8 nos permite calcular $\mathbb{P}(X_0 \in A_0, \cdots, X_n \in A_n)$ para subconjuntos arbitrarios $A_0, \cdots, A_n$ de $\mathbf{I}$. A su vez, esto nos permite calcular la distribución del vector aleatorio $(X_{n_1}, \cdots, X_{n_k})$, que llamamos *distribución finito-dimensional* de la cadena de Markov.

Por tanto, la proposición anterior nos dice que las distribuciones finito-dimensionales de una cadena de Markov quedan determinadas por su espacio de estados, su matriz de transición y su distribución inicial.

Clasificación de estados.

Definición 5. *Sean i, j dos estados de una cadena de Markov. Decimos que*

(a) de i se llega a j y escribimos $i \to j$ si existe un natural $n \geq 0$ que verifica $p_{ij}^n > 0$.

(b) los estados i, j se comunican si cada estado es alcanzable desde el otro. Es decir, $i \to j$ y $j \to i$. Se denota como $i \leftrightarrow j$

Es decir, $i \to j$ si empezando desde el estado i es posible (con probabilidad positiva) entrar en el estado j en un número finito de transiciones.

Proposición 20. *Dada una cadena de Markov con espacio de estados $\mathbf{I}$, la relación $i \leftrightarrow j$ define una relación de equivalencia en $\mathbf{I} \times \mathbf{I}$*

Demostración. Veamos que cumple las propiedades reflexiva, simétrica y transitiva.

- *Reflexividad.* Dado $i \in \mathbf{I}$ tenemos que $p_{ii}^0 = \mathbb{P}(X_0 = i | X_0 = i) = 1$.
  Luego existe $n \in \mathbb{N}$ tal que $i \leftrightarrow i$.
- *Simetría.* Por definición $i \leftrightarrow j \iff j \leftrightarrow i$
- *Transitividad.* Veamos que si $i \to j$ y $j \to k$, entonces $i \to k$.
  En efecto, si $i \to j$ y $j \to k$, existen $r, s > 0$ tales que $p_{ij}^r > 0$ y $p_{jk}^s > 0$.
  Luego, por la ecuación A.3 de Kolmogorov-Chapman tenemos

  $$p_{ik}^{r+s} = \sum_{l \in \mathbf{I}} p_{il}^r p_{lk}^s \geq p_{ij}^r p_{jk}^s > 0 \Rightarrow i \to k$$

  De forma análoga se prueba que $k \to i$.

□

Definición 6. *Decimos que un estado i es esencial si para todo j tal que $i \to j$ se verifica $j \to i$. En caso contrario decimos que i es un estado no esencial.*

Como consecuencia de la proposición anterior tenemos que el conjunto de los estados esenciales de una cadena de Markov se descompone en una unión disjunta de subconjuntos, llamados *clases irreducibles*, tales que i, j están en la misma clase si y sólo si $i \leftrightarrow j$. Si el espacio de estados es la única clase irreducible, decimos que la cadena de Markov es *irreducible*.

Recurrencia.

Consideremos una cadena de Markov $X_0, X_1, \cdots$, con espacio de estados I, matriz de transición $\mathcal{P}$, y distribución inicial π.

Definición 7. *Definimos la probabilidad de la primera transición de i a j en n pasos como*

$$f_{ij}^n = \mathbb{P}_i(X_1 \neq j, \cdots, X_{n-1} \neq j, X_n = j) = \mathbb{P}(X_1 \neq j, \cdots, X_{n-1} \neq j, X_n = j | X_0 = i)$$

donde $\mathbb{P}_i(X_1 \neq j, \cdots, X_{n-1} \neq j, X_n = j) = \mathbb{P}(X_1 \neq j, \cdots, X_{n-1} \neq j, X_n = j | X_0 = i)$. En particular, si $i = j$, f_{ii}^n es la probabilidad del primer retorno a i en n pasos.

Sea $A_n = \{X_1 \neq j, \cdots, X_{n-1} \neq j, X_n = j\}$. Como $A_n \subset \{X_n = j\}$, tenemos que

$$f_{ij}^n = \mathbb{P}_i(A_n) \leq \mathbb{P}_i(X_n = j) = p_{ij}^n$$

Por otro lado, como los sucesos $A_1, A_2, \cdots$, son incompatibles dos a dos, tenemos que

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^n$$

es la probabilidad de visitar j partiendo de i. En el caso en el que $i = j$, f_{ii} es la probabilidad de retornar a i.

Definición 8. *Decimos que un estado i es recurrente cuando $f_{ii} = 1$. De lo contrario, si $f_{ii} < 1$ decimos que i es un estado transitorio.*

Consideremos para cada estado j la variable aleatoria

$$\tau_j = \text{ínf}\{n \geq 1 : X_n = j\} \tag{A.9}$$

Si el conjunto en A.9 es vacío, ponemos $\tau_j = \infty$.

La variable aleatoria τ_j es el *tiempo del primer pasaje por j* cuando la distribución inicial es arbitraria, y es el *tiempo del primer retorno a j* cuando la cadena de Markov parte del estado j.

Observar que $A_n = \{X_1 \neq j, \cdots, X_{n-1} \neq j, X_n = j\} = \{\tau_j = n\}$, y por tanto

$$f_{ij}^n = \mathbb{P}_i(\tau_j = n)$$

$$f_{ij} = \mathbb{P}_i(\tau_j < \infty)$$

En conclusión:

- Un estado i es recurrente cuando $f_{ii} = \mathbf{P}_i(\tau_i < \infty) = 1$
- Un estado i es transitorio cuando $f_{ii} = \mathbf{P}_i(\tau_i < \infty) < 1$

Proposición 21. *Sean i, j estados. Entonces vale la siguiente fórmula*

$$p_{ij}^n = \sum_{m=1}^{n} f_{ij}^m p_{jj}^{n-m} \tag{A.10}$$

Demostración. Aplicando la fórmula de la probabilidad total tenemos

$$\begin{aligned} p_{ij}^n = \mathbb{P}_i(X_n = j) &= \mathbb{P}_i\left(X_n = j, \bigcup_{m=1}^{n} \{\tau_j = m\}\right) \\ &= \sum_{m=1}^{n} \mathbb{P}_i(X_n = j, \tau_j = m) \\ &= \sum_{m=1}^{n} \mathbb{P}_i(X_n = j | \tau_j = m)\mathbb{P}_i(\tau_j = m) \\ &= \sum_{m=1}^{n} \mathbb{P}_i(X_n = j | X_m = j)\mathbb{P}_i(\tau_j = m) \\ &= \sum_{m=1}^{n} p_{jj}^{n-m} f_{ij}^m \end{aligned}$$

□

Usaremos la fórmula A.10 para demostrar la siguiente proposición.

Proposición 22. *Consideremos una cadena de Markov con espacio de estados $\mathbf{I}$, matriz de distribución $\mathcal{P}$ y distribución inicial π.*

(a) Un estado i es recurrente si y solo si se verifica $\sum_{n=1}^{\infty} p_{ii}^n = \infty$.

(b) Si j es transitorio, entonces $\sum_{n=1}^{\infty} p_{ij}^n < \infty$, lo que implica que $\lim_{n\to\infty} p_{ij}^n = 0$

Demostración. Supongamos que $a_{ij} = \sum_{n=1}^{\infty} p_{ij}^n < \infty$.

Aplicando la fórmula A.10 y reordenando tenemos

$$\begin{aligned} a_{ij} = \sum_{n=1}^{\infty} p_{ij}^n = \sum_{n=1}^{\infty}\left(\sum_{m=1}^{n} f_{ij}^m p_{jj}^{n-m}\right) = \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} f_{ij}^m p_{jj}^{n-m} &= \sum_{m=1}^{\infty} f_{ij}^m \sum_{n=0}^{\infty} p_{jj}^n \\ &= f_{ij}(1 + a_{jj}) \end{aligned} \tag{A.11}$$

donde en la última igualdad utilizamos que $p_{jj}^0 = 1$.

Ahora demostremos (a). Supongamos que $a_{ii} = \sum_{n=1} p_{ii}^n < \infty$. Aplicando la fórmula A.11 con $i = j$ tenemos que

$$f_{ii} = \frac{a_{ii}}{1 + a_{ii}} < 1$$

entonces i es transitorio.

Por otro lado, supongamos que $a_{ii} = \sum_{n=1} p_{ii}^n = \infty$, veamos que i es recurrente.

Para cada natural N tenemos

$$\sum_{n=1}^{N} p_{ii}^n = \sum_{n=1}^{N} \left(\sum_{m=1}^{n} f_{ii}^m p_{ii}^{n-m} \right) = \sum_{m=1}^{N} \sum_{n=m}^{N} f_{ii}^m p_{ii}^{n-m} \leq \sum_{m=1}^{N} f_{ii}^m \sum_{n=0}^{N} p_{ii}^n$$

De aquí obtenemos la acotación

$$f_{ii} \geq \sum_{m=1}^{N} f_{ii}^m \geq \frac{\sum_{n=1}^{N} p_{ii}^n}{\sum_{n=0}^{N} p_{ii}^n} = \frac{\sum_{n=1}^{N} p_{ii}^n}{1 + \sum_{n=1}^{N} p_{ii}^n} \to 1 \quad N \to \infty$$

Por tanto, $f_{ii} = 1$, y el estado i es recurrente.

Ahora demostremos (b).

Como j es transitivo, por (a) tenemos que $a_{jj} = \sum_{n=1}^{\infty} p_{jj}^n < \infty$. Aplicando la fórmula A.11 tenemos que

$$a_{ij} = f_{ij}(1 + a_{jj}) \leq 1 + a_{jj} < \infty$$

□

Cadenas de Markov absorbentes.

Vimos que el conjunto de los estados esenciales de una cadena de Markov se descompone en una unión disjunta de clases irreducibles. Si una clase irreducible está formada por un único estado, decimos que el estado es *absorbente*.

Definición 9. *Una cadena de Markov se dice absorbente si posee al menos un estado absorbente, y si desde cada estado es posible llegar a uno absorbente.*

Consideremos una cadena de Markov $X_0, X_1, \cdots$, con espacio de estados $\mathbf{I}$ finito, matriz de transición $\mathcal{P}$, y distribución inicial π. Supongamos que la cadena es absorbente. En este caso, todos los estados no absorbentes serán transitorios: no puede haber estados recurrentes no absorbentes, pues la probabilidad de pasar de un estado no absorbente a uno absorbente es positiva, pero la probabilidad de salir del estado absorbente es nula.

Podemos reenumerar los estados y colocar los transitorios primero. Si hay r estados absorbentes y t transitorios, la matriz de transición tendrá la siguiente *forma canónica*:

$$\mathcal{P} = \left(\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \end{array} \right) \tag{A.12}$$

donde $\mathbf{I}$ es la matriz identidad de dimensión $r \times r$, $\mathbf{0}$ es una matriz de ceros de dimensión $r \times t$, $\mathbf{R}$ es una matriz de dimensión $t \times r$ y $\mathbf{Q}$ es una matriz de dimensión $t \times t$.

Multiplicando por bloques tenemos:

$$\mathcal{P}^n = \left(\begin{array}{c|c} \mathbf{Q}^n & \tilde{\mathbf{R}} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right) \tag{A.13}$$

donde $\tilde{\mathbf{R}}$ depende de $\mathbf{Q}$ y $\mathbf{R}$.

Teorema 12. *En una cadena de Markov absorbente con espacio de estados finito, la probabilidad de absorción del proceso es 1. Es decir,* $\lim_{n\to\infty} \mathbf{Q}^n = 0$.

Demostración. La entrada ij de la matriz $\mathbf{Q}^n$ es p_{ij}^n: la probabilidad de estar un estado transitorio j luego de n pasos, empezando en i.

En la proposición 12 probamos que si j es transitorio entonces $\lim_{n\to\infty} p_{ij}^n = 0$. Por tanto, $\lim_{n\to\infty} \mathbf{Q}^n = 0$. □

Apéndice B

Martingalas

Una *martingala* es un proceso cuyo valor medio se mantiene constante en un sentido que definiremos a continuación. Para hacerlo, primero definiremos el concepto de *esperanza condicional.*

Definición 10. *Consideremos un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, una variable aleatoria X y otra variable aleatoria Y con esperanza $\mathbb{E}(Y)$. La esperanza condicional de Y dada X, que designamos $\mathbb{E}(Y|X)$ es una variable aleatoria $g(X)$ donde la función $g(X)$ verifica la propiedad*

$$\mathbb{E}(\mathbb{1}_{\{X\in I\}}Y) = \mathbb{E}(\mathbb{1}_{\{X\in I\}}g(X)) \tag{B.1}$$

para todo intervalo $I = (a, b]$ de la recta real.

La propiedad B.1 exige que las esperanzas de las variables aleatorias Y y $g(X)$ coincidan en los sucesos generados por X, es decir, en los sucesos de la forma $\{\omega : X(\omega) \in (a, b]\}$. En este sentido, $\mathbb{E}(Y|X)$ es la función de X que mejor aproxima a Y.

En estadística matemática se dice que $\mathbb{E}(Y|X)$ es un estimador de la variable aleatoria Y, cuando observamos la variable aleatoria X.

Veamos algunas propiedades de la esperanza condicional.

Propiedad 1. *(Linealidad). Consideremos dos variables aleatorias Y_1, Y_2 con esperanzas respectivas $\mathbb{E}(Y_1), \mathbb{E}(Y_2)$, una variable aleatoria X y dos constantes a, b. Entonces*

$$\mathbb{E}(aY_1 + bY_2|X) = a\mathbb{E}(Y_1|X) + b\mathbb{E}(Y_2|X) \tag{B.2}$$

Demostración. Verifiquemos que la función $h(x) = ag_1(x)+bg_2(x)$ cumple la ecuación B.1, donde $\mathbb{E}(Y_k|X) = g_k(X)$, $k = 1, 2$. En efecto, dado $I = (a, b]$, tenemos

$$\begin{aligned}\mathbb{E}(\mathbb{1}_{\{X\in I\}}(aY_1 + bY_2)) &= a\mathbb{E}(\mathbb{1}_{\{X\in I\}}Y_1) + b\mathbb{E}(\mathbb{1}_{\{X\in I\}}Y_2)\\ &= a\mathbb{E}(\mathbb{1}_{\{X\in I\}}g_1(X)) + b\mathbb{E}(\mathbb{1}_{\{X\in I\}}g_2(X))\\ &= \mathbb{E}(\mathbb{1}_{\{X\in I\}}h(X))\end{aligned}$$

Entonces

$$\mathbb{E}(aY_1 + bY_2|X) = h(X) = ag_1(X) + bg_2(X) = a\mathbb{E}(Y_1|X) + b\mathbb{E}(Y_2|X)$$

lo que concluye la demostración. □

Propiedad 2. *(Monotonía). Consideremos una variable aleatoria X y otra variable aleatoria Y con esperanza $\mathbb{E}(Y)$. Si $Y \geq 0$ entonces $\mathbb{E}(Y|X) \geq 0$.*

Demostración. Si el vector aleatorio (X, Y) tiene distribución discreta, y toma los valores (x_k, y_j) $(k, j = 1, 2, \cdots)$, tenemos $y_j \geq 0$, por lo que $g(x) \geq 0$ y $\mathbb{E}(Y|X) = g(X) \geq 0$ (ver ecuación 9.4 de [26]). Si el vector aleatorio (X, Y) tiene distribución absolutamente continua con densidad $p(x, y)$, la condición $Y \geq 0$ implica que $p(x, y) = 0$, si $y \leq 0$. Entonces $g(x) \geq 0$ (ver 9.5 de [26]) y $\mathbb{E}(Y|X) = g(X) \geq 0$.

En el caso general, en la demostración del teorema 9.1 de [26] se define una medida μ que es positiva, y también lo es la derivada de Radon-Nikodym $g(x)$; concluyendo que $\mathbb{E}(Y|X) = g(X) \geq 0$ c.s. □

Propiedad 3. *Consideremos una variable aleatoria X, y otra variable aleatoria Y con esperanza $\mathbb{E}Y$. Entonces*

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}Y \tag{B.3}$$

La igualdad B.3 se denomina la fórmula de la esperanza total.

Demostración. Sea $g(x)$ tal que $\mathbb{E}(Y|X) = g(X)$, y consideremos $I = (-\infty, \infty)$ en la propiedad B.1. Como $\{\omega : X(\omega) \in I\} = \Omega$, tenemos $\mathbb{1}_{\{X \in I\}} = 1$, y

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{1}_{\{X \in I\}} Y) = \mathbb{E}(\mathbb{1}_{\{X \in I\}} g(X)) = \mathbb{E}(\mathbb{E}(Y|X))$$

lo que demuestra la propiedad. □

Propiedad 4. *Consideremos dos variables aleatorias X, Y y una función $h(x)$. Supongamos que existen las esperanzas $\mathbb{E}Y, \mathbb{E}(h(x)Y)$. Entonces*

$$\mathbb{E}(h(X)Y|X) = h(X)\mathbb{E}(Y|X) \tag{B.4}$$

Demostración. Consideremos primero un vector aleatorio (X, Y) con distribución discreta que toma los valores (x_k, y_j), $k, j = 1, 2, \cdots$, la función $g(x)$ en 9.4 de [26] y veamos que la función $h(x)g(x)$ verifica B.1. En efecto, si $I = (a, b]$, tenemos

$$\begin{aligned}
\mathbb{E}(\mathbb{1}_{\{X \in I\}} h(X)Y) &= \sum_{k: a < x_k \leq b}^{\infty} h(x_k) \sum_{j=1}^{\infty} y_j \mathbb{P}(Y = y_j | X = x_k) \mathbb{P}(X = x_k) \\
&= \sum_{k: a < x_k \leq b}^{\infty} h(x_k) g(x_k) \mathbb{P}(X = x_k) \\
&= \mathbb{E}(\mathbb{1}_{\{X \in I\}} h(X) g(X))
\end{aligned}$$

Entonces $\mathbb{E}(h(X)Y|X) = h(X)g(X) = h(X)\mathbb{E}(Y|X)$.

Si el vector (X, Y) tiene densidad $p(x, y)$, la función $h(x)g(x)$ verifica B.1 con $g(x)$ definida en 9.5 de [26]. En efecto, si $I = (a, b]$ tenemos

$$\begin{aligned}
\mathbb{E}(\mathbb{1}_{\{X \in I\}} h(X)Y) &= \int_a^b \left(\int_{-\infty}^{\infty} h(x) y r(y|x) dy \right) p_1(x) dx \\
&= \int_a^b h(x) g(x) p_1(x) dx \\
&= \mathbb{E}(\mathbb{1}_{X \in I} h(X) g(X))
\end{aligned}$$

En el caso general, sea $\mathbb{E}(Y|X) = g(X)$. Como la propiedad B.1 se verifica para todo conjunto boreliano B de la recta real, la ecuación B.4 se verifica si $h(x) = \sum_{k=1}^{K} c_k \mathbb{1}_{\{x \in B_k\}}$ es una función simple, donde $c_1, \cdots, c_K$ son reales arbitrarios y $B_1, \cdots, B_K$ son conjuntos borelianos arbitrarios. La demostración concluye considerando una sucesión de funciones simples $\{h^n(x)\}$ que verifica $|h^n(x)| \leq |h(x)|$ y $h^n(X) \to h(X)$ $(n \to \infty)$ c.s., y aplicando el teorema de convergencia dominada. □

Propiedad 5. *(Telescópica). Consideremos los vectores aleatorios $F_n = (X_1, \cdots, X_n)$, $F_{n+m} = (X_1, \cdots, X_n, \cdots, X_{n+m})$ y una variable aleatoria Y con esperanza $\mathbb{E}(Y)$. Entonces se verifican las siguientes igualdades*

(a) $\mathbb{E}(\mathbb{E}(Y|F_n)|F_{n+m}) = \mathbb{E}(Y|F_n)$

(b) $\mathbb{E}(\mathbb{E}(Y|F_{n+m})|F_n) = \mathbb{E}(Y|F_n)$

Demostración. (a) Sea $g(x)$ $(x \in \mathbb{R}^n)$ tal que $\mathbb{E}(Y|F_n) = g(F_n)$. Considerando $g(F_n)$ función del vector aleatorio F_{n+m} y aplicando la fórmula B.4 tenemos

$$\mathbb{E}(\mathbb{E}(Y|F_n)|F_{n+m}) = \mathbb{E}(g(F_n|F_{n+m})) = g(F_n) = \mathbb{E}(Y|F_n)$$

lo que concluye la demostración de (a).

(b) Consideremos la función $h(x)$ $(x \in \mathbb{R}^{n+m})$ tal que $\mathbb{E}(Y|F_{n+m]}) = h(F_{n+m})$, y sea $k(z)$ $(z \in \mathbb{R}^n)$ tal que $\mathbb{E}(h(F_{n+m})|F_n) = k(F_n)$. Veamos que la función $k(z)$ verifica la ecuación 9.7 de [26]. En efecto, dado $I = (a_1, b_1] \times (a_n, b_n]$, tenemos $\{F_n \in I\} = \{F_{n+m} \in I \times \mathbb{R}^m\}$, y entonces

$$\begin{aligned}
\mathbb{E}(\mathbb{1}_{\{X_n \in I\}} k(F_n)) = \mathbb{E}(\mathbb{1}_{\{X_n \in I\}} \mathbb{E}(h(F_{n+m})|F_n)) &= \mathbb{E}(\mathbb{E}(\mathbb{1}_{\{X_n \in I\}} h(F_{n+m})|F_n)) \\
&= \mathbb{E}(\mathbb{E}(\mathbb{1}_{\{X_{n+m} \in I \times \mathbb{R}^m\}} h(F_{n+m})|F_n)) \\
&= \mathbb{E}(\mathbb{1}_{\{X_{n+m} \in I \times \mathbb{R}^m\}} h(F_{n+m})) \\
&= \mathbb{E}(\mathbb{1}_{\{X_{n+m} \in I \times \mathbb{R}^m\}} Y) \\
&= \mathbb{E}(\mathbb{1}_{\{X_n \in I\}} Y)
\end{aligned}$$

□

En adelante supondremos que $X_0, X_1, \cdots$ es una sucesión de variables aleatorias, cuyas propiedades especificaremos en cada situación, y consideraremos la sucesión de vectores aleatorios $F_n = (X_0, \cdots, X_n)$ $(n = 0, 1, \cdots)$.

Se dice que una sucesión $Y_0, Y_1, \cdots$ es *adaptada* a $F_0, F_1, \cdots$ cuando cada variable aleatoria Y_n es función del vector aleatorio F_n para cada $n = 0, 1, \cdots$. Es decir, para cada $n = 0, 1, \cdots$ existe una función $f_n(x)$ $(x \in \mathbb{R}^{n+1})$ tal que $Y_n = f_n(F_n)$.

Decimos que $Y_0, \cdots, Y_N$ son *adaptadas* a $\{F_0, \cdots, F_N\}$ cuando cada Y_n es función de F_n $(n = 0, \cdots, N)$.

Definición 11. *Decimos que una sucesión de variables aleatorias $Y_0, Y_1, \cdots$ con esperanzas $\mathbb{E}(Y_0), \mathbb{E}(Y_1), \cdots$ respectivamente y adaptada a $\{F_n\}$ es una martingala cuando se verifica*

$$\mathbb{E}(Y_{n+1}|F_n) = Y_n \tag{B.5}$$

para $n = 0, 1, \cdots$. La condición B.5 se llama propiedad de martingala.

Decimos que $\{Y_n\}$ es una submartingala si vale $\mathbb{E}(Y_{n+1}|F_n) \geq Y_n$, y que es una supermartingala si vale $\mathbb{E}(Y_{n+1}|F_n) \leq Y_n$.

De esta definición podemos deducir las siguientes propiedades.

Proposición 23. *Sea $\{Y_n\}$ un martingala respecto a $\{F_n\}$.*

(a) $\mathbb{E}(Y_n|F_m) = Y_m$ *para todo entero no negativo* $m \leq n$

(b) $\mathbb{E}(Y_n) = Y_0$

Demostración. (a) Dados $n \leq m$ tenemos $n = m + k$. Luego, usando la propiedad teléscópica tenemos y la fórmula de la esperanza total

$$\begin{aligned}
\mathbb{E}(Y_n|F_m) &= \mathbb{E}(\mathbb{E}(Y_n|F_{n-1})|F_m) = \mathbb{E}(Y_{n-1}|F_m) \\
&= \mathbb{E}(\mathbb{E}(Y_{n-1}|F_{n-2})|F_m) = \mathbb{E}(Y_{n-2}|F_m) \\
&\quad \vdots \\
&= \mathbb{E}(\mathbb{E}(Y_{n-(k-2)}|F_{n-2})|F_m) = \mathbb{E}(Y_{n-(k-1)}|F_m) \\
&= \mathbb{E}(Y_{m+1}|F_m) = Y_m
\end{aligned}$$

(b) Aplicando (a) a $m = 0$ obtenemos $\mathbb{E}(Y_n|F_0) = Y_0$. Tomando esperanza de ambos lados

$$\mathbb{E}(Y_n) = \mathbb{E}[\mathbb{E}(Y_n|F_0)] = \mathbb{E}(Y_0)$$

□

Apéndice C

Análisis de componentes principales

El análisis de componentes principales (PCA) es una técnica de análisis multivariado introducida por Karl Pearson en 1901, y desarrollado de manera independiente por Harold Hotelling en 1933. La idea central detrás de este método es reducir la dimensionalidad de un conjunto de datos en el que hay un número alto de variables correlacionadas, conservando al mismo tiempo la mayor cantidad de variación posible presente en el conjunto. Esta reducción se consigue transformando las variables originales en un nuevo conjunto de variables, las componentes principales, no correlacionadas, y ordenadas de tal forma que las primeras retienen la mayor parte de la variación presente en las variables originales.

El cálculo de las componentes principales se reduce a la solución de un problema de valores y vectores propios para una matriz simétrica semidefinida positiva.

Sea $\{\mathbf{x_1}, \cdots, \mathbf{x_n}\}$ n vectores en $\mathbb{R}^d$, queremos hallar el subespacio afín $\mathcal{S} \subset \mathbb{R}^d$ de dimensión $p < d$ que mejor aproxima el conjunto, en un sentido que definiremos más adelante.

Consideremos que cada $\mathbf{x_j}$ puede ser escrito como

$$\mathbf{x_j} = \mu + \mathbf{U}\mathbf{y_j} \quad \text{con } j = 1, \cdots, n \tag{C.1}$$

donde $\mu \in \mathcal{S}$, $\mathbf{U} \in \mathbb{R}^{d \times p}$ es una matriz cuyas columnas forman una base de $\mathcal{S}$, e $\mathbf{y_j} \in \mathbb{R}^p$ es el vector de coordenadas de $\mathbf{x_j}$ en $\mathcal{S}$. Es decir, estamos suoiniendo que la aproximación es exacta.

Ahora bien, la representación de los puntos $\mathbf{x_j}$ no es única: si consideramos $\mathbf{y_0} \in \mathbb{R}^d$ arbitrario, entonces

$$\mathbf{x_j} = (\mu + \mathbf{U}\mathbf{y_0}) + \mathbf{U}(\mathbf{y_j} - \mathbf{y_0}) \tag{C.2}$$

Es decir, podemos encontrar infinitos espacios afines que cumplan (C.1). Una forma de resolverlo es realizar una restricción en el conjunto $\{\mathbf{y_1}, \cdots, \mathbf{y_n}\}$

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{y_j} = 0$$

Esta elección tendrá sentido al momento de resolver el problema de optimización.

Por otro lado, para cualquier matriz $\mathbf{A} \in \mathbb{R}^{p\times p}$ invertible tenemos

$$\mathbf{x_j} = \mu + \mathbf{U}\mathbf{A}\mathbf{A}^{-1}\mathbf{y_j} = \mu + \tilde{\mathbf{U}}\tilde{\mathbf{y}}_\mathbf{j} \tag{C.3}$$

con $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{A}$ y $\tilde{y_j} = \mathbf{A}\mathbf{y_j}$. Es decir, la matriz $\mathbf{U}$ y los vectores $\mathbf{y_j}$ no son únicos. Una forma de resolver esto es requerir que la matriz $\mathbf{U}$ sea ortogonal, es decir, $\mathbf{U}\mathbf{U}^T = \mathbf{I}_p$. Esto resuelve el problema de ambigüedad a menos de isometrías: si $\mathbf{U}$ es como en la ecuación (C.1) y $\mathbf{R} \in \mathbb{R}^{p\times p}$ es una matriz ortogonal, entonces $\mathbf{x_j} = \mu + \mathbf{U}\mathbf{R}\mathbf{R}^{-1}\mathbf{y_j}$. Luego $\mathbf{U}\mathbf{R}$ es también solución.

El modelo anterior asume que el conjunto $\{\mathbf{x_1}, \cdots, \mathbf{x_n}\}$ pertenece al espacio afín de forma perfecta, pero esto en general no es así. Para modelarlo, asumamos que cada punto $\mathbf{x_j}$ pertenece al subespacio $\mathcal{S}$ pero está perturbado por un error aditivo $\epsilon_j \in \mathbb{R}$

$$\mathbf{x_j} = \mu + \mathbf{U}\mathbf{y_j} + \epsilon_j \tag{C.4}$$

La mejor aproximación será aquella que minimice el error. Es decir, la que minimice

$$\sum_{j=1}^{n} \|\epsilon_j\|_2^2 = \sum_{j=1}^{n} \|\mathbf{x_j} - \mu - \mathbf{U}\mathbf{y_j}\|_2^2 \tag{C.5}$$

Por tanto, podemos plantear el problema de optimización de encontrar el subespacio afín óptimo como

$$\begin{aligned} \text{minimizar}_{\ \mu,\mathbf{U},\{\mathbf{y_j}\}_{j\in\{1,\cdots,n\}}} \quad & \sum_{j=1}^{n} \|\mathbf{x_j} - \mu - \mathbf{U}\mathbf{y_j}\|_2^2 \\ \text{sujeto a} \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I}_p \\ & \sum_{j=1}^{n} \mathbf{y_j} = 0 \end{aligned} \tag{C.6}$$

Las columnas de $\mathbf{U}$ son las llamadas *componentes principales*. Al resolver el problema (C.6) estamos realizando un *análisis de componentes principales* (PCA) sobre la matriz $\mathbf{X} = [\mathbf{x_1}, \cdots, \mathbf{x_n}] \in \mathbb{R}^{d\times n}$.

La solución del problema está estrechamente relacionada con la descomposición en valores singulares (SVD) de una matriz.

C.1. PCA vía SVD

Para resolver el problema de optimización usaremos que si $\mathbf{A} \in \mathbb{R}^{n\times d}$ y $f : \mathbb{R}^n \to \mathbb{R}^d$ dada por $f(x) = \mathbf{A}x$, entonces

$$\frac{\partial f(x)}{\partial x} = \mathbf{A} \tag{C.7}$$

y que la derivada de la norma 2 de un vector $f(x)$ es

$$\frac{\partial}{\partial x}\|f(x)\|_2^2 = f(x)^T \frac{\partial f(x)}{\partial x} \tag{C.8}$$

Para resolver (C.6), planteamos la función lagrangiana del problema

$$\begin{aligned}\mathcal{L} &= \mathcal{L}(\mu, \mathbf{U}, \mathbf{y_1}, \cdots, \mathbf{y_n}, \gamma, \Delta) \\ &= \sum_{j=1}^{n} \|\mathbf{x_j} - \mu - \mathbf{U}\mathbf{y_j}\|_2^2 + \gamma^T \sum_{j=1}^{n} \mathbf{y_j} + tr((\mathbf{I}_d - \mathbf{U}^T\mathbf{U})\Delta)\end{aligned} \tag{C.9}$$

donde $\gamma \in \mathbb{R}^p$ y $\Delta \in \mathbb{R}^{p\times p}$.

Usando (C.8) derivamos $\mathcal{L}$ respecto a μ e igualamos a 0

$$\frac{\partial}{\partial \mu}\mathcal{L} = -2\sum_{j=1}^{n}(\mathbf{x_i} - \mu - \mathbf{U}\mathbf{y_j})^T = \mathbf{0}_p^T$$

Luego

$$\left(\sum_{j=1}^{n} \mathbf{x_j}\right) - n\mu - \mathbf{U}\left(\sum_{j=1}^{n} \mathbf{y_j}\right) = \mathbf{0}_p$$

y como $\sum_{j=1}^{n} \mathbf{y_j} = 0$ tenemos

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_j} \tag{C.10}$$

Por otro lado, si derivamos la función respecto a $\mathbf{y_j}$ para cada $j = 1, \cdots, n$, e igualamos a 0, usando (C.7) y (C.8) tenemos

$$2(\mathbf{x_j} - \mu - \mathbf{U}\mathbf{y_j})^T(-\mathbf{U}) + \gamma^T = \mathbf{0}_p^T \tag{C.11}$$

Sumando las derivadas

$$\underbrace{-2\sum_{j=1}^{n} \mathbf{x_j}\mathbf{U}^T + 2\sum_{j=1}^{n} \mu^T\mathbf{U}}_{S_1} + \underbrace{2\sum_{j=1}^{n} \mathbf{y_j}^T\mathbf{U}^T\mathbf{U}}_{S_2} + \gamma^T = S_1 + S_2 + \gamma^T$$

Luego, utilizando (C.10) tenemos

$$S_1 = -2n\mu^T\mathbf{U} + 2n\mu^T\mathbf{U} = \mathbf{0}_p^T$$

Por otro lado, como $\mathbf{U}$ es ortogonal y $\sum_{j=1}^{n} \mathbf{y_j} = 0$ tenemos

$$S_2 = 2\sum_{j=1}^{n} \mathbf{y_j}^T = \mathbf{0}_p^T$$

Es decir, $S_1 = S_2 = 0_p^T$, y entonces el vector de multiplicadores de Lagrange es $\gamma = \mathbf{0}_p$

Sustituyendo lo anterior en (C.11) obtenemos

$$\mathbf{y_j} = \mathbf{U}^T(\mathbf{x_j} - \mu) \tag{C.12}$$

Remplazando los valores hallados para $\mathbf{y_j}$ en (C.6)

$$\begin{aligned}&\text{minimizar}_{\ \mu,\mathbf{U},\{\mathbf{y_j}\}_{j\in\{1,\cdots,n\}}} && \sum_{j=1}^{n} \|\mathbf{x_j} - \mu - \mathbf{U}\mathbf{U}^T(\mathbf{x_j} - \mu)\|_2^2 \\ &\text{sujeto a} && \mathbf{U}^T\mathbf{U} = \mathbf{I}_p\end{aligned} \tag{C.13}$$

Resta entonces encontrar la matriz $\mathbf{U}$ óptima del problema. Para esto utilizaremos la descomposición en valores singulares de la matriz $\tilde{\mathbf{X}} = \mathbf{X} - \mu$, donde $\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_j}$.

Definición 12. *(Valor singular de una matriz). Dada $A \in \mathbf{R}^{n\times d}$, decimos que σ es un valor singular de $\mathbf{A}$ si σ^2 es un valor propio de $\mathbf{A}\mathbf{A}^T$ (o equivalentemente, de $\mathbf{A}^T\mathbf{A}$).*

Teorema 13. *(Descomposición en valores singulares). Dada $\mathbf{A}$ una matriz de rango r, existen $\mathbf{U} \in \mathbb{R}^{n\times n}$ y $\mathbf{V} \in \mathbb{R}^{d\times d}$ matrices ortogonales, y $\mathbf{\Sigma} \in \mathbb{R}^{n\times d}$ una matriz formada por los r valores singulares de $\mathbf{A}$ en su diagonal principal, tales que $\mathbf{A}$ puede ser escrita como*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{C.14}$$

Una demostración del teorema anterior se puede encontrar en [12].

Teorema 14. *Sea $\tilde{\mathbf{X}} \in \mathbb{R}^{d\times n}$ la matriz formada por los datos centrados puestos como vectores columna. Considermos $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ la SVD de X tal que*

$$\Sigma = diag(\sigma_1, \cdots, \sigma_r) \quad con\ \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

Luego para $p < d$, la solución óptima de $\mathbf{U}$ para el problema (C.13) está dada por las primeras p columnas de $\mathbf{U}$, la solución óptima para $\mathbf{y_j}$ está dada por la j-ésima colmuna de la submatriz $\mathbf{\Sigma}\mathbf{V}^T$ y el valor óptimo de la función objetivo del problema (C.13) está dado por $\sum_{j=p+1}^{d} \sigma_j^2$, donde σ_j es el j-ésimo valor singular de $\mathbf{X}$.

Una demostración del teorema anterior se puede encontrar en [17].

Bibliografía

[1] Buneman, P., 1971. *The Recovery of Trees from Measures of Dissimilarity.* In Mathematics the the Archeological and Historical Sciences: Proceedings of the Anglo-Romanian Conference, Mamaia, 1970 (pp. 387-395). Edinburgh University Press.

[2] Buneman, P., 1974 *A note on the metric properties of trees.* J. Comb. Theory Ser. B 17: 48–50.

[3] Cavalli-Sforza, L. L. and A. W. F. Edwards, *Phylogenetic analysis: models and estimation procedures.* Evolution 21: 550–570, 1967.

[4] Cavalli-Sforza, L. L., and A. Piazza *Analysis of evolution: evolutionary rates, independence and treeness.* Theor. Popul. Biol. 8: 127–165, 1975.

[5] Crow, J.F. and Kimura, M. *An introduction in Population Genetics Theory.* Harper and Row, New York, 1970.

[6] Dobrow, R. P., *Introduction to Stochastic Processes with R*, John Wiley & Sons, Inc., 2016.

[7] Durrett, R. *Probability Models for DNA Sequence Evolution*, Segunda Edición, 2008.

[8] Futuyma, D. J. *Evolution.* Sinauer Associates, Inc., 2005.

[9] Gillespie, J. H. *Population Genetics - A Concise Guide*, Johns Hopkins University Press, Baltimore and London, 1998.

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

[11] Hein, J., Schierup, M. H. and Wiuf, C. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*, Oxford University Press, 2004.

[12] Jolliffe, I.T. *Principal component analysis.* New York, NY: Springer Science & Business Media, 2013.

[13] Lessa, E. P. *Guía de estudio de Genética de Poblaciones*, Laboratorio de Evolución de Facultad de Ciencias, Montevideo, Uruguay, 2004.

[14] Lewontin RC. 1972 *The apportionment of human diversity.* In Evolutionary biology (eds WC Steere, T Dobzhansky, MK Hecht), pp. 381–398. New York, NY: Springer.

[15] Lipson, M., P.-R. Loh, A. Levin, D. Reich, N. Patterson et al., 2013 *Efficient moment-based inference of admixture parameters and sources of gene flow.* Mol. Biol. Evol. 30: 1788–1802

[16] Lipson M., 2020 *Applying f4-statistics and admixture graphs: Theory and examples.* Molecular Ecology Resources. 20:1658–55 1667.

[17] Martínez, G. *Análisis de componentes principales para datos genómicos en presencia de datos faltantes.*, Tesis de Maestría en Ingeniería Matemática, Universidad de la República, 2021.

[18] Norris, J. Frontmatter. In *Markov Chains* (Cambridge Series in Statistical and Probabilistic Mathematics, pp. I-Vi). Cambridge: Cambridge University Press, 1997.

[19] Novembre J. 2022 *The background and legacy of Lewontin's apportionment of human genetic diversity.* Phil. Trans.R.Soc. B 377: 20200406. https://doi.org/10.1098/rstb.2020.0406

[20] Oteo-Garcia G, Oteo JA. *A geometrical framework for f-statistics.* Bull. Math. Biol. 83, 1–22, 2021. (doi:10.1007/s11538-020-00850-8)

[21] Patterson N, Price AL, Reich D. *Population structure and eigenanalysis.* PLoS Genet. 2, e190, 2006. (doi:10.1371/journal.pgen.0020190)

[22] Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. *Ancient admixture in human history.* Genetics 192, 1065–1093, 2012. (doi:10.1534/genetics.112.145037)

[23] Peter, Benjamin M, *Admixture, Population Structure, and F-Statistics*, Genetics, Volume 202, Issue 4, 1 April 2016, Pages 1485–1501, https://doi.org/10.1534/genetics.115.183913

[24] Peter Benjamin M. 2022 *A geometric relationship of F_2, F_3 and F_4-statistics with principal component analysis*, Philosophical Transactions of the Royal Society B: Biological Sciences, 377(1852): 20200413. https://doi.org/10.1098/rstb.2020.0413

[25] Pierce, B.A. *Genética. Un enfoque conceptual.* 2ª. Edición. Editorial Médica Panamericana, 2009.

[26] Petrov, V. V., Mordecki, E. *Teoría de la Probabilidad*, Segunda Edición, DIRAC - Facultad de Ciencias, 2008.

[27] Raghavan, M., P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen et al., 2014 *Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans.* Nature 505: 87–91.

[28] Reich, D., Thangaraj, K., Patterson, N. et al. *Reconstructing Indian population history.* Nature 461, 489–494 (2009). https://doi.org/10.1038/nature08365

[29] Reich, D., N. Patterson, D. Campbell, A. Tandon, S. Mazieres et al., 2012 *Reconstructing Native American population history.* Nature 488: 370–374.

[30] Slatkin, M., 1991 *Inbreeding coefficients and coalescence times.* Genet. Res. 58: 167–175.

[31] Strobeck, C., 1987 *Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision.* Genetics 117: 149–153.

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

[33] Tavaré, S., 1984 *Line-of-descent and genealogical processes, and their applications in population genetics models.* Theor. Popul. Biol. 26: 119–164.

[34] Tavaré, S. 2001 *Ancestral Inference in Population Genetics.* Springer Lecture Notes in Mathematics.

[35] The 1000 Genomes Project Consortium. *A global reference for human genetic variation.* Nature 526, 68–74 (2015). https://doi.org/10.1038/nature15393

[36] Wahlund, S. 1928 *Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus Betrachtet.* Hereditas 11: 65-106.

[37] Wakeley J. *Coalescent Theory: An Introduction*, 2009. Roberts & Co. Greenwood Village, CO.