FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Métodos matemáticos en el estudio del parentesco

María Vidal García

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Métodos matemáticos en el estudio del parentesco

María Vidal García

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estadística e Investigación Operativa
Título: Métodos matemáticos en el estudio del parentesco
Breve descrición do contido
El objetivo de este trabajo es estudiar los conceptos y métodos matemáticos que se usan para determinar el parentesco a partir de muestras de ADN.

Índice general

Resumen

El presente trabajo tiene como objetivo principal revisar los conceptos básicos relativos al cálculo de probabilidades en al área de la genética forense y su aplicación en problemas de búsqueda de parentesco. Para ello en primer lugar se introducirán algunas nociones básicas de biología y las herramientas matemáticas empleadas. A continuación se desgranará el proceso de traducción del problema genético a un lenguaje matemático y se implementarán los distintos objetos estadísticos. Dado que el enfoque del trabajo es predominantemente práctico, no solo se ilustrará el proceso anterior con numerosos ejemplos resueltos a mano, sino que se presentará Familias, un paquete de funciones implementadas en el software estadístico R que permite resolver los problemas expuestos. Se explicará su funcionamiento en profundidad y se aplicará para resolver problemas más complejos. Por último, se presentará el fundamento teórico subyacente al modelo para reforzar y complementar la comprensión del tema. La fuente fundamental de este trabajo es el libro *Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics* de Thore Egeland, Daniel Kling y Petter Mostad [7].

O presente traballo ten como obxectivo principal revisar os conceptos básicos relativos ó cálculo de probabilidades na área da xenética forense e a súa aplicación na procura de parentesco. Para iso, comezamos introducindo algunhas nocións básicas de bioloxía e as ferramentas matemáticas que empregaremos. Deseguido, explicarase polo miúdo o proceso de tradución do problema xenético á linguaxe matemática e implementaranse os distintos obxectos estatísticos. Como o enfoque do traballo é predominantemente práctico, non só se ilustrará o proceso anterior con numerosos exemplos resoltos a man, senón que se presentará Familias, un paquete de funcións implementadas no software estatístico R que permite resolver os problemas expostos. Explicarase o seu funcionamento en profundidade e aplicarase para resolver problemas máis complexos. Por último, presentarase o fundamento teórico sobre o que descansa o modelo para reforzar e complementar a comprensión do tema. A fonte fundamental deste traballo é o libro *Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics* de Thore Egeland, Daniel Kling e Petter Mostad [7].

Abstract

The main purpose of this work is to review basic concepts related to the assessment of probability in forensic genetics and their application to kinship problems. To this end, basic notions of biology and required mathematical skills will be presented in the first place. Then, we will study the process of expressing the genetic problems in mathematical language and previously presented statistical concepts will be implemented. Due to our mainly practical approach, every explanation will be accompanied by examples solved by hand and with Familias, a package belonging to the statistical software R. We will explain in detail its functioning and use it to solve more complicated problems. Lastly, theoretical foundation of the model will be presented to reinforce the comprehension. The main source of the content of this work is the book *Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics* from Thore Egeland, Daniel Kling and Petter Mostad [7].

Introducción

En los últimos tiempos, el estudio de la genética ha experimentado un gran crecimiento. Su gran potencial en los ámbitos de la medicina y la criminología hacen que de manera recurrente se oiga hablar de estudios de ADN en los medios de comunicación. Aunque prácticamente todo el mundo tiene cierta idea sobre este tema, no es tan común tener una visión rigurosa al menos a nivel conceptual sobre su funcionamiento, sus implicaciones y sus limitaciones. En este trabajo vamos a tratar de dar una visión general, básica pero rigurosa, de la relación entre las matemáticas y la genética.

Para un matemático, el ADN puede ser visto simplemente como un conjunto de información que sigue ciertas reglas en cuanto a su estructura y conservación. Con esta premisa tan sencilla, resulta inmediato que la estadística ha de jugar un papel crucial en el desarrollo de técnicas que permitan analizar e interpretar esta información.

Trabajar en este ámbito no es fácil, la naturaleza es tremendamente compleja y no conocemos en profundidad los mecanismos que intervienen a la hora de transmitir la información a la descendencia. Así, cualquier planteamiento o modelo ha de ser por fuerza simplificador. Donde en matemáticas el término *general* significa aplicable siempre, en biología (en particular en genética) quiere decir el caso más simple. El **objetivo** de este texto es precisamente explicar con detalle el razonamiento base, dando intuición sobre las ideas subyacentes e ilustrándolas con ejemplos.

La **estructura** del texto consta de 5 capítulos que se pueden dividir temáticamente en tres bloques. El primer bloque es introductorio y comprende los capítulos 1 y 2: en el primero se explican brevemente las nociones básicas de genética necesarias para poder seguir el desarrollo de la teoría. El segundo es una recopilación de herramientas básicas de probabilidad y estadística que se utilizarán a lo largo del documento.

La segunda parte es el núcleo del trabajo, se centra en exponer los principios fundamentales del cálculo de probabilidades en problemas de determinación de parentesco y sus aplicaciones. En el capítulo 3 establecemos la relación entre matemáticas y estadística, explicando como expresar los datos en un lenguaje apropiado que nos permita extraer conclusiones. Explicamos este proceso sobre el caso más sencillo de determinación de parentesco en genética forense, un problema conocido como Standard Duo Case. A continuación se presentan tres ejemplos que servirán para ilustrar algunas ampliaciones del modelo original a lo largo de la segunda parte de este capítulo. El capítulo 4 presenta la herramienta informática que vamos a utilizar para ilustrar los conceptos aprendidos. Se trata del paquete Familias para R, del que se detallarán sus funciones

y su implementación. Finalmente se aplicará tanto para comprobar algunos resultados obtenidos a mano en secciones previas como para resolver casos que requieren cálculos más extensos.

Por último, el último bloque presenta los fundamentos teóricos del modelo estadístico utilizado. Consta de un único capítulo, el capítulo 5. En él se profundiza en las raíces del contenido previo.

El **contenido** está basado principalmente en el libro *Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics* por Thore Egeland, Daniel Kling y Petter Mostad [7]. Se ha hecho una selección dentro de su extenso contenido y se ha completado con numerosas fuentes complementarias (ver Bibliografía para más información).

Pese a tener un **enfoque** eminentemente práctico, se ha intentado en todo momento dotar de interpretación a todos los conceptos usados, cuando no ha sido posible una introducción teórica rigurosa se ha tratado al menos de dar la idea intuitiva, y presentar ejemplos que faciliten la comprensión. Este texto pretende ser un punto de partida a un tema enormemente complejo como es el análisis de la información genética. Por razones de espacio algunas extensiones del modelo base como mutaciones, dependencia entre marcadores o marcadores no autosómicos[1] no se verán o se presentarán superficialmente.

Por último, cabe subrayar algunos **aspectos éticos** del lenguaje que emplearemos. La mayor parte del lenguaje del campo semántico de la genética fue establecido en el siglo pasado. Es común, especialmente al hacer referencia a información asociada a manifestaciones físicas típicas o a relaciones tradicionalmente establecidas, que gran parte del lenguaje se haya tomado directamente del lenguaje común. Con la evolución de la sociedad algunos términos pueden presentar hoy día ciertos conflictos éticos que es conveniente aclarar antes de comenzar la lectura.

En primer lugar uno de los datos básicos sobre los individuos involucrados en cualquier problema genético es su **sexo**. Aunque el término está lleno de connotaciones, en este contexto se utilizará en un sentido estricto como clasificación en base a la información genética de los cromosomas del par 23 (comúnmente denominado par sexual): una dotación XX clasifica al individuo como mujer y una dotación XY como varón (no se plantean problemas que involucren hermafroditismos ni entrarán en consideración las características físicas o la identidad de género de los individuos en cuestión).

Otra cuestión es la relativa a las **relaciones** que se establecen entre los individuos. La nomenclatura utilizada es la tradicional *padre*, *madre* y *hijo/a*. Se utilizará con el significado estricto de parentesco genético, es decir, la relación se ciñe a la transmisión/recepción de material genético. A veces se utilizan las expresiones *padre biológico* o *padre verdadero* para subrayar la hipótesis de existencia de parentesco genético. Estos términos en modo alguno tienen significación afectiva. Del mismo modo, para denotar la relación entre dos individuos que tienen descendencia se utilizan los términos *matrimonio* o se dice que son *cónyuges*, pero nuevamente se trata de nomenclatura heredada que pierde su dimensión social y legal en este contexto. Por último, en muchos problemas se detalla el sexo de algunos individuos aunque después no se utiliza (hay que recordar que no vamos a explorar el potencial de los marcadores no autosómicos, en un caso real

[1] Marcadores vinculados al sexo del individuo.

sí se tendrían en cuenta). Por eso en la mayor parte de los problemas el sexo de la descendencia no tiene importancia y se podría considerar indistintamente *niño* o *niña*, *hijo* o *hija*.

Por último, esperamos que al final de este trabajo, además del contenido específico que aparece detallado a lo largo de sus páginas, hayamos sido capaces de transmitir el papel fundamental que juegan y deben seguir jugando las matemáticas a la hora de afrontar los retos y nuevos horizontes en el tratamiento de la información genética.

Capítulo 1

Fundamentos de Biología

A la hora de estudiar las características y evolución de cualquier organismo vivo, quizás el aspecto más importante que debemos observar es su **genoma**, es decir, el conjunto de toda su información genética.

El parecido entre padres e hijos o la herencia de algunas enfermedades son dos ejemplos tradicionalmente conocidos y aceptados de que hay *algo*, algún agente biológico que transmite información de los progenitores a la descendencia. Este planteamiento, en apariencia tan sencillo, no aparece hasta finales del siglo XIX, como consecuencia del impulso que experimenta el razonamiento científico durante la Ilustración.

El primero en estudiar el fenómeno de la herencia de manera sistemática fue el fraile agustino Gregor Mendel. Su trabajo, publicado en 1866 [25], se limita a describir los patrones de herencia observados en la cosecha de guisantes. Tras él, en los años siguientes, se producen avances importantes en el ámbito de la bioquímica molecular que permiten finalmente identificar la estructura molecular encargada de contener y transmitir la información: el **ADN**[1].

El ADN se encuentra en general almacenado en el núcleo celular y en algunos orgánulos. La célula presenta además proteínas encargadas de tomar la información del núcleo y materializarla.

La estructura del ADN fue determinada por los biólogos James Watson y Francis Crick en 1953, gracias a las aportaciones de la cristalógrafa Rosalind Franklin [25]. Lograron determinar que el ADN es una doble hélice formada por dos cadenas de ácidos nucleicos. La clave de esta estructura, y lo que va a permitir codificar la información, es que cada ácido nucleico presenta una **base** nitrogenada de alguno de los siguientes tipos: adenina (A), guanina (G), citosina (C) y timina (T). Así distintas secuencias de bases nitrogenadas contienen información diferente.

Llamaremos **gen** a un segmento de ADN funcional, es decir, que codifica y se hereda en bloque. Normalmente un individuo posee toda su información genética duplicada, pues hereda la mitad del genoma de su padre y la otra mitad de su madre. En particular para cada gen tendrá dos versiones que pueden o no coincidir. Cada posible secuencia que pueda presentar un gen se llama **alelo**. Si un individuo presenta dos alelos iguales para un determinado gen se dice que es **homocigoto** para ese gen, de lo contrario se dice que es **heterocigoto**. Cuando nos referimos a

[1] Ácido desoxirribonucleico.

los alelos particulares que presenta una persona para un gen específico hablaremos de **genotipo** para ese gen. Al conjunto de manifestaciones perceptibles de la información genética se le llama **fenotipo** (por ejemplo el color de ojos o el grupo sanguíneo).

En los seres humanos, el material genético se condensa de forma natural formando unas estructuras llamadas **cromosomas**. Por norma general hay en total 46 cromosomas de 23 tipos: 22 pares de autosomas (cromosomas que no contienen información relativa al sexo del individuo) y un par sexual. Cada pareja de cromosomas se caracteriza por presentar los mismos genes, es decir, información para los mismos rasgos. A la localización de un gen en un cromosoma se le llama **locus**, a veces pueden identificarse ambos términos.

La herencia de la información de genes se produce a través de la copia. Durante la formación del óvulo y el espermatozoide, tiene lugar un proceso de mezcla y división de la información. Al final cada una de esas células posee una dotación genética de 23 cromosomas: un ejemplar de cada gen escogido de forma aleatoria de entre las dos opciones del individuo. Así, al juntarse ambas células obtenemos la dotación completa. Sin embargo, durante la mezcla, división y copia de la información pueden producirse errores que modifiquen la carga genética heredada. Los errores más comunes son las **mutaciones**, que son fallos a la hora de copiar la información. Algo tan simple como cambiar, añadir u omitir una base puede llegar a tener consecuencias fenotípicas.

Ejemplo 1.1. El gen HBB es uno de los más cortos. Cuenta con menos de 4000 bases. Se encuentra en el cromosoma 11 y contiene la información relativa a la síntesis de una proteína que forma parte de la hemoglobina, la $\beta - globina$ [26].

La presencia de mutaciones en este gen presenta un impacto variable. Mientras que siempre pueden existir mutaciones que pasen inadvertidas, al modificar alguna base en un gen tan pequeño difícilmente se dará ese caso. Existen diversas mutaciones de distinta extensión que se manifiestan en forma de beta-talasemia, una enfermedad en la que no se reparte correctamente el oxígeno a los distintos tejidos. En función de como sea la mutación la enfermedad puede ser más o menos grave. Otros tipos de mutaciones provocan lo que se conoce como anemia falciforme, una enfermedad que afecta a los glóbulos rojos pero que solo se manifiesta en individuos homocigotos.

Como consecuencia de todo lo anterior, quedó patente desde un principio que el análisis del ADN tenía mucho que aportar a las ciencias forenses (además de a muchos otros campos como la medicina o la biología). Esta disciplina experimenta un crecimiento muy importante a partir de 1984, año en el que el genetista británico Alec Jeffreys desarrolla la primera técnica de análisis de la huella genética, lo que hoy se conoce como prueba de ADN [9]. El potencial de este descubrimiento se puso de manifiesto de inmediato, siendo aplicado en una investigación criminal por primera vez en el caso de Colin Pitchfork (1986), y posibilitando la detención del verdadero responsable del delito. La disciplina de la genética forense ha experimentado grandes avances desde entonces [24].

Otra de las consecuencias del desarrollo de la genética ha sido la aparición y florecimiento de

la genética clínica y otros análisis genéticos. Las pruebas de determinación de parentesco (siendo la más conocida popularmente la prueba de paternidad) son problemas pertenecientes a estas áreas que se utilizarán como ejemplo principal en este trabajo.

En genética forense se estudiarán problemas que involucran a varios individuos. La herramienta que utilizaremos para resolverlos será el análisis de la información genética de algunos loci[2] concretos para cada sujeto involucrado, lo que se conoce como **perfil genético**, que serán obtenidos a partir de la secuenciación de muestras.

¿Y dónde entra la estadística en todo esto? Teniendo en cuenta como se codifica la información utilizando secuencias de cuatro bases (A,G,C,T), así como los patrones de herencia alélica, está claro que las técnicas de probabilidad y estadística pueden ayudarnos a determinar la plausibilidad de ciertas hipótesis. Así, partiendo de perfiles genéticos, se puede estimar el grado de confianza que depositamos en que un suceso (existencia de parentesco, dos muestras provienen de la misma fuente, etc) sea cierto.

Sin embargo, antes de plantear las técnicas y herramientas para modelizar la herencia genética, hay que analizar como es la información con la que contaremos.

En el genoma humano existen entre veinte mil y veinticinco mil genes, y la mayor parte del genoma está formado por secciones enteras de ADN no codificante. En general no es necesario analizar toda esta información, lo que se hace es seleccionar locus concretos que presenten determinadas características deseables y trabajar con ellos para extraer conclusiones. Estas localizaciones prefijadas y estudiadas es lo que se conoce como **marcadores genéticos** o directamente **locus**.

Veamos los criterios de selección utilizados para elegir marcadores genéticos:

- **Polimorfismo**: el polimorfismo hace referencia a la variabilidad que presenta esa localización concreta del genoma, es decir, a la cantidad de alelos que puede presentar. Cuanto mayor polimorfismo presente el marcador, más información nos aporta sobre el problema.

- **Tasa de mutación**: La existencia de mutaciones complica el proceso de determinar parentescos, por lo que idealmente intentaremos escoger marcadores cuya tasa de mutación no sea muy alta, aunque igualmente deberemos encontrar la forma de tener en cuenta esta posibilidad en nuestro modelo estadístico.

- **Rasgos fenotípicos**: tradicionalmente, por motivos tanto legales como éticos, se evita trabajar con marcadores que tengan manifestaciones en el fenotipo. En los últimos años sin embargo están empezando a desarrollarse técnicas para poder utilizar este tipo de marcadores.

- **Independencia**: se intenta recurrir a marcadores que no estén relacionados entre sí, es decir, que los alelos que presente cada uno de ellos sea independiente del que presenten los demás. De no ser así se dice que los alelos están relacionados y existen métodos específicos para ellos.

[2] Plural de *locus*.

- **Técnicas de determinación**: deben existir recursos tecnológicos para secuenciar el genoma de los locus escogidos con un coste tanto económico como temporal razonable. Esto es especialmente problemático cuando se trata con muestras de ADN degradadas o insuficientes.
- **Bases de datos**: el investigador ha de tener a su disposición bases de datos completas sobre la frecuencia en la población de interés de cada alelo de los marcadores utilizados.
- **Variación interpoblacional**: relacionado con el criterio anterior, en general se preferirán marcadores cuya frecuencia no varíe mucho de unas poblaciones a otras.

Encontrar marcadores que se ajusten a todas estas exigencias no es fácil, en particular, si un marcador presenta mucho polimorfismo suele ser porque su tasa de mutación es alta (por eso a lo largo de las generaciones han aparecido tantas variantes). Los principales tipos de marcadores con los que vamos a tratar son dos: los polimorfismos de un solo nucleótido **SNPs**[3] y las repeticiones de una única secuencia **STRs**[4].

Los SNPs son marcadores que constan de una sola base nitrogenada (aunque existen algunos que son secuencias muy cortas). Cada alelo sería simplemente una posible base. Se considera SNP siempre que se presente en al menos el 1 % de la población, de lo contrario se considera mutación puntual ([27]).

Ejemplo 1.2. Los marcadores $rs6311$ y $rs6313$ en el cromosoma 13 son dos localizaciones concretas que pueden presentar las variantes C (citosina) y T (timina) [27].

Los STRs son secuencias cortas de ADN que se repiten una cantidad variable de veces. La cantidad de copias que haya determina ante que alelo nos encontramos (pueden existir incluso variantes con repeticiones incompletas que también se consideran alelos distintos).

Ejemplo 1.3. El marcador $D3S1768$ es muy utilizado en pruebas de paternidad. El patrón que se repite es $AGAT$ y tiene 11 alelos diferentes [13].

Ambos tipos de marcadores se usan frecuentemente en genética forense. Ambos presentan tasas de mutación bastante altas, pero los SNPs sobresalen en este aspecto (casi el 90 % de la variabilidad genética humana proviene de mutaciones en una sola base [12]). La ventaja es que se cuenta con multitud de técnicas diferentes que permiten analizarlos masivamente [10]. Por su parte los STRs presentan mayor polimorfismo (normalmente entre 6 y 40 alelos) por lo que suelen ser preferibles.

[3]Del inglés *single nucleotide polymorphism* (polimorfismo de un solo nucleótico).
[4]Del inglés *sigle tandem repeats* (repeticiones de una única secuencia).

Capítulo 2

Fundamentos de Matemáticas

A lo largo de este capítulo vamos a presentar los resultados necesarios para poder desarrollar el contenido de las secciones posteriores. El objetivo es contar con herramientas suficientes que permitan dar un enfoque formal al estudio del parentesco La estructura y la selección del contenido a incluir se basan en [9] y en [7]. El contenido se ha construido a partir de [11] y [14].

2.1. Espacio de probabilidad

La teoría de la probabilidad se ocupa de estudiar los posibles resultados de un **experimento aleatorio**, es decir, un experimento del que no podemos predecir el resultado antes de realizarlo. Que no se pueda saber a ciencia cierta el resultado, no obstante, no significa que no se pueda obtener información sobre él a priori.

Denotaremos por **espacio muestral** Ω de un experimento aleatorio al conjunto de sus posibles resultados. El espacio muestral puede ser discreto (finito o infinito) o continuo, se modeliza de una u otra forma atendiendo a las características específicas de cada experimento.

Vamos a trabajar sobre cualquier subconjunto del espacio muestral $A \subset \Omega$, que denominaremos **suceso**. En genética normalmente los sucesos considerados serán los genotipos de los individuos involucrados en cada problema obtenidos a partir de las muestras. La noción de suceso lleva implícita cierta estructura, que es lo que hace que sea útil en el contexto del estudio de probabilidades. En particular, nos interesa que el conjunto de sucesos sea cerrado para operaciones usuales del álgebra de conjuntos (tales como la unión numerable, intersección numerable y complementaridad). Por ejemplo, si conozco la probabilidad de que un niño sea hijo biológico de una mujer concreta, y la probabilidad de que sea hijo de un hombre concreto, es interesante poder obtener la probabilidad de que lo sea de ambos, es decir, de la intersección de ambos sucesos, de lo contrario la utilidad de la función de probabilidad sería mínima.

El objetivo es de alguna forma *cuantificar* la tendencia que tiene el experimento a presentar uno u otro resultado. Para ello se busca asignar a cada suceso una magnitud, la *probabilidad* del suceso. Esto se consigue definiendo la función de probabilidad, que no es más que una medida definida sobre el conjunto de sucesos (con algunas particularidades que en seguida veremos).

A la hora de definir rigurosamente la probabilidad, las consideraciones hechas son suficientes para espacios muestrales discretos, sin embargo para espacios muestrales continuos no resulta útil trabajar sobre un conjunto tan grande como $\mathcal{P}(\Omega)$. Por tanto, la definición rigurosa de probabilidad requiere encontrar algún subconjunto de $\mathcal{P}(\Omega)$ que cumpla ciertos requisitos estructurales, es decir, una σ-álgebra del espacio muestral:

Definición 2.1. Dado un conjunto arbitrario Ω, llamamos σ-álgebra a una familia $\mathcal{A}$ de subconjuntos de Ω si verifica:

- El conjunto vacío está en $\mathcal{A}$: $\emptyset \in \mathcal{A}$
- Para cualquier $A \in \mathcal{A}$, se cumple $A^c \in \mathcal{A}$.
- Dada una colección numerable $\{A_n\} \subset \mathcal{A}$, entonces $\cup_n A_n \in \mathcal{A}$

Para un espacio muestral discreto, la σ-álgebra que consideraremos será directamente $\mathcal{P}(\Omega)$. Para espacios muestrales continuos suelen tomarse algún subconjunto propio, por ejemplo, para el espacio muestral $\mathbb{R}$ se suele utilizar la σ-álgebra de Borel.

Ahora, con todas estas consideraciones formales, estamos en condiciones de definir la función que mida la posibilidad de que, entre todos los posibles sucesos que se pueden dar al realizar el experimento, obtengamos precisamente el que nos interesa.

Definición 2.2. (Definición axiomática de Kolmogorov, 1933). Dado un espacio muestral Ω y una σ-álgebra del mismo, una **probabilidad** (o **medida de probabilidad**) es una aplicación $Pr : \mathcal{A} \longrightarrow [0, 1]$ que verifica:

- $Pr(A) \geq 0$ para todo $A \in \mathcal{A}$.
- Pr es numerablemente aditiva, es decir, dada cualquier colección numerable de sucesos $\{A_n\} \subset \mathcal{A}$ disjuntos entre sí, se tiene $Pr(\cup_n A_n) = \sum_n Pr(A_n)$.
- $Pr(\Omega) = 1$.

La terna $(\Sigma, \mathcal{A}, P)$ se conoce como **espacio de probabilidad** y para cada suceso $A \in \mathcal{A}$, $Pr(A)$ se llama probabilidad de A.

Esta definición axiomática es por un lado consistente con los resultados de teoría de la medida y por otro es coherente con los intentos previos de asignar probabilidades. Estos primeros enfoques, que aparecen ligados a los juegos de azar (experimentos aleatorios con espacios muestrales discretos), incluyen la regla de Laplace y el enfoque frecuentista:

- **Regla de Laplace**
  Cuando todos los posibles resultados del experimento son igualmente probables (esto se puede deducir de las características del experimento, por ejemplo lanzar un dado), podemos asignar a cada suceso $A \subset \Omega$ la siguiente probabilidad

$$Pr(A) = \frac{\#\{\text{casos favorables al suceso A}\}}{\#\{\text{casos posibles}\}}. \tag{2.1}$$

Un ejemplo que encaja en esta situación y para el que usaremos este enfoque es aquel en el que estamos estudiando un marcador que puede presentar k alelos diferentes, todos ellos con la misma probabilidad $\frac{1}{k}$.

- **Enfoque frecuentista**
  Este enfoque, puramente empírico, se basa en la repetición del experimento para tratar de obtener información sobre la probabilidad de un resultado a partir de las realizaciones muestrales (muestras) obtenidas. La magnitud empírica que empleamos para estimar la probabilidad es la frecuencia relativa, ya que para experimentos reproducibles la ley de estabilidad de las frecuencias garantiza que:

$$\begin{gathered} N\text{: número de experimentos realizados.} \\ n\text{: número de resultados de tipo } i \text{ obtenidos.} \\ f_i^N = \tfrac{n}{N}\text{: frecuencia relativa del resultado } i. \end{gathered}$$

$$p_i = \lim_{N\to\infty} f_i^N. \tag{2.2}$$

Por tanto, podemos concluir que la definición general de probabilidad es a la vez coherente con la definición de medida (sujeta a la restricción de que el espacio total mide 1) y generaliza las propiedades de la frecuencia relativa para espacios muestrales de todo tipo.

2.1.1. Probabilidad en genética forense

En la práctica las técnicas estadísticas que indagan en el parentesco se basan en el análisis de marcadores genéticos. Por tanto, se trabaja con espacios de probabilidad discretos, donde el experimento aleatorio es la secuenciación del genotipo de uno o varios individuos en uno o varios marcadores, y el espacio muestral es el conjunto de posibles alelos que puede presentar.

A la hora de afrontar un problema, antes siquiera de plantearnos la obtención de conclusiones a partir de los datos, debemos conocer la probabilidad con la que cada posible alelo puede aparecer en un locus determinado en la población. La importancia de estas probabilidades es sencilla de intuir; por ejemplo en un caso de pruebas de paternidad supongamos que el supuesto padre y el niño comparten un mismo alelo en un locus específico. Si este alelo es muy común en la población de la que proceden, esta coincidencia apenas nos proporciona información a favor o en contra de la paternidad, si por el contrario es un alelo muy poco común, la coincidencia nos invita a pensar que la hipótesis que involucra el parentesco es más probable.

Un planteamiento habitual pasa por utilizar el enfoque frecuentista y aproximar estas probabilidades poblacionales por las frecuencias relativas obtenidas al analizar grandes muestras de la población.

Dado un marcador G, un individuo presenta en general dos copias del mismo que denotaremos G_1/G_2, una heredada por vía materna y otra por vía paterna. Si llamamos A al conjunto de alelos

que se pueden presentar en dicho marcador y tomamos un elemento $a \in A$, vamos a asumir entonces:

$$Pr(G_1 = a) = p_a, \tag{2.3}$$

donde estamos denotando por $Pr(G_1 = a)$ la probabilidad de encontrar un alelo de tipo a en el marcador G en un individuo aleatorio de la población, y p_a denota la frecuencia relativa del alelo a en la base de datos utilizada. Notar que no estamos dando ningún significado a la ordenación de las dos copias G_1 y G_2.

Por tanto, ante un problema de parentesco lo primero con lo que se debe contar es con una base de datos completa que recoja las frecuencias en la población de los posibles alelos de los marcadores elegidos. Esto puede ser un criterio a la hora de decidir qué marcadores utilizar, si escogemos uno que no esté suficientemente estudiado podemos extraer conclusiones falsas.

2.2. Probabilidad condicionada

La probabilidad asignada a un resultado de un experimento aleatorio puede verse influida por otros resultados relacionados o por información extra que hayamos obtenido sobre el mismo. Por ejemplo, supongamos que tomamos parte en un proceso de identificación de restos de un accidente en el que se ha producido una víctima mortal. La policía ha registrado la denuncia de dos familias que denuncian la desaparición de un ser querido en el accidente, un chico y una chica respectivamente. En principio, sin más información, se podría asumir que la probabilidad a priori de que los restos hallados pertenezcan a cada familia es de 1/2. Supongamos ahora que en un examen forense preliminar se determina el sexo de la víctima. La probabilidad ahora será cero para una de las familias. La nueva información sobre los perfiles genéticos involucrados ha alterado (condicionado) las probabilidades.

Definición 2.3. Dado un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ y dados dos sucesos A y B tal que $Pr(B) > 0$, se define la **probabilidad de A condicionada a B** como:

$$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)}. \tag{2.4}$$

Es inmediato ver que en efecto la probabilidad condicionada es probabilidad, pues verifica los tres axiomas enunciados por Kolmogorov.

Los tres resultados siguientes serán cruciales en los cálculos posteriores.

2.2.1. Teorema de factorización

De la expresión (2.4) para la probabilidad condicionada obtenemos la siguiente expresión para la probabilidad de la intersección de dos sucesos:

$$Pr(A \cap B) = Pr(A \mid B)Pr(B). \tag{2.5}$$

Generalizando para una cantidad numerable de sucesos obtenemos el siguiente resultado:

Teorema 2.4 (Teorema de Factorización o Regla del producto)**.** *Si* $\{A_i\}_{i=1}^n$ *son sucesos en un espacio de probabilidad* $(\Omega, \mathcal{A}, P)$ *tales que* $P\left(\cap_{i=1}^{n-1} A_i\right) > 0$ *se cumple:*

$$P\left(\cap_{i=1}^{n} A_i\right) = Pr(A_n \mid \cap_{i=1}^{n-1} A_i) Pr(A_{n-1} \mid \cap_{i=1}^{n-2} A_i) \cdots Pr(A_2 \mid A_1) Pr(A_1). \tag{2.6}$$

Demostración. Basta aplicar reiteradamente la expresión (2.5). Así tomando $A = A_n$, $B = \cap_{i=1}^{n-1} A_i$ obtenemos:

$$P\left(\cap_{i=1}^{n} A_i\right) = Pr(A_n \mid \cap_{i=1}^{n-1} A_i) Pr(\cap_{i=1}^{n-1} A_i),$$

tomando ahora $A = A_{n-1}$, $B = \cap_{i=1}^{n-2} A_i$ obtenemos:

$$Pr(\cap_{i=1}^{n-1} A_i) = Pr(A_{n-1} \mid \cap_{i=1}^{n-2} A_i) Pr(\cap_{i=1}^{n-2} A_i).$$

Continuando iterativamente llegamos a la expresión buscada. □

2.2.2. Teorema de las probabilidades totales

El siguiente teorema nos permite expresar la probabilidad de un suceso en términos de probabilidades condicionadas.

Teorema 2.5 (Teorema de las probabilidades totales)**.** *Si tomamos una partición* $A_1, A_2, ...A_n$ *de* Ω *tal que* $Pr(A_i) > 0$ *para cada* A_i*, entonces para cualquier suceso* $B \in \mathcal{A}$ *se cumple:*

$$Pr(B) = \sum_{i=1}^{n} Pr(B \mid A_i) Pr(A_i). \tag{2.7}$$

Demostración. Basta observar que

$$Pr(B) = Pr(B \cap \Omega) = Pr(B \cap \cup_{i=1}^{n} A_i) = Pr(\cup_{i=1}^{n} (B \cap A_i)) = \sum_{i=1}^{n} Pr(B \cap A_i).$$

Utilizando ahora el la regla del producto (2.6):

$$Pr(B) = \sum_{i=1}^{n} Pr(B \cap A_i) = \sum_{i=1}^{n} Pr(B \mid A_i) Pr(A_i).$$

□

2.2.3. Teorema de Bayes

Tomemos una partición del espacio muestral $A_1, A_2, ...A_n$ tal que $Pr(A_i) > 0$ para todo $i \in \{1, ...n\}$. Supongamos que podemos calcular la probabilidad de cualquier suceso B condicionado a cualquiera de los anteriores. Entonces, combinando la definición de probabilidad condicionada, la conmutatividad de la intersección y el teorema de factorización (2.6) podemos calcular también la probabilidad cuando la influencia se da en sentido contrario:

Proposición 2.6 (Fórmula de Bayes)**.** *Si $A_1, A_2, ...A_n$ es una partición del espacio muestral tal que todos los sucesos tienen probabilidad positiva $Pr(A_i) > 0$ y son disjuntos dos a dos, y B es un suceso con probabilidad positiva $Pr(B) > 0$, entonces:*

$$Pr(A_i \mid B) = \frac{Pr(B \mid A_i)Pr(A_i)}{Pr(B)}. \tag{2.8}$$

- $Pr(A_i)$ es la probabilidad **a priori** del suceso A_i no condicionada, es decir, asignada antes de tener en cuenta cualquier otro suceso.
- $Pr(B \mid A_i)$ es la probabilidad de B condicionada a A_i.
- $Pr(B)$ es la probabilidad de B, luego teniendo en cuenta el teorema de las probabilidades totales $Pr(B) = Pr(B \mid A_1)Pr(A_1) + \cdots + Pr(B \mid A_n)Pr(A_n)$.
- $Pr(A_i \mid B)$ probabilidad **a posteriori** del suceso A_i, es decir, la probabilidad asignada al suceso teniendo en cuenta que se ha producido el suceso B.

Este teorema es la base de la inferencia bayesiana. En esta área de la estadística, se consideran las variables θ y $\boldsymbol{x}$ que son respectivamente el parámetro de una distribución de familia paramétrica conocida y un vector de observaciones extraídas de una población que sigue la distribución anterior. El teorema previo nos permite entonces calcular la verosimilitud del parámetro θ, es decir, su probabilidad condicionada a la muestra:

$$Pr(\theta \mid \boldsymbol{x}) = \frac{Pr(\boldsymbol{x} \mid \theta)Pr(\theta)}{Pr(\boldsymbol{x})}. \tag{2.9}$$

Para realizar este tipo de cálculos nuevamente debe especificarse una probabilidad a priori razonable para el parámetro (en el caso continuo se sustituye por la función de densidad del parámetro $\pi(\theta)$), la probabilidad de la muestra condicionada al parámetro (suele ser evidente por como se produce la recogida de datos) y la probabilidad total de obtener la muestra en la población, que en el caso de que la distribución sea discreta sigue siendo una suma como hemos visto, y en caso de que sea continua se transforma en una integral sobre los posibles valores del parámetro θ.

Para el caso continuo la fórmula sería:

$$Pr(\theta \mid \boldsymbol{x}) = \frac{Pr(\boldsymbol{x} \mid \theta)\pi(\theta)}{\int_\theta Pr(\boldsymbol{x} \mid \theta)\pi(\theta)d\theta}. \tag{2.10}$$

2.3. Independencia

Cuando dos sucesos no se influyen mutuamente se dice que son independientes. Expresado en términos de probabilidad, la definición de independencia es la que sigue:

Definición 2.7. Dada una colección de sucesos $\{A_i\}_{i \in I}$ en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$, se dicen **independientes** si para cualquier subconjunto finito $F \subset I$ se verifica:

$$P\left(\cap_{i \in F} A_i\right) = \prod_{i \in F} Pr(A_i), \tag{2.11}$$

Es decir, la probabilidad de la intersección de sucesos independientes es el producto de sus probabilidades.

En el caso de las probabilidades de los genotipos, la hipótesis de independencia impone restricciones en dos sentidos. Por un lado hemos de exigir la independencia entre marcadores, es decir, que observar un alelo u otro en un marcador no tenga efecto sobre las probabilidades de los alelos de otro. Un caso en el que esto no se cumple sería tomar un marcador relativo al color del cabello y otro al color de la piel, ya que por ejemplo tener el cabello pelirrojo está ligado a presentar la piel y los ojos claros. Son alelos que tienen tendencia a heredarse conjuntamente.

Por otro lado, cada individuo presenta dos copias de cada marcador, necesitamos que la probabilidad de observar un alelo en alguna de esas dos posiciones no dependa de cual sea el otro alelo observado. En genética forense esta hipótesis se denomina **equilibrio de Hardy-Weinberg (HWE)**.

Por tanto, si establecemos algún tipo de orden en los alelos de individuo utilizaremos la notación $\check{G} = \check{G}_1/\check{G}_2$ donde ahora el primero es el alelo paterno y el segundo el materno. En este caso los cálculos de las probabilidades de encontrar aleatoriamente en la población un individuo homocigoto a/a y heterocigoto a/b quedarían respectivamente:

$$\begin{aligned} &Pr(\check{G} = \check{a}/\check{a}) = {p_a}^2, \\ &Pr(\check{G} = \check{a}/\check{b}) = Pr(G = \check{b}/\check{a}) = p_a p_b. \end{aligned} \tag{2.12}$$

En la práctica las técnicas de secuenciación detectan qué alelos están presentes en el genotipo de un individuo, pero no son capaces de determinar su procedencia. Por lo tanto en realidad cuando hablemos de genotipo el orden no tiene ningún significado: un genotipo a/b significa que el individuo presenta ambos alelos, y a la hora de calcular la probabilidad de este suceso hay que tener en cuenta los dos escenarios posibles:

$$\begin{aligned} &Pr(G = a/a) = {p_a}^2, \\ &Pr(G = a/b) \equiv Pr(G = b/a) = 2p_a p_b. \end{aligned} \tag{2.13}$$

2.4. Contrastes de hipótesis

Una cuestión habitual dentro del área de la inferencia estadística es el **contraste de hipótesis**. El objetivo es comprobar, a partir de una muestra, si una hipótesis es o no cierta para la población de la que se ha extraído. Se basa en un planteamiento dicotómico: por un lado la hipótesis que queremos comprobar y por otro su complementaria.

El objetivo de un contraste de hipótesis clásico es probar una de las propuestas más allá de una duda razonable. Para lograrlo, empezamos por establecer una diferencia entre ambas: la hipótesis que queremos comprobar se denomina **hipótesis alternativa** (H_a), mientras que la otra recibe el nombre de **hipótesis nula** (H_0). Para aceptar la primera debemos considerar que existen pruebas significativas en su favor, mientras que para aceptar la segunda basta con que no tengamos evidencias suficientes como para aceptar la alternativa.

Ejemplo 2.8.

$$\begin{cases} H_0: \text{ La población es una distribución normal de media 0.} \\ \\ H_a: \text{ La población es una distribución normal de media distinta de 0.} \end{cases}$$

Para decidir si las muestras obtenidas de la población suponen una prueba concluyente construimos estadísticos, es decir funciones de la muestra, cuyo valor numérico nos indicará si estamos en región de aceptación (de la hipótesis nula) o de rechazo (aceptación de la hipótesis alternativa). Para decidir a partir de qué valores se considera una u otra debemos fijar el **nivel de significación** α del test. Este nivel indica cual es el riesgo que corremos de rechazar la hipótesis nula siendo cierta (error de tipo I). La razón por la que se controla este error a la hora de definir el test es porque se considera más grave rechazar la hipótesis nula falsamente que la alternativa, por lo tanto lo primero ha de ser garantizar que este nivel no es muy alto.

Capítulo 3

De la genética a la estadística

Con ayuda de las herramientas del capítulo anterior, vamos ahora a expresar los problemas de genética forense en términos estadísticos. Para ello vamos a presentar el modelo básico para este tipo de problemas a través de ejemplos sencillos y luego exploraremos algunas situaciones más complejas y los recursos necesarios para afrontarlas.

3.1. Planteamiento general

Los problemas de búsqueda de relaciones parten de una sospecha sobre la existencia de una o varias posibles relaciones de parentesco. A fin de decantarnos a favor o en contra de esta sospecha, se hace acopio de muestras de los individuos involucrados y de ellas se extrae información sobre el genotipo de distintos marcadores previamente seleccionados. A partir de estos datos, el objetivo es construir herramientas estadísticas que nos permitan extraer conclusiones.

Para introducir las peculiaridades del planteamiento de este tipo de problemas, vamos a comenzar presentando un caso sencillo e ilustrativo.

3.1.1. Modelo básico: Standard Duo Case

Se solicita una prueba de paternidad para determinar si el sujeto AF[1] es en efecto el padre biológico de un niño CH[2]. Para determinarlo, se toman muestras del supuesto padre (AF) y del hijo (CH), y se determina su genotipo para una colección de marcadores prefijada.

El **planteamiento de la hipótesis** del problema es en este caso muy sencillo:

$$\begin{cases} H_1: \text{ El supuesto padre (AF) es el padre biológico del niño.} \\ \\ H_2: \text{ El supuesto padre (AF) no es el padre biológico del niño.} \end{cases}$$

Sobre este planteamiento cabe destacar tres aspectos. En primer lugar, vemos que las hipótesis están planteadas verbalmente en vez de en términos matemáticos, algo típico en problemas de

[1] Del inglés *Alleged Father* (supuesto padre).
[2] Del inglés Child (niño/a)

genética forense. En segundo lugar, como hemos mencionado en la sección 2.4, en general en un contraste ambas hipótesis deben ser mutuamente excluyentes y sus probabilidades deben sumar 1. Para aplicar los recursos presentados en este capítulo la situación no es exactamente esta, pues como hipótesis de no paternidad (H_2) se incluirán los casos en los que el verdadero padre biológico es un individuo no relacionado con AF, es decir, un individuo aleatorio de la población. Si por ejemplo se sospechase que el hermano del individuo AF pudiera ser el padre el tratamiento del problema cambiaría. Ante cualquier problema es por tanto necesario tener cuidado y empezar por determinar si encaja en el planteamiento anterior. Y en tercer lugar, y más importante, no es un contraste de hipótesis al uso, pues carecemos de lo que se suele denominar hipótesis nula.

El contraste de hipótesis tradicional descansa sobre la idea de determinar si existen pruebas significativas suficientes para probar la hipótesis alternativa más allá de una duda razonable. La consideración de una hipótesis nula en un problema implica automáticamente una asimetría entre las hipótesis involucradas; esta se supone cierta de entrada y la carga de la prueba recae sobre la alternativa. El eje central es controlar que las probabilidad de rechazar falsamente la hipótesis nula (error tipo I) no sea muy alta. Para ello se fija el nivel de significación del test estableciendo zonas de aceptación y zonas de rechazo para el valor del estadístico. La decisión se toma evaluando el estadístico para los valores de la muestra.

En genética forense esta asimetría no existe pues se colocan todas las hipótesis al mismo nivel. Por tanto los mecanismos habituales no sirven: no hablaremos de aceptar o rechazar una hipótesis, en su lugar trataremos de encontrar la hipótesis más probable (aunque en general la diferencia entre las probabilidades de una y otra suele ser tan grande que no suele quedar ninguna duda razonable). La herramienta que utilizaremos para cuantificar y comparar estas probabilidades es la **razón de verosimilitudes (LR)**

Razón de verosimilitudes

La **razón de verosimilitudes (LR)** es el cociente de la función de verosimilitud de los datos bajo una hipótesis y bajo la otra, su expresión general es:

$$LR = \frac{Pr(datos \mid H_1)}{Pr(datos \mid H_2)}. \tag{3.1}$$

La función de verosimilitud de los datos bajo una hipótesis H es la probabilidad de obtener dichos datos si la población de la que se extraen cumple con los supuestos de H.

La razón de verosimilitudes se interpreta entonces como una medida de cuánto más probables son los datos obtenidos bajo H_1 que bajo H_2. Evidentemente si es inferior a 1 la hipótesis del numerador hace la muestra menos probable que la del denominador, y viceversa si LR supera la unidad. En nuestro caso, este cálculo nos dirá cuánto más probables son los genotipos observados bajo la hipótesis de paternidad que bajo la de no paternidad.

Descomponemos esta probabilidad utilizando la regla del producto (2.6) como sigue

$$LR = \frac{Pr(datos \mid H_1)}{Pr(datos \mid H_2)} = \frac{Pr(G_{AF}, G_{CH} \mid H_1)}{Pr(G_{AF}, G_{CH} \mid H_2)} = \frac{Pr(G_{CH} \mid G_{AF}, H_1)}{Pr(G_{CH} \mid G_{AF}, H_2)} \cdot \frac{Pr(G_{AF} \mid H_1)}{Pr(G_{AF} \mid H_2)}. \tag{3.2}$$

Teniendo en cuenta que hemos dicho que en estos problemas ambas hipótesis son simétricas en cuanto al tratamiento matemático, el lector podría preguntarse que significa considerar H_1 en el numerador y H_2 en el denominador, si tiene algún significado o si son intercambiables. En general en problemas como este donde solo planteamos dos hipótesis se suele tomar en el denominador la hipótesis de no efecto (no relación). El problema surge cuando tenemos más de dos hipótesis, en este caso habría que especificar con respecto a cual de ellas estamos calculando los LR.

Vamos a asumir por el momento las siguientes condiciones razonables:

- La probabilidad del genotipo observado para el supuesto padre (AF) no depende de la hipótesis considerada:

$$Pr(G_{AF} \mid H_1) = Pr(G_{AF} \mid H_2) = Pr(G_{AF}). \tag{3.3}$$

- La probabilidad de observar los genotipos del supuesto padre (AF) y del niño (CH) son independientes bajo H_2, como consecuencia:

$$Pr(G_{CH} \mid G_{AF}, H_2) = Pr(G_{CH}). \tag{3.4}$$

Esta última simplificación pone de manifiesto de nuevo la importancia de determinar correctamente las diferentes hipótesis a las que nos enfrentamos; si, por ejemplo, un familiar del supuesto padre (AF) pudiese ser también el padre biológico, esta segunda simplificación sería falsa y los resultados no serían válidos.

Ahora, imponiendo las dos condiciones previas en (3.2), obtenemos la siguiente expresión simplificada de la razón de verosimilitudes para el Standard Duo Case:

$$LR = \frac{Pr(G_{CH} \mid G_{AF}, H_1)}{Pr(G_{CH})}. \tag{3.5}$$

En el caso de que analicemos varios marcadores, que es el escenario más habitual, cobra relevancia la condición de **independencia** que hemos impuesto sobre ellos, pues esto nos permite calcular la razón de verosimilitud total de los datos como el producto de la obtenida para cada marcador.

Es decir, si consideramos el perfil genético para $\{1, 2, ..., m\}$ marcadores independientes, y si LR_i representa la razón de verosimilitud para el marcador $i \in \{1, 2, ..., m\}$ calculada como acabamos de ver, la razón de verosimilitud total la obtendremos como:

$$LR = \prod_{i=1}^{m} LR_i. \tag{3.6}$$

Como nota final, otro caso conocido es el Standard Trio Case, que corresponde a una situación totalmente análoga a la planteada en esta sección pero donde el genotipo de la madre es también conocido. Podemos asumir las mismas simplificaciones, lo que sucederá es que tendremos más información para calcular las probabilidades.

Probabilidad a posteriori

Una vez obtenido el LR, es posible transformar esta magnitud en la **probabilidad a posteriori** de cada una de las hipótesis involucradas. En particular nos interesa la probabilidad de la paternidad (H_1) dada la evidencia genética, conocida como **Índice de Essen-Möller (W)**

$$W = Pr(H_1 \mid datos). \tag{3.7}$$

Para ello solo necesitamos la probabilidad a priori de cada una de las hipótesis y el teorema de Bayes.

- La **probabilidad a priori** es la probabilidad asignada a un suceso simplemente razonando con información cualitativa sobre las circunstancias en que ocurre o sobre sus características, es decir, sin tener en cuenta los datos. Si no hay razones para preferir una hipótesis sobre la otra, se asume que ambas son equiprobables:

$$Pr(H_1) = Pr(H_2) = \frac{1}{2}. \tag{3.8}$$

- El **teorema de Bayes**, aunque ya lo hemos presentado en su versión habitual en el capítulo 2, puede formularse para el caso general en el que existen k hipótesis. Sea $i \in \{1, 2, ..., k\}$:

$$Pr(H_i \mid datos) = \frac{Pr(datos \mid H_i)Pr(H_i)}{Pr(datos)} = \frac{Pr(datos \mid H_i)Pr(H_i)}{\sum_{j=1}^{k} Pr(datos \mid H_j)Pr(H_j)}. \tag{3.9}$$

Donde hemos utilizado el teorema de la probabilidad total para calcular $Pr(datos)$. Aplicando la condición (3.8) al caso particular del Standard Duo Case:

$$W = \frac{Pr(datos \mid H_1)Pr(H_1)}{Pr(datos \mid H_1)Pr(H_1) + Pr(datos \mid H_2)Pr(H_2)} = \frac{Pr(datos \mid H_1)}{Pr(datos \mid H_1) + Pr(datos \mid H_2)} = \frac{LR}{LR+1}. \tag{3.10}$$

La ventaja de este índice es que se trata de una probabilidad sencilla de interpretar, la desventaja, que para obtenerla estamos obligados a especificar las probabilidades a priori que, como ya hemos visto, se basan en suposiciones.

3.1.2. Ejemplo 1. Standard Duo Case

Nos encontramos ante un problema de determinación de paternidad. Para estudiar este caso hemos obtenido el genotipo del supuesto padre (AF) y del hijo (CH) para dos marcadores independientes: $D3S1358$ y $TPOX$. Los datos sobre las frecuencias de cada alelo están extraídos de una base de datos de población noruega. Estas cifras, además de los ejemplos que utilizaremos se pueden encontrar en [7].

Planteamos el contraste de hipótesis:

$$\begin{cases} H_1: \text{ El supuesto padre (AF) es el padre biológico del niño.} \\ \\ H_2: \text{ El padre biológico no es AF (sino un individuo no relacionado con AF)} \end{cases}$$

La Figura 3.1 representa la situación bajo cada una de las hipótesis. Bajo la hipótesis H_2 se coloca un individuo desconocido como padre biológico del niño (*added 1*). Cada genotipo se denotará por G y un subíndice indicando el individuo al que pertenece. En este caso $G_{AF} = 17/18$, $G_{CH} = 17/17$ para el marcador $D3S1358$ y $G_{AF} = 8/8$, $G_{CH} = 8/8$ para el $TPOX$. El genotipo de la madre se desconoce, como corresponde al Standard Duo Case.

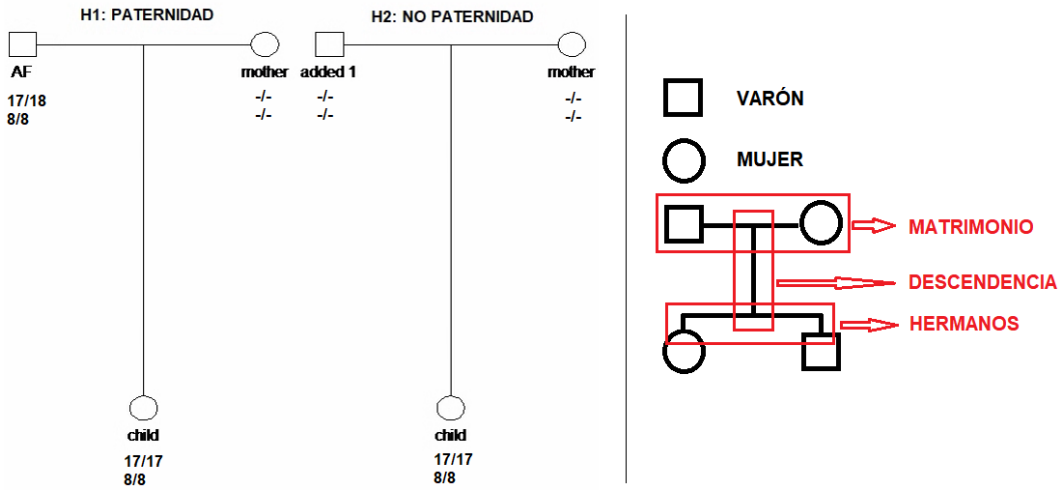

Figura 3.1: Ejemplo 1. Standard Duo Case

Los datos sobre las frecuencias poblacionales para los alelos involucrados son $p_{17} = 0{,}2040$, $p_{18} = 0{,}1394$ y $p_8 = 0{,}5539$ [3].

Razón de verosimilitudes (LR)

El objetivo es obtener e interpretar la razón de verosimilitudes para este contraste. Comenzamos calculando la razón de verosimilitudes para cada marcador en solitario. Como se trata de un Standard Duo Case y vamos a considerar ambas hipótesis equiprobables a priori, utilizaremos la ecuación (3.5):

$$LR = \frac{Pr(G_{CH} \mid G_{AF}, H_1)}{Pr(G_{CH})}. \tag{3.11}$$

Marcador 1: D3S1358 Los genotipos involucrados son $G_{AF} = 17/18$ y $G_{CH} = 17/17$

- Calculamos $Pr(G_{CH} \mid G_{AF}, H_1)$
  Bajo la hipótesis H_1, el niño tendría que haber heredado uno de sus alelos 17 del padre (AF) y el otro de la madre. Si suponemos equilibrio de Hardy-Weinberg (HWE):

$$Pr(G_{CH} \mid G_{AF}, H_1) = Pr(\text{hereda alelo 17 de AF}) \cdot Pr(\text{hereda alelo 17 de la madre}). \tag{3.12}$$

  La probabilidad del primer suceso es $1/2$, pues estamos considerando un caso sencillo y sin mutaciones con lo que la descendencia hereda cualquiera de los dos alelos del progenitor

con la misma probabilidad. La probabilidad del segundo suceso, al no tener datos sobre el genotipo de la madre, vamos a calcularla como la probabilidad de encontrar un alelo 17 de manera aleatoria en la población: p_{17}. Obtenemos finalmente:

$$Pr(G_{CH} \mid G_{AF}, H_1) = \frac{1}{2}p_{17}. \tag{3.13}$$

- Calculamos $Pr(G_{CH})$
  Como carecemos de información sobre la madre, calcular la probabilidad de que el niño presente este genotipo bajo la hipótesis de no paternidad es equivalente a calcular la probabilidad de encontrar este genotipo entre la población de forma aleatoria. Basta aplicar de nuevo que si asumimos HWE ambos alelos son independientes y obtenemos:

$$Pr(G_{CH}) = {p_{17}}^2. \tag{3.14}$$

Como vemos, únicamente nos hace falta conocer la probabilidad en la población del alelo 17. Consultando la base de datos del problema encontramos que $p_{17} = 0{,}2040$, con lo que finalmente:

$$LR_1 = \frac{\frac{1}{2}p_{17}}{{p_{17}}^2} = \frac{1}{2p_{17}} = 2{,}4504. \tag{3.15}$$

Si a mayores queremos conocer la verosimilitud de alguna de las hipótesis, se calcula fácilmente teniendo en cuenta de nuevo la regla del producto (2.6) y todos los genotipos compatibles con los alelos observados:

$$\begin{aligned}
&Pr(datos \mid H_1) = Pr(G_{CH}, G_{AF} \mid H_1) = Pr(G_{CH} \mid G_{AF}, H_1) \cdot Pr(G_{AF} \mid H_1) = \\
&= Pr(G_{CH} = 17/17 \mid \check{G}_{AF} = \check{1}\check{7}/\check{1}\check{8}, H_1) Pr(\check{G}_{AF} = \check{1}\check{7}/\check{1}\check{8} \mid H_1) \\
&+ Pr(G_{CH} = 17/17 \mid \check{G}_{AF} = \check{1}\check{8}/\check{1}\check{7}, H_1) Pr(\check{G}_{AF} = \check{1}\check{8}/\check{1}\check{7} \mid H_1) = \\
&= \frac{1}{2}p_{17} Pr(\check{G}_{AF} = \check{1}\check{7}/\check{1}\check{8} \mid H_1) + \frac{1}{2}p_{17} Pr(\check{G}_{AF} = \check{1}\check{8}/\check{1}\check{7} \mid H_1) \mid H_1) = \\
&= p_{18}p_{17}^2 = 0{,}1394 \cdot 0{,}2040^2 = 0{,}00580127,
\end{aligned} \tag{3.16}$$

$$\begin{aligned}
&Pr(datos \mid H_2) = Pr(G_{CH}, G_{AF} \mid H_2) = Pr(G_{CH} \mid G_{AF}, H_2) \cdot Pr(G_{AF} \mid H_2) = \\
&= Pr(G_{CH} = 17/17 \mid \check{G}_{AF} = \check{1}\check{7}/\check{1}\check{8}, H_2) Pr(\check{G}_{AF} = \check{1}\check{7}/\check{1}\check{8} \mid H_2) \\
&+ Pr(G_{CH} = 17/17 \mid \check{G}_{AF} = \check{1}\check{8}/\check{1}\check{7}, H_2) Pr(\check{G}_{AF} = \check{1}\check{8}/\check{1}\check{7} \mid H_2) = \\
&= Pr(G_{CH} = 17/17) \cdot \left[Pr(\check{G}_{AF} = \check{1}\check{7}/\check{1}\check{8}) + Pr(\check{G}_{AF} = \check{1}\check{8}/\check{1}\check{7} \right] = \\
&= p_{17}^2 \cdot 2p_{18}p_{17} = 0{,}2040^2 \cdot 2 \cdot 0{,}2040 \cdot 0{,}1394 = 0{,}002366918.
\end{aligned}$$

Marcador 2: TPOX Los genotipos involucrados son $G_{AF} = 8/8$ y $G_{CH} = 8/8$

- Calculamos $Pr(G_{CH} \mid G_{AF}, H_1)$
  Bajo la hipótesis H_1, el niño tendría que haber heredado uno de sus alelos 8 del padre (AF)

y el otro de la madre. Bajo la suposición de existencia de equilibrio de Hardy-Weinberg (HWE):

$$Pr(G_{CH} \mid G_{AF}, H_1) = Pr(\text{alelo 8 paterno}) \cdot Pr(\text{alelo 8 materno}). \tag{3.17}$$

La probabilidad del primer suceso es 1 ya que AF es homocigoto para el alelo de interés. La probabilidad del segundo suceso, al no tener datos sobre el genotipo de la madre, vamos a calcularla como la probabilidad de encontrar un alelo 8 de manera aleatoria en la población: p_8. Obtenemos finalmente:

$$Pr(G_{CH} \mid G_{AF}, H_1) = p_8. \tag{3.18}$$

- Calculamos $Pr(G_{CH})$
  Aplicando exactamente el mismo razonamiento que en el marcador anterior:

$$Pr(G_{CH}) = {p_8}^2. \tag{3.19}$$

La única probabilidad poblacional que es preciso conocer es la del alelo 8, $p_8 = 0{,}55391$, con lo que finalmente:

$$LR_2 = \frac{p_8}{{p_8}^2} = \frac{1}{p_8} = 1{,}8053. \tag{3.20}$$

Obtenemos la verosimilitud de cada hipótesis siguiendo el mismo procedimiento que hemos utilizado en el caso del marcador anterior:

$$\begin{aligned} &Pr(datos \mid H_1) = Pr(G_{CH} = 8/8 \mid G_{AF} = 8/8, H_1) Pr(G_{AF} = 8/8 \mid H_1) = \\ &= 1 \cdot p_8 \cdot Pr(G_{AF} = 8/8 \mid H_1) = p_8^3 = 0{,}5539^3 = 0{,}1699394, \\ &Pr(datos \mid H_2) = Pr(G_{CH} = 8/8 \mid G_{AF} = 8/8, H_2) Pr(G_{AF} = 8/8 \mid H_2) = \\ &= Pr(G_{CH} = 8/8) \cdot Pr(G_{AF} = 8/8) = p_8^4 = 0{,}5539^4 = 0{,}09412944. \end{aligned} \tag{3.21}$$

Cálculos globales

Finalmente obtenemos la razón de verosimilitud total para estos datos, es decir, la combinación de la obtenida para ambos marcadores. Como dichos marcadores son independientes, basta aplicar la expresión (3.6) y obtenemos:

$$LR = LR_1 \cdot LR_2 = 2{,}4504 \cdot 1{,}8053 = 4{,}4237. \tag{3.22}$$

A la vista de los resultados, es entre 4 y 5 veces más probable obtener los datos de este problema (es decir, los perfiles genéticos de los dos individuos involucrados) si la hipótesis del parentesco es cierta que si no lo es. Evidentemente estas evidencias son demasiado débiles como para pronunciarnos a favor o en contra de la veracidad de las hipótesis, pero ilustra el esqueleto del proceso que se lleva a cabo en los laboratorios con decenas de marcadores en vez de un par de ellos y con modelos un poco más sofisticados.

De igual modo se puede calcular la verosimilitud global de cada hipótesis (H) considerando los datos de ambos marcadores. Como ambos marcadores son independientes, utilizando la expresión (2.11), obtenemos que la verosimilitud global es el producto de las verosimilitudes para ambos marcadores:

$$\begin{aligned} &Pr(datos \mid H) = Pr(G_{AF} = 17/18, G_{CH} = 17/17, G_{AF} = 8/8, G_{CH} = 8/8 \mid H) = \\ &= Pr(G_{AF} = 17/18, G_{CH} = 17/17 \mid H) \cdot Pr(G_{AF} = 8/8, G_{CH} = 8/8 \mid H). \end{aligned} \tag{3.23}$$

Utilizando entonces las verosimilitudes de las hipótesis para ambos marcadores (calculadas en (3.16) y (3.21) respectivamente), obtenemos las verosimilitudes globales de la hipótesis de paternidad (H_1) y de no paternidad (H_2):

$$\begin{aligned} &Pr(datos \mid H_1) = Pr(datosD3S1358 \mid H_1) \cdot Pr(datosTPOX \mid H_1) = \\ &= 0{,}00580127 \cdot 0{,}1699394 = 0{,}0009858643, \\ &Pr(datos \mid H_2) = Pr(datosD3S1358 \mid H_2) \cdot Pr(datosTPOX \mid H_2) = \\ &= 0{,}094134933 \cdot 0{,}169946848 = 0{,}01599794. \end{aligned} \tag{3.24}$$

Probabilidad de la paternidad

Continuamos aplicando las herramientas explicadas a este caso del Standard Duo Case. Así, nos interesa ahora transformar la razón de verosimilitud obtenida en probabilidad a posteriori, es decir, en una magnitud que exprese la plausibilidad de las hipótesis anteriores a la vista de los datos obtenidos.

En particular, en este tipo de situaciones lo interesante es conocer la probabilidad de la paternidad condicionada a los datos obtenidos. Como ya hemos visto, a esta probabilidad se la conoce como índice de Essen-Möller (W), y nos exige estimar una probabilidad a priori para cada una de las hipótesis. Asumiendo que ambas hipótesis son igualmente probables, podemos aplicar la expresión (3.10) y obtenemos:

$$W = \frac{LR}{LR+1} = 0{,}8156. \tag{3.25}$$

Es decir, si inicialmente partimos de que la hipótesis de paternidad tiene un 50 % de probabilidad de ser cierta, teniendo en cuenta los datos esta probabilidad crece hasta superar el 80 %. Por tanto los datos son coherentes y reafirman la hipótesis de paternidad.

3.1.3. Ejemplo 2. Standard Trio Case

Este problema se trata de determinar si un hombre (AF) es el padre biológico de un niño. Se plantean las siguientes hipótesis:

$$\begin{cases} H_1: \text{ El supuesto padre (AF) es el padre biológico del niño.} \\ \\ H_2: \text{ El padre biológico no es AF (sino un individuo no relacionado con AF)} \end{cases}$$

En este caso contamos con información sobre los genotipos relativos a un único marcador del niño CH, la madre indiscutida MO y el supuesto padre AF. Los genotipos involucrados son $G_{AF} = A/A$, $G_{MO} = B/B$ y $G_{CH} = A/B$ (Figura 3.2). Las frecuencias poblaciones de los alelos involucrados vamos a suponer que son $p_A = p_B = 0{,}05$.

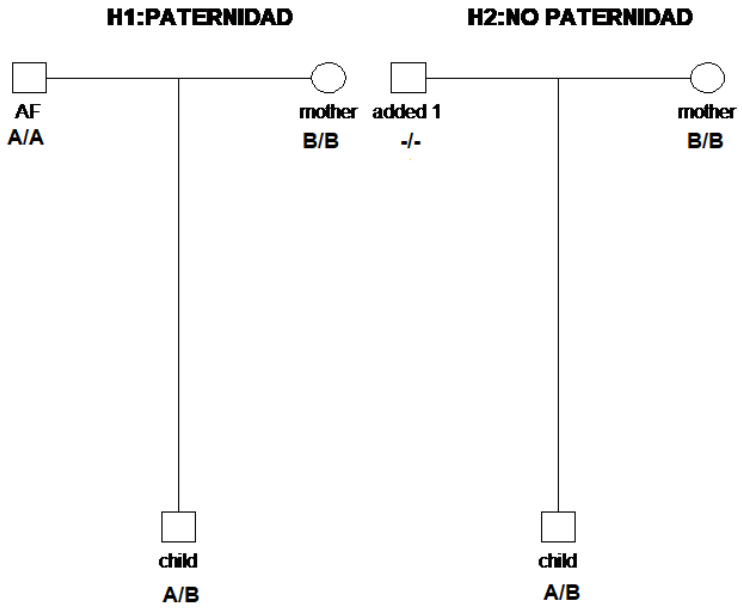

Figura 3.2: Ejemplo 2. Standard Trio Case

Vamos a calcular la razón de verosimilitudes para este marcador: Recuperamos la definición general (3.1)

$$LR = \frac{Pr(datos \mid H_1)}{Pr(datos \mid H_2)}, \tag{3.26}$$

y reescribimos cada término en base a nuestros datos utilizando el teorema de factorización (2.6):

$$\begin{aligned} &Pr(datos \mid H_1) = Pr(G_{CH}, G_{AF}, G_{MO} \mid H_1) = \\ &= Pr(G_{CH} \mid G_{AF}, G_{MO}, H_1) \cdot Pr(G_{AF} \mid G_{MO}, H_1) \cdot Pr(G_{MO} \mid H_1). \end{aligned} \tag{3.27}$$

Como estamos considerando que la maternidad no se pone en duda y la madre es homocigota para el alelo B, que el alelo B del niño es materno está fuera de discusión. Considerando esto, el único suceso que realmente está condicionando la probabilidad del genotipo del hijo es el genotipo del supuesto padre AF (que consideramos el verdadero padre bajo H_1).

$$Pr(G_{CH} \mid G_{AF}, G_{MO}, H_1) = Pr(G_{CH} \mid G_{CH} = B/-, G_{AF}, H_1). \tag{3.28}$$

Ahora bien, el niño ha de heredar el alelo A del padre. En este caso, como el padre también es homocigoto, cualquiera de sus alelos encaja en este escenario, y como la probabilidad de heredar cada uno de ellos es $1/2$ obtenemos:

$$Pr(G_{CH} \mid G_{AF}, H_1) = \frac{1}{2} + \frac{1}{2} = 1. \tag{3.29}$$

Dicho de otra forma, bajo la hipótesis de paternidad el niño tendría el genotipo que ya presenta y es el único genotipo compatible con esta hipótesis.

Tanto el genotipo de la madre como del padre vamos a considerarlos independientes entre sí e independientes de la hipótesis que estemos asumiendo. Por tanto:

$$Pr(G_{AF} \mid G_{MO}, H_1) = Pr(G_{AF}) \text{ y } Pr(G_{MO} \mid H_1) = Pr(G_{MO}), \tag{3.30}$$

entonces sustituyendo (3.28),(3.29) y (3.30) en (3.27) obtenemos:

$$Pr(datos \mid H_1) = Pr(G_{AF}) \cdot Pr(G_{MO}). \tag{3.31}$$

Los cálculos para $Pr(datos \mid H_2)$ son análogos y se basan en la expresión (3.27) particularizada para H_2. El único cálculo cuyo resultado se ve alterado es el de $Pr(G_{CH} \mid G_{AF}, G_{MO}, H_2)$. En este caso hemos de tener en cuenta que el genotipo de AF no condiciona en absoluto al del niño, por lo que (como el alelo B es materno), la probabilidad del genotipo del niño se calcula como la probabilidad de hallar un alelo A en la población:

$$Pr(G_{CH} \mid G_{AF}, G_{MO}, H_2) = Pr(G_{CH} \mid G_{MO}, H_2) = Pr(G_{CH} \mid G_{CH} = B/-) = p_A, \tag{3.32}$$

por tanto, volviendo a la expresión del LR (3.26), obtenemos:

$$LR = \frac{1 \cdot Pr(G_{AF})Pr(G_{MO})}{p_B Pr(G_{AF})Pr(G_{MO})} = \frac{1}{p_A} = \frac{1}{0{,}05} = 20, \tag{3.33}$$

y concluimos que los datos obtenidos para este marcador son 20 veces más probables si la hipótesis de paternidad es cierta que si no lo es.

3.1.4. Ejemplo 3. Caso degenerado

Se plantea el problema de determinar si un individuo (AF) es el padre de un niño (CH). Para ello, de nuevo hemos obtenido muestras y extraído el genotipo de ambos para el marcador $D3S1358$. En este caso los perfiles obtenidos son: $G_{AF} = 14/15$ y $G_{CH} = 16/17$.

La situación se corresponde de nuevo con un Standard Duo Case, donde las hipótesis son las siguientes (Figura 3.3):

$$\begin{cases} H_1 : \text{AF es el padre biológico del niño.} \\ \\ H_2 : \text{El padre biológico del niño es un individuo no relacionado con } AF). \end{cases}$$

Es inmediato a la vista de los datos que con el modelo de herencia adoptado hasta ahora nuestros datos son incompatibles con la hipótesis de paternidad.

En términos de la razón de verosimilitud, los cálculos son sencillos, usando (3.5):

$$Pr(G_{CH} \mid G_{AF}, H_1) = 0 \Longrightarrow LR = 0. \tag{3.34}$$

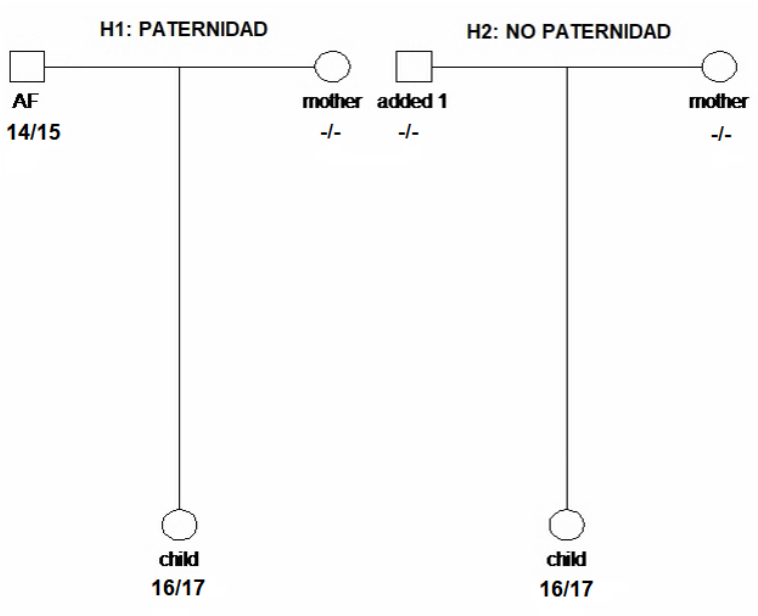

Figura 3.3: Ejemplo 3. Standard Duo Case (caso degenerado)

3.2. Modelo completo

Ya hemos visto el esqueleto de los problemas presentando casos muy sencillos. Ahora vamos a ver como podemos adaptarlo a situaciones más complejas. El razonamiento subyacente es muy similar, los cálculos son más laboriosos e introducimos parámetros nuevos. A cambio el modelo ampliado nos permite considerar la existencia de alelos silenciosos, dropouts o mutaciones que pueden explicar en ocasiones la aparente inconsistencia de los datos. Para ilustrar cada caso vamos a utilizar los ejemplos anteriores.

3.2.1. Mutaciones

En general en los análisis de laboratorio es conveniente tener en cuenta la posibilidad de mutación[3]. A continuación se presentan algunos recursos para incluir esta situación en nuestro planteamiento.

Una forma de hacerlo es fijar una **tasa de mutación** R y considerar todas las mutaciones igualmente probables. Dependiendo de la precisión que necesitemos este recurso puede ser demasiado simple.

Un enfoque más completo es a través de una **matriz de mutaciones**. Supongamos que $A = \{a, b, c, d\}$ el conjunto de alelos que puede presentar el marcador que estamos analizando. Entonces podemos construir la siguiente tabla:

[3]Hay dos tipos de mutaciones: germinales y somáticas. Las segundas afectan únicamente a un individuo mientras que las primeras tienen lugar en las células sexuales y se transmiten a la descendencia. Las mutaciones germinales son en general las que tienen importancia en problemas de genética.

	a	b	c	d
a	$m_{1,1}$	$m_{1,2}$	$m_{1,3}$	$m_{1,4}$
b	$m_{2,1}$	$m_{2,2}$	$m_{2,3}$	$m_{2,4}$
c	$m_{3,1}$	$m_{3,2}$	$m_{3,3}$	$m_{3,4}$
d	$m_{4,1}$	$m_{4,2}$	$m_{4,3}$	$m_{4,4}$

En ella cada fila indica el alelo antes de la mutación y la columna el resultado tras la mutación. La magnitud $m_{i,j}$ expresa la probabilidad de que el alelo i sea heredado como j en la siguiente generación (la diagonal recoge el caso de no mutación):

$$m_{i,j} = Pr(i \to j) \tag{3.35}$$

Si incluimos en la tabla todas las variantes posibles que puede presentar el marcador en cuestión, todas las filas deben sumar uno. A esta matriz cuadrada $M = (m_{i,j})_{i,j=1}^{4}$ se la conoce como matriz de mutaciones.

Detallar estas matrices es arduo y muy costoso computacionalmente, por lo tanto se suele optar por una solución intermedia: la introducción de un **modelo paramétrico**. Estos modelos proporcionan un algoritmo capaz de definir toda la matriz de mutaciones a partir de unos cuantos parámetros. Además deben cumplir los siguientes requisitos: ser matemáticamente consistentes, estar formulado en términos de parámetros interpretables y ser biológicamente razonables.

En particular, veamos un modelo paramétrico muy común para **marcadores STR**. Estos marcadores consisten, como ya hemos dicho, en secuencias de bases repetidas. La cantidad de repeticiones que presenta el marcador es lo que diferencia a cada uno de los alelos posibles. Así, las mutaciones son simplemente fallos a la hora de replicar el ADN, resultando en que la copia presenta secuencias de más o de menos respecto al alelo original. En base a la descripción anterior, resulta intuitivo pensar que no todas las mutaciones son igualmente probables, sino que cuanto más similar sea el número de copias de dos alelos, más probable es pasar de uno a otro por error. Los modelos que siguen esta planteamiento se denominan modelos **stepwise**, ya que tienen en cuenta cuantos *pasos* tiene que dar una mutación para pasar de un alelo a otro.

Los parámetros utilizados en estos modelos son los siguientes:

- Una **tasa de mutación** R para pasar del alelo original a otro con una copia más o una copia menos, es decir, mutaciones de un paso.
- El **rango de mutación** r representa la probabilidad de una mutación de n pasos con respecto a la de $n-1$ pasos.

En general para realizar los cálculos de un modelo con mutación hace falta ayuda informática. La excepción puede ser el Standard Duo Case, cuya expresión es relativamente sencilla.

Ejemplo 3.1. Vamos a trabajar con el problema planteado en el Ejemplo 3. Standard Duo Case degenerado (sección 3.1.4). Vamos a introducir la posibilidad de que se haya producido una

mutación. En primer lugar vamos a obtener la expresión de la verosimilitud de los datos bajo la hipótesis de paternidad en términos de la matriz de mutaciones M:

$$Pr(datos \mid H_1) = Pr(G_{CH}, G_{AF} \mid H_1) = Pr(G_{CH} \mid G_{AF}, H_1) \cdot Pr(G_{AF} \mid H_1). \tag{3.36}$$

Al considerar mutaciones lo que varía es el cálculo del primer término, pues ahora debemos considerar la posibilidad de que cada alelo del padre haya sido heredado como cualquiera de los alelos del hijo:

$$\begin{aligned} &Pr(G_{CH} \mid G_{AF}, H_1) = \\ &= \frac{1}{2}[Pr(14 \to 16)p_{17} + Pr(14 \to 17)p_{16} + Pr(15 \to 16)p_{17} + Pr(15 \to 17)p_{16}] = \\ &= m_{14,16}p_{17} + m_{14,17}p_{16} + m_{15,16}p_{17} + m_{15,17}p_{16}. \end{aligned} \tag{3.37}$$

Como el resto de términos se calcula de la manera habitual obtenemos:

$$LR = \frac{p_{16}(m_{14,17} + m_{15,17}) + p_{17}(m_{14,16} + m_{15,16})}{4p_{16}p_{17}}. \tag{3.38}$$

Como podemos ver, si no tenemos en cuenta la presencia de mutaciones, habremos descartado la hipótesis de paternidad asignándole una probabilidad de cero cuando todavía es viable considerando modelos más complejos.

A veces solo interesa estudiar algunos alelos del marcador escogido. Un posible planteamiento es incluir los alelos de interés y uno a mayores que englobe a todos los restantes para que las probabilidades totales sumen 1, es lo que se conoce como modelo minimal. La ventaja es que agiliza mucho los cálculos especialmente en modelos con mutaciones, pero los resultados pueden verse afectados al considerar este tipo de simplificaciones. En general se recomienda escoger el modelo antes de conocer los datos, de lo contrario puede parecer que los datos condicionan el modelo y esto compromete los resultados. Lo más fiable es utilizar siempre el modelo completo. Además, se recomienda comprobar la robustez de los resultados obtenidos comparándolos con los que se obtienen al implementar otros modelos para las mutaciones.

3.2.2. Subpoblaciones: corrección theta

Hasta ahora hemos estado considerando en todos los ejemplos la existencia de equilibrio de Hardy-Weinberg HWE (ver sección 2.3). Sin embargo, hay muchas situaciones en las que esta suposición no se ajusta a la realidad. En la práctica, una población no es una masa homogénea en la que cada individuo se relaciona con igual probabilidad con cualquier otro. Existen factores como la proximidad, la edad u otros factores étnicos, culturales, socioeconómicos o políticos que pueden inducir la formación de subpoblaciones, es decir, de subgrupos dentro de la población original cuyos individuos tienen mayor tendencia a relacionarse entre sí. Esto influye en las probabilidades de cada alelo en la población, pero además el hecho de que esta tendencia se prolongue a lo largo de varias generaciones hace que aumente la endogamia, en el sentido de que incluso dos progenitores que no se conocen es probable que estén relacionados en algún grado.

Naturalmente esto es cierto en global para toda la población pero a gran escala se difumina tanto que despreciamos este efecto. A pequeña escala por el contrario, esta situación hace que aumente la probabilidad de homocigosis con respecto a una población estándar (HWE).

Para modelizar la estratificación de la población, se crea un medidor del grado de relación intrapoblacional, el **coeficiente de coancestralidad** $\theta \in [0, 1]$. El objetivo de esta magnitud es indicar el grado de separación entre la subpoblación de interés y la población global a la que pertenece. Un valor de cero indica que no hay estratificación y se corresponde a una situación de equilibrio de Hardy-Weinberg (HWE)

Para implementar esta magnitud vamos a considerar que vamos obteniendo los alelos de los genotipos involucrados de forma secuencial. Es decir, dotamos a la secuencia genética de orden y denotamos por $j = 1, 2, ...n$ la posición que ocupa cada alelo.

Vamos a considerar que la probabilidad de obtener el primer alelo es la probabilidad poblacional p_i. A partir de aquí, para las posiciones $j \geq 2$. Si el alelo en la posición j es de tipo i, vamos a añadir un contador que tenga en cuenta la cantidad de alelos de tipo i observados en la secuencia hasta la posición $j - 1$ incluida. Vamos a penalizar la probabilidad de que aparezcan alelos distintos y a reforzar la probabilidad de homocigosis en función del valor de θ. Es decir, la probabilidad de obtener en la posición j un alelo de tipo i es:

$$p'_i = \frac{b_j\theta + (1-\theta)p_i}{1 + (j-2)\theta}, \qquad j \geq 2. \tag{3.39}$$

Como podemos ver, si hemos observado muchos alelos de un determinado tipo esta fórmula favorece la observación de otro más, por el contrario si se han observado pocos la penaliza. Cuanto mayor sea θ, mayor será este efecto. El origen de esta expresión se detallará en el Capítulo 5, sección 5.2.1.

Ejemplo 3.2. Recuperamos el Ejemplo 2 (sección 3.1.3), vamos a calcular la razón de verosimilitud nuevamente, esta vez considerando que los individuos involucrados pertenecen a una subpoblación con relación intrapoblacional dada por θ. Calculemos la expresión teórica.

Calculamos las probabilidades de los datos condicionados a cada hipótesis utilizando el teorema de las probabilidades totales como hicimos en (3.27):

$$Pr(datos \mid H_1) = Pr(G_{CH} \mid G_{AF}, G_{MO}, H_1)Pr(G_{AF} \mid G_{MO}, H_1)Pr(G_{MO} \mid H_1). \tag{3.40}$$

El cálculo de $Pr(G_{CH} \mid G_{AF}, G_{MO}, H_1) = 1$ ya lo hemos hecho y no se ve alterado. Las simplificaciones hechas para los demás factores también siguen vigentes de forma que conforme a (3.31): $Pr(datos \mid H_1) = Pr(G_{AF}) \cdot Pr(G_{MO})$ (para ver cuales son los cambios respecto a la situación de HWE ver la nota 3.3). Con respecto a la hipótesis de no paternidad H_2:

$$Pr(datos \mid H_2) = Pr(G_{CH} \mid G_{AF}, G_{MO}, H_2)Pr(G_{AF} \mid G_{MO}, H_2)Pr(G_{MO} \mid H_2). \tag{3.41}$$

Las simplificaciones de los últimos factores siguen vigentes, pero el cálculo que sí varía si consideramos la corrección theta es el de la probabilidad de $Pr(G_{CH} \mid G_{AF}, G_{MO}, H_2)$, veamos como se hace.

Recordemos que ahora estamos considerando los genotipos de forma secuencial, luego para este cálculo en particular contamos con la información sobre los genotipos de AF y la madre:

$$\text{Secuencia: } G_{AF}, G_{MO} = A/A/B/B \tag{3.42}$$

Estos son nuestros datos, con ellos debemos calcular la verosimilitud del genotipo del niño en el supuesto de que AF no es su padre (al no estar emparentados, bajo HWE la información G_{AF} no influiría en este cálculo, pero ahora importa cualquier observación que hayamos extraído de la población).

Por un lado, la maternidad no está discutida por lo que el niño ha de haber heredado un alelo de la madre. En este caso, como es homocigota, el alelo B del niño es el alelo materno con probabilidad 1. Por tanto:

$$Pr(G_{CH} \mid G_{AF}, G_{MO}, H_2) = 1 \cdot Pr(G_{CH} = A/B \mid G_{AF} = B/-, G_{AF}, G_{MO}, H_2) = p'_A. \tag{3.43}$$

La probabilidad de que el genotipo del niño sea el del enunciado es la probabilidad de que el verdadero padre (individuo aleatorio de la subpoblación considerada) le haya legado un alelo A. Esta probabilidad p'_A se calcula como la probabilidad de obtener un alelo A en la subpoblación de interés. Como no conocemos más información, esto es equivalente a secuenciar otro alelo A en nuestra subpoblación de interés:

$$\begin{aligned} \textit{Secuencia} :& A/A/B/B/\underline{A}, \\ & j = 5 \text{ y } b_j = 2. \end{aligned} \tag{3.44}$$

Calculamos entonces p'_A utilizando la corrección theta acorde a (3.39):

$$p'_A = \frac{2\theta + (1-\theta)p_A}{1+3\theta}. \tag{3.45}$$

Recapitulando, obtenemos finalmente:

$$LR = \frac{1 \cdot Pr(G_{AF}) \cdot Pr(G_{MO})}{\frac{2\theta+(1-\theta)p_A}{1+3\theta} \cdot Pr(G_{AF}) \cdot Pr(G_{MO})} = \frac{1+3\theta}{2\theta+(1-\theta)p_A}. \tag{3.46}$$

Nota 3.3. *Cabe observar que, aunque para resolver este problema no nos hace falta la expresión explícita de $Pr(G_{AF})$ o $Pr(G_{MO})$, estas probabilidades son diferentes de las que obtendríamos bajo el planteamiento inicial del problema, es decir, bajo HWE. Vamos a ilustrar esto, por ejemplo, sobre el genotipo de AF, $G_{AF} = A/A$:*

- *Bajo HWE: $Pr(A/A) = p_A p_A = {p_A}^2$.*
- *Con la corrección theta: tengamos en cuenta que ahora estamos considerando los alelos secuenciados en orden, y a partir de la posición 2, cada probabilidad se calcula teniendo en cuenta la información anterior:*

$$Pr(A/A) = Pr(\underline{A}/-) \cdot Pr(A/\underline{A}) = p_A \cdot \frac{1 \cdot \theta + (1-\theta)p_A}{1+(2-2)\theta} = \theta p_A + (1-\theta){p_A}^2. \tag{3.47}$$

3.2.3. Alelo silencioso

Para obtener un perfil genético a partir de una muestra se existen varias técnicas biológicas específicas, uno de los pasos que requieren es amplificar la muestra mediante la técnica PCR[4]. Cuando el marcador que tratamos de identificar es de tipo STR puede pasar que la amplificación falle en algunas zonas del gen omitiendo la amplificación de algunas copias. El verdadero alelo queda entonces enmascarado para los científicos en forma de otro alelo ya que estamos utilizando una técnicas incapaz de detectarlo. Como resultado obtenemos un perfil genético falso. Si además el individuo en cuestión posee otro ejemplar de este último, en los resultados aparecerá un solo alelo presente en la muestra, es decir, una falsa homocigosis.

Existe una corrección para esta última situación. Cuando ante una homocigosis se sospecha la existencia de un alelo silencioso S, el problema se modeliza incluyendo dicho alelo como una opción más para el marcador. Esto requiere reajustar las probabilidades alélicas, que recordemos deben sumar uno para que el modelo sea válido.

Ilustremos la situación suponiendo que hemos genotipado a un individuo como $B/-$. Entonces para calcular cualquier probabilidad con este genotipo debemos considerar:

$$Pr(B/-) = Pr(B/B) + Pr(B/S) = Pr(B/B) + Pr(\check{B}/\check{S}) + Pr(\check{S}/\check{B}). \tag{3.48}$$

Ejemplo 3.4. Planteamos una situación análoga al ejemplo 2 pero modificando ligeramente los genotipos. Vamos a suponer que hemos genotipado lo siguiente (Figura 3.4):

$$G_{AF} = B/-, \qquad G_{MO} = A/-, \qquad G_{CH} = A/- \,. \tag{3.49}$$

Ahora calculamos las verosimilitudes de ambas hipótesis.

$$\begin{aligned} &Pr(datos \mid H_1) = \\ &= Pr(G_{CH} = A/-, G_{AF} = B/-, G_{MO} = A/- \mid H_1) = \\ &= Pr(G_{CH} = A/S, G_{AF} = B/S, G_{MO} = A/S \mid H_1) \\ &+ Pr(G_{CH} = A/S, G_{AF} = B/S, G_{MO} = A/A \mid H_1) = \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot 2p_B p_S \cdot 2p_A p_S + 1 \cdot \frac{1}{2} \cdot 2p_B p_S \cdot {p_A}^2 = p_B p_S p_A (p_S + p_A), \end{aligned} \tag{3.50}$$

[4]Técnica biológica que consiste en desnaturalizar el ADN de una muestra y con ayuda de enzimas como la ADN-polimerasa copiar la información contenida en él. El objetivo es amplificar una muestra para tener más material sobre el que realizar análisis.

$$
\begin{aligned}
&Pr(datos|H_2) = \\
&= Pr(G_{CH} = A/-, G_{AF} = B/-, G_{MO} = A/-|H_2) = [Pr(G_{CH} = A/S|G_{AF} = B/-, G_{MO} = A/A, H_2) \\
&+ Pr(G_{CH} = A/S|G_{AF} = B/-, G_{MO} = A/A|H_1)] \cdot Pr(B/-) \cdot Pr(A/A) \\
&+ [Pr(G_{CH} = A/S|G_{AF} = B/-, G_{MO} = A/S, H_2) + Pr(G_{CH} = A/S|G_{AF} = B/-, G_{MO} = A/S|H_1)] \\
&\cdot Pr(B/-) \cdot Pr(A/S) = [1 \cdot p_S + 1 \cdot p_A] \cdot [{p_B}^2 + 2p_B p_S] \cdot {p_A}^2 \\
&+ \frac{1}{2} \cdot p_S + \frac{1}{2} \cdot p_A + \frac{1}{2} \cdot ({p_B}^2 + 2p_B p_S) \cdot 2p_A p_S = \\
&= [(p_A + p_S)^2(p_B + 2p_S) + p_A p_S(p_B + 2p_S)]p_A p_B.
\end{aligned}
\tag{3.51}
$$

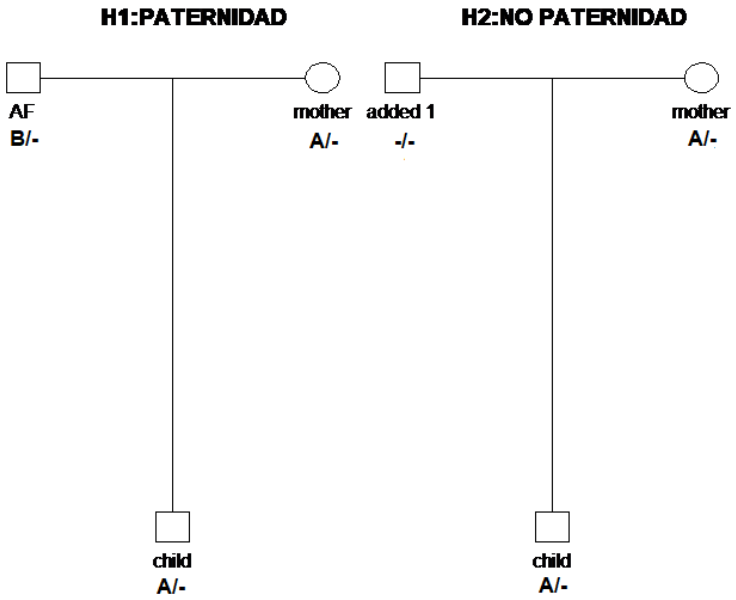

Figura 3.4: Ejemplo. Alelo silencioso.

Las probabilidades anteriores han sido obtenidas utilizando el teorema de las probabilidades totales, como ya empieza a ser habitual. Calculamos finalmente la razón de verosimilitudes y obtenemos la expresión simplificada:

$$
LR = \frac{p_S(p_S + p_A)}{(p_A + p_S)^2(p_B + 2p_S) + p_A p_S(p_B + 2p_S)}. \tag{3.52}
$$

Otra forma de asegurar que no estamos omitiendo un alelo silencioso es, si hay muestra suficiente y es posible, realizar el genotipado con varias técnicas diferentes para maximizar las posibilidades de que alguna de ellas sea sensible al alelo en cuestión.

3.2.4. Dropout

Se habla de dropout cuando falla la amplificación de un alelo. Como el otro sí se amplifica y no podemos distinguir el paterno del materno, el resultado se traducirá en nuestro perfil como homocigosis para ese marcador cuando la realidad puede no ser esta (el alelo que estamos dejando de detectar puede ser igual o distinto). Esta situación es muy común cuando la muestra es pequeña o está muy degradada, algo habitual en problemas de criminalística.

Para incluir esta posibilidad, el abordaje más sencillo pasa por fijar una probabilidad d de que un alelo no sea observado. Por ejemplo, apliquemos esta modificación para calcular la probabilidad de que, dado un marcador dialélico con variantes A, B, obtengamos para un individuo aleatorio el genotipo $B/-$:

$$\begin{aligned} Pr(B/-) = Pr(B/- \mid B/B)Pr(B/B) + Pr(B/- \mid \check{B}/\check{A})Pr(\check{B}/\check{A}) + Pr(B/- \mid \check{A}/\check{B})Pr(\check{A}/\check{B}) = \\ = (1-d^2){p_B}^2 + 2d(1-d)p_A p_B, \end{aligned} \tag{3.53}$$

si nos fijamos en cómo se calcula el primer sumando vemos que se debe tener en cuenta la posibilidad de que se expresen ambos alelos, solo el primero o solo el segundo ya que los tres casos son compatibles con un perfil genético $B/-$.

Contemplar o no la posibilidad de introducir los dropouts en el modelo depende de la calidad del ADN. En casos de determinación de parentesco no suele haber problemas pues las muestras suelen ser abundantes y tener una calidad aceptable, pero por ejemplo en el análisis genético en casos criminales se trata habitualmente con muestras degradadas o insuficientes. Por esto el desarrollo de técnicas que permitan trabajar con estas muestras están cobrando especial relevancia en los últimos años.

Ejemplo 3.5. Vamos a trabajar nuevamente sobre el ejemplo 2, de nuevo bajo las hipótesis planteadas en el apartado anterior (Figura 3.4), pero sin tener en cuenta el genotipo de la madre ($G_{MO} = -/-$). Además vamos a considerar la presencia de dropouts en los genotipos tanto del padre como del niño.

Para simplificar el problema, vamos también a suponer que el marcador considerado es dialélico, es decir, que solo puede presentar dos variantes: A y B. (De nuevo esto requeriría redefinir las frecuencias alélicas para que $p_A + p_B = 1$).

Vamos a indicar entonces como se calcularía la razón de verosimilitudes en este caso. Empezamos calculando el numerador

$$Pr(datos \mid H_1) = Pr(G_{CH} = A/-, G_{AF} = B/- \mid H_1), \tag{3.54}$$

donde nuevamente hemos de descomponer la probabilidad anterior utilizando la ley de probabilidades totales. Para obtener la suma de todas las posibilidades hay que contemplar las probabilidades recogidas en el Cuadro 3.1 donde las columnas se corresponden con los posibles genotipos de AF

	B/B	$\check{A}/\check{B}$	$\check{B}/\check{A}$
A/A	0	$\frac{1}{2}p_A$	$\frac{1}{2}p_A$
$\check{A}/\check{B}$	0	$\frac{1}{2}p_B$	$\frac{1}{2}p_B$
$\check{B}/\check{A}$	p_A	$\frac{1}{2}p_A$	$\frac{1}{2}p_A$

Cuadro 3.1: Las columnas indican los posibles genotipos de AF, las filas lo posibles genotipos de CH y cada elemento representa la probabilidad de cada G_{CH} condicionada al respectivo G_{AF}

y las filas con los del niño CH y cada elemento de la tabla es la probabilidad del genotipo del hijo condicionada al genotipo del padre. Obtenemos:

$$Pr(datos \mid H_1) = 2p_A p_B(p_B + p_A) + 2d{p_A}^2 p_B(1-d) - d^2 p_A {p_B}^2. \tag{3.55}$$

Por otra parte, el cálculo del genotipo del niño bajo la hipótesis de no paternidad, como no tenemos otra información a mayores, es la probabilidad de obtener el genotipo $A/-$ de forma aleatoria en la población. Este cálculo ya lo hemos hecho en (3.53) y resulta:

$$Pr(datos \mid H_2) = Pr(G_{CH} = A/- \mid H2) = Pr(A/-) = (1-d^2){p_A}^2 + 2d(1-d)p_B p_A, \tag{3.56}$$

con lo que finalmente simplemente dividiendo obtendríamos:

$$LR = \frac{2p_B(p_B + p_A) + 2dp_A p_B(1-d) - d^2 {p_B}^2}{(1-d^2)p_A + 2d(1-d)p_B}. \tag{3.57}$$

3.2.5. Relaciones de parentesco complejas

En esta sección simplemente vamos a presentar algunas situaciones basadas en los planteamientos anteriores pero que requieren de cálculos más complejos. En la sección siguiente presentaremos ejemplos que incluyen algunas de estas situaciones y los resolveremos utilizando el paquete Familias de R.

No obstante comenzaremos viendo un recurso utilizado en multitud de ocasiones para extraer información de otras relaciones que no sean de paternidad.

Cálculos IBD

Hasta ahora solo se ha utilizado para extraer información sobre el genotipo de una persona la información genética de sus ascendentes directos. Cuando se desconoce el genotipo de un progenitor, procediendo como venimos haciendo, la única opción es sumar sobre todas las opciones para dicho genotipo. Sin embargo, se puede extraer información sobre las probabilidades de genotipos desconocidos a partir de otros tipos de parentesco no directo, por ejemplo de la relación de hermanos.

La idea fundamental que subyace bajo este tipo de razonamientos es la existencia de alelos **idénticos por descendencia (IBD)**[5]. Un alelo de un individuo es IBD a otro alelo de un

[5]Del inglés *Identical By Descent* (idénticos por descendencia).

individuo diferente si ambos proceden del mismo alelo ancestral común. Para considerar que dos alelos son IBD no basta que sean iguales en ambos individuos y que estos individuos estén emparentados, pues podríamos estar ante una situación de alelos idénticos por estado (IBS)[6], es decir, que son iguales pero no han sido heredados del mismo antepasado.

Para visualizar la diferencia supongamos por un momento que dotamos de sentido al orden de los alelos de cada individuo, de forma que si el genotipo para un marcador es $G = (G_1, G_2)$, G_1 es el alelo paterno (heredado del padre) y G_2 es el alelo materno (heredado de la madre). En estas circunstancias podemos ver la diferencia entre alelos IBD y alelos IBS en la Figura 3.5:

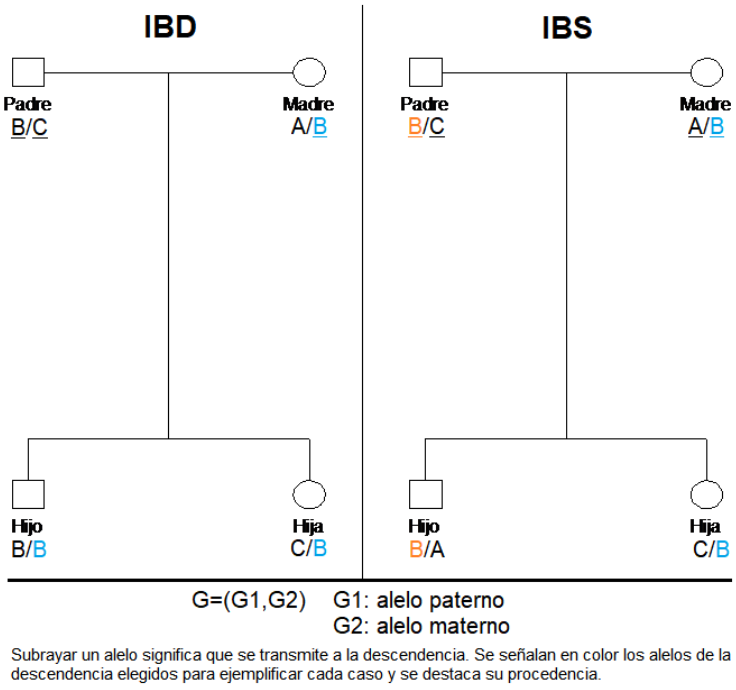

Figura 3.5: Ejemplo IBD vs IBS

Ahora bien, en la práctica con las técnicas de secuenciación existentes solo podemos determinar el tipo de alelo que presenta un individuo, no su origen (materno o paterno). Así, los casos en que dos hermanos poseen un alelo que solo se presenta en uno de los progenitores en heterocigosis, podemos saber que sus alelos son IBD, pero en el resto de casos no hay manera de determinarlo.

Por tanto, a la hora de extraer información de una relación fraternal, debemos considerar todas las posibilidades:

- **IBD=0:**
  Si un hermano ha heredado los alelos (G_1, G_2) (alguna de las cuatro combinaciones posibles de alelos materno y paterno), el otro hermano no ha heredado ninguno de ellos. Sumando

[6]Del inglés *Identical By State* (idénticos por estado).

sobre todas las posibilidades:

$$\begin{aligned} &Pr(IBD = 0) = \\ &= 4 \cdot Pr(\text{un hermano hereda } G_1 \text{ y } G_2) \cdot Pr(\text{el otro hermano no hereda } G_1 \text{ ni } G_2) = \\ &= 4 \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{4}. \end{aligned}$$

- **IBD=1:**
  Si un hermano ha heredado los alelos (G_1, G_2) (alguna de las cuatro combinaciones posibles de alelos materno y paterno), el otro hermano ha heredado uno en común con los anteriores. Sumando sobre todas las posibilidades:

$$\begin{aligned} &Pr(IBD = 1) = \\ &= 4Pr(\text{un hermano hereda } G_1 \text{ y } G_2) \cdot \left[Pr(\text{comparten solo } G_1) + Pr(\text{comparten solo } G_2)\right] = \\ &= 4 \cdot \frac{1}{4} \cdot \left[\frac{1}{4} + \frac{1}{4}\right] = \frac{1}{2}. \end{aligned}$$

- **IBD=2:**
  Si un hermano ha heredado los alelos (G_1, G_2) (alguna de las cuatro combinaciones posibles de alelos materno y paterno), el otro hermano hereda los mismos. Es decir:

$$\begin{aligned} &Pr(IBD = 2) = \\ &= 4 \cdot Pr(\text{un hermano hereda } G_1 \text{ y } G_2) \cdot Pr(\text{el otro hermano hereda } G_1 \text{ y } G_2) = \\ &= 4 \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{4}. \end{aligned}$$

Si se posee información sobre el genotipo de ambos progenitores, está claro que la información genética que puede proporcionar el ADN de un hermano es redundante (excluyendo casos raros como existencia de mutaciones o errores de genotipado). Sin embargo, cuando uno de los progenitores o ambos son desconocidos, esta información nos permite calcular la probabilidad de un genotipo determinado para un individuo de interés. Así, si H es la hipótesis que incluye la relación de hermanos:

$$\begin{aligned} &Pr(datos \mid H) = Pr(datos \mid IBD = 0) \cdot Pr(IBD = 0) \\ &+ Pr(datos \mid IBD = 1) \cdot Pr(IBD = 1) + Pr(datos \mid IBD = 2) \cdot Pr(IBD = 2) = \\ &= \frac{1}{4}Pr(datos \mid IBD = 0) + \frac{1}{2}Pr(datos \mid IBD = 1) + \frac{1}{4}Pr(datos \mid IBD = 2). \end{aligned} \tag{3.58}$$

Ejemplo 3.6. Este ejemplo se verá en detalle en el capítulo siguiente, en el que utilizaremos R para resolverlo. De momento basta decir que se han hecho pruebas y determinado los genotipos de una madre (AM) y su hija (AS) para un marcador específico: $G_{AM} = 6/8$ y $G_{AS} = 6/9$. Queremos comprobar la verosimilitud de los datos bajo la hipótesis de que unos restos humanos de genotipo $G_{Hueso} = 6/9$ pertenecen a la otra hija de AM, hermana completa de AS.

Como no estamos contemplando la posibilidad de mutaciones, la única opción compatible con esta hipótesis es que tanto AS como la hermana a la que pertenecen los restos hayan heredado el único alelo 6 de la madre. Es decir, los alelos 6 de ambas hermanas son IBD, por lo tanto:

$$Pr(datos \mid IBD = 0) = 0, \tag{3.59}$$

$$\begin{aligned} &Pr(datos \mid IBD = 1) = Pr(G_{AM}, G_{AS}, G_{Hueso} \mid IBD = 1) = \\ &= Pr(G_{Hueso} \mid G_{Hueso} = 6/-, G_{AS}, IBD = 1) = Pr(G_{AF} = 9/9) = p_9^2, \end{aligned} \tag{3.60}$$

$$\begin{aligned} &Pr(datos \mid IBD = 2) = Pr(G_{AM}, G_{AS}, G_{Hueso} \mid IBD = 1) = \\ &= Pr(G_{Hueso} \mid G_{Hueso} = 6/-, G_{AS}, IBD = 2) = Pr(G_{AF} = 9/-) = p_9. \end{aligned} \tag{3.61}$$

Ahora sustituyendo (3.59), (3.60) y (3.61) en la expresión (3.58) obtenemos el resultado buscado:

$$Pr(datos \mid H) = \frac{1}{4} \cdot 0 + \frac{1}{2} \cdot p_9^2 + \frac{1}{4} \cdot p_9 = \frac{1}{4} p_9(2p_9 + 1). \tag{3.62}$$

Resulta inmediato entender por qué este recurso resulta tan ventajoso a la hora de realizar los cálculos. En este ejemplo, de no aplicar este método habría que recorrer todos los posibles genotipos del padre y sumar sobre ellos para obtener el resultado buscado. De esta forma los cálculos sencillos y directos. El paquete Familias, al igual que la mayoría de programas pensados para usar en laboratorio, lleva implementado este recurso.

Relaciones complejas

Empezamos considerando el caso del **parentesco lejano**. Bajo esta denominación se engloba cualquier posible escenario que involucre más de una generación, desde un problemad de determinación de parentesco entre primos o abuelos y nietos, hasta estudios genéticos sobre la evolución que se remontan cientos de generaciones atrás (para estos casos se utilizan otro tipo de marcadores y técnicas específicas).

Veremos algún ejemplo en la sección siguiente de problemas de parentesco que involucran a dos generaciones.

Otro contexto en el que la complejidad de los cálculos se puede ver incrementada es el de los problemas de **múltiples hipótesis**. Hasta ahora el planteamiento que hemos propuesto para los problemas de determinación de parentesco pasa por considerar dos hipótesis, de las que una es la hipótesis de no relación. Sin embargo cabe plantearse qué sucede si sospechamos de la existencia de parentesco pero no sabemos precisar de qué tipo: por ejemplo, barajamos la posibilidad de que dos individuos sean padre e hijo (Standard Duo Case) o hermanos.

Un posible enfoque es plantear una hipótesis por cada posible relación que se baraja. Añadimos la hipótesis de no relación y planteamos el contraste.

Así, dados dos individuos con perfiles G_1 y G_2, consideramos que pueden estar emparentados a través de una serie de relaciones. Si denotamos por $R = \{r_1, r_2, ..., r_{s-1}\}$ al conjunto de dichas

relaciones el problema se plantearía como sigue:

$$\begin{cases} H_1: \ G_1 \text{ y } G_2 \text{ están relacionados a través de } r_1. \\ \\ H_2: \ G_1 \text{ y } G_2 \text{ están relacionados a través de } r_2. \\ \\ \vdots \\ \\ H_s: \ G_1 \text{ y } G_2 \text{ no están emparentados.} \end{cases}$$

Sin embargo, hasta ahora se ha utilizado la razón de verosimilitudes LR para extraer conclusiones, y el cálculo de dicha magnitud involucra exclusivamente dos hipótesis. Lo que haremos será entonces escoger una de las hipótesis y calcular el LR de cada una de las hipótesis restantes escalando frente a la escogida.

Lo habitual es contrastar frente a la hipótesis de no efecto, es decir, tomando $i \in \{1, 2, ..., s\}$

$$LR_{i,s} = \frac{Pr(datos \mid H_i)}{Pr(datos \mid H_s)}. \tag{3.63}$$

Ahora considerando el teorema de Bayes en su versión general (3.9), y asumiendo que todas las hipótesis son igualmente probables a priori obtenemos la expresión de la probabilidad a posteriori:

$$Pr(H_i \mid datos) = \frac{LR_{i,s} Pr(H_i)}{\sum_{j=1}^{s} LR_{j,s} Pr(H_j)} = \frac{LR_{i,s}}{\sum_{j=1}^{s} LR_{j,s}}. \tag{3.64}$$

Si se presenta esta situación es importante señalar siempre qué hipótesis hemos escogido para normalizar las razones de verosimilitud.

Otra opción es trabajar directamente con las probabilidades a posteriori, la ventaja es que así ya no hay denominadores y no hay peligro de confusión con respecto a qué hipótesis estamos comparando. La desventaja vuelve a ser que se nos exige fijar probabilidades a priori para cada hipótesis.

Por último, un tercer factor que incrementa la complejidad de un problema es la existencia de **endogamia**, es decir, de la existencia de progenie entre individuos ya emparentados en las generaciones previas. La dificultad radica en este caso en calcular las probabilidades, pues aunque el razonamiento no cambia las relaciones entre los distintos genotipo requieren cálculos cuidadosos.

Capítulo 4

Cálculos con R

A lo largo del presente capítulo vamos a examinar el paquete de R Familias. Este paquete provee al investigador de algunas herramientas sencillas para realizar los cálculos anteriores utilizando el software R. El paquete ha sido creado por Petter Mostad, Thore Egeland e Ivar Simonsson y tiene un fin académico. Vamos a presentar las funciones que contiene y a ver como se aplicarían para realizar los cálculos (si se desea tener toda la información se puede consultar el manual [6] o la fuente principal [7] en la que se implementan algunos ejemplos).

El paquete Familias es gratuito y de uso libre, por lo que se puede descargar directamente de la red. Algunas de sus funcionalidades requieren el uso de funciones de otros paquetes más antiguos relacionados con la determinación de parentesco (entre los que destaca el kinship2 [1]). Por ello para lograr el pleno rendimiento de Familias es necesario descargar otros paquetes complementarios. Para descargarlo directamente de la red, basta con introducir la siguiente línea de código en el terminal y seleccionar uno de los repositorios de R.

```
#Paquetes complementarios
install.packages(c("kinship2","Matrix","quadprog","paramlink","Rsolnp"))
#Paquete principal
install.packages("Familias")
```

Para utilizarlo basta simplemente con invocarlo mediante el siguiente comando:

```
library(Familias)
```

Nota 4.1. *En la práctica los laboratorios no utilizan este paquete (creado específicamente para ilustrar los ejemplos de [7]). En su lugar se utilizan el conjunto de paquetes* ***ped suite****: pedtools, pedprobr, ribd, y forrel (fuente [4]). Para ver más se remite al lector a los respectivos manuales de cada uno de los paquetes: [17], [16],[18] y [15].*

4.1. Standard Duo Case

Vamos a empezar trabajando nuevamente con el ejemplo base: Ejemplo 1. Standard Duo Case (sección 3.1.2). Para cada función utilizada introduciremos un breve explicación sobre

su funcionamiento y sintaxis e ilustraremos su implementación sobre este ejemplo. Finalmente calcularemos la razón de verosimilitud (LR) con R y podremos comprobar que en efecto coincide con el resultado que hemos obtenido manualmente.

4.1.1. Introducir la genealogía: FamiliasPedigree

Lo primero que vamos a necesitar es introducir los datos del problema de una forma comprensible para las funciones encargadas de trabajar con ellos. Para ello vamos a crear un objeto capaz de almacenar la información sobre las **relaciones** entre los individuos involucrados.

```
#Individuos involucrados en nuestro caso: supuesto padre (AF) y el hijo.
```

Es evidente, en vista del planteamiento habitual de los problemas de contraste de parentesco que vamos a necesitar tantos objetos como hipótesis hayamos planteado, pues en cada una de ellas estamos trabajando con una genealogía diferente (recordamos la Figura 3.1).

La función encargada de construir estos objetos es **FamiliasPedigree**:

```
#Comenzaremos con la genealogia bajo H1: Paternidad
FamiliasPedigree(id, dadid, momid, sex)
```

Los argumentos anteriores son:

- **id:** vector con los identificadores de los individuos involucrados.

  ```
  id<-c("hijo","AF")
  ```

- **dadid:** vector en el que se indica qué individuos tienen a su padre biológico entre los individuos involucrados. Para indicarlo, en la posición correspondiente a cada uno se introduce o bien el valor NA, o bien el identificador del padre biológico.

  ```
  dadid<-c("AF",NA)
  ```

- **momid:** vector análogo al anterior pero con información sobre las madres biológicas.

  ```
  momid<-c(NA,NA) #La madre no esta genotipada asi que no la incluimos
  ```

- **sex:** vector en el que se introduce la información sobre el sexo genético[1] de cada uno de los individuos involucrados. Los únicos valores que se pueden introducir son "male"(varón) y "female"(mujer).

  ```
  sex<-c("male","male")
  ```

Finalmente creamos el primer perfil genalógico para nuestros datos y examinamos el objeto resultante:

[1] Esta información, pese a ser útil en muchos casos, es más bien inútil para marcadores autosómicos.

```
> ped1<-FamiliasPedigree(id,dadid,momid,sex)
> ped1
$id
[1] "Hijo" "AF"

$findex
[1] 2 0

$mindex
[1] 0 0

$sex
[1] "male" "male"

attr(,"class")
[1] "FamiliasPedigree"
```

Como podemos ver en la última línea, el objeto creado es una lista de la clase FamiliasPedigree que contiene básicamente la información que hemos introducido: el vector de identificadores, los vectores de índices indicando las relaciones de maternidad y paternidad y el sexo.

Procedemos análogamente para crear el linaje bajo cada posible hipótesis.

```
#Bajo H2:No paternidad
ped2<-FamiliasPedigree(id,dadid=c(NA,NA),momid=c(NA,NA),sex)
```

Por último, creamos una lista donde asignamos a cada genealogía un nombre representativo (esto facilitará la interpretación de la salida de los datos). Enseguida estudiaremos el comando encargado de calcular la razón de verosimilitudes, pero es importante saber que por defecto coloca en el denominador (recordar expresión (3.9)) la verosimilitud bajo la primera hipótesis de la lista. Por supuesto podemos personalizar este argumento, pero una solución sencilla es ser cuidadosos y colocar siempre la hipótesis frente a la que queremos normalizar verosimilitudes en primera posición.

```
#Calculamos los LR frente a la hipotesis de no parentesco
genealogias<-list(noPadre=ped2, Padre=ped1)
```

Como ya hemos mencionado, muchas de las funcionalidades implementadas por Familias aprovechan el trabajo de otros paquetes previos como kinship2 o paramlink. En particular, esta función recicla en gran medida el funcionamiento de una función previa de kinship2, pero la simplifica (elimina marcadores para enfermedades, que no son útiles en el estudio del parentesco pues recordemos que no se pueden utilizar marcadores con manifestaciones fenotípicas) y la generaliza (permite la introducción de individuos con un solo progenitor en la genealogía como es nuestro caso).

Gráficos: árbol genealógico

Un árbol genealógico es una representación gráfica de las relaciones entre varios individuos. Mediante simbología sencilla (fijada por convenio, ver Figura 3.1), permite resumir de forma compacta el parentesco a lo largo de generaciones. En nuestro caso, es particularmente útil a la hora de entender las distintas hipótesis que entran en juego en cada contraste.

R nos permite representar directamente los linajes que hemos introducido utilizando el comando plot:

```
plot(ped1)
#Representamos linajes bajo las dos hipotesis
par(mfrow=c(1,2))
plot(ped1);plot(ped2)
```

Como resultado obtendremos precisamente la Figura 3.1 (sin información sobre los genotipos, dicha información la hemos incluido sobreescribiendo la imagen). Hemos estado usando desde el principio este tipo de representaciones pues su interpretación es muy intuitiva y condensa los datos de partida de cualquier problema. El paquete pedtools [17] amplia estos recursos gráficos de forma que se puede incluir directamente la información sobre los genotipos entre otras funciones más enfocadas a la genética médica).

4.1.2. Definir marcadores: FamiliasLocus

Lo siguiente que debemos hacer es definir nuestros marcadores y la información que conocemos sobre ellos. Como nuestro objetivo es puramente matemático, no nos interesa la información médica o biológica de los mismos, sino aquella que podemos utilizar a la hora de obtener información sobre la plausibilidad de las hipótesis: los posibles alelos y sus frecuencias.

La introducción de estos datos puede darse de dos formas: bien manualmente utilizando la función FamiliasLocus, o bien a partir de una base de datos (como es nuestro caso en el ejemplo de referencia).

Manualmente: FamiliasLocus

Para definir un marcador introduciendo directamente los datos se utiliza la función **FamiliasLocus**. Los argumentos básicos para definir un marcador sin mutaciones son los siguientes:

```
marcador<-FamiliasLocus(frequencies,allelenames,name)
```

- **frequencies:** vector con las frecuencias de cada alelo considerado. Debemos tener en cuenta que estas frecuencias han de sumar uno, por lo que si estos no se cumple para las frecuencias de los alelos considerados la mejor solución es crear un nuevo alelo artificial que acumule las frecuencias que faltan.
- **allelenames:** vector con los nombre de los alelos.

- **name:** Nombre del marcador, si no se indica se toma el nombre del vector de frecuencias.
- **modelo con mutaciones:** aunque no vamos a centrarnos en este aspecto, la función FamiliasLocus permite incluir multitud de argumentos para construir modelos de mutación complejos. Hay dos argumentos básicos, por un lado MutationModel, que indica como construir la matriz de mutaciones, bien introduciéndola directamente (Custom) o bien teniendo en cuenta las frecuencias de los alelos y otros datos como la forma del modelo o las tasas de mutación (serían necesarios más argumentos). El otro parámetro es Stabilization, que indican como funciona el modelo cuando se acumulan mutaciones sobre un mismo alelo).

Como ejemplo vamos a implementar los marcadores de nuestro problema de esta forma. Los datos de las frecuencias de cada alelo involucrado son (recogidos de [3]) son $p_17 = 0{,}2040$, $p_18 = 0{,}1394$ y $p_8 = 0{,}5539$. Empezamos con el primer marcador, crearemos un tercer alelo artificial "A" para poder definir nuestro vector de frecuencias:

```
#Marcador D3S1358
> frecuencias<-c(0.2040,0.1394,1-(0.2040+0.1394))
> nombrealelos<-c(17,18,"A")
> M1<-FamiliasLocus(frecuencias,nombrealelos,"D3S1358")
> M1
$locusname
[1] "D3S1358"

$alleles
    17     18      A
0.2040 0.1394 0.6566

$femaleMutationType
[1] "No␣mutations"

$femaleMutationMatrix
   17 18 A
17  1  0 0
18  0  1 0
A   0  0 1

$maleMutationType
[1] "No␣mutations"

$maleMutationMatrix
   17 18 A
17  1  0 0
18  0  1 0
A   0  0 1
```

```
$simpleMutationMatrices
[1] TRUE

$Stabilization
[1] "NONE"

attr(,"class")
[1] "FamiliasLocus"
```

El objeto creado es de tipo FamiliasLocus y condensa la información sobre el marcador D3S1358: nos proporciona su nombre, una matriz con los alelos y sus frecuencias y varios argumentos sobre como se modelizan las mutaciones (en este caso estamos considerando modelo sin mutaciones, por lo que la matriz de mutaciones es siempre la identidad). Le hemos llamado $M1$ para poder manipular el objeto con más comodidad.

Análogamente construimos todos los marcadores involucrados en nuestro problema y los agrupamos en una lista:

```
> frecuencias2<-c(0.5539,1-0.5539)
> nombrealelos2<-c(8,"A")
> M2<-FamiliasLocus(frecuencias2,nombrealelos2,"TPOX")
> marcadores<-list(M1,M2)
```

Base de datos

Cuando contamos con una base de datos debidamente estructurada en un formato legible para R, podemos cargar directamente los marcadores ya constituidos. Por ejemplo, en nuestro caso los datos están recogidos y ordenados en la base de datos NorwegianFrequencies. Esta base de datos viene cargada con el paquete Familias, por lo que es incluso más sencillo, pero se podría leer información de forma análoga de cualquier base de datos en formato lista.

```
> class(NorwegianFrequencies)
[1] "list"
> names(NorwegianFrequencies)
 [1] "D3S1358"  "TH01"     "D21S11"   "D18S51"   "PENTA_E"  "D5S818"
 [7] "D13S317"  "D7S820"   "D16S539"  "CSF1PO"   "PENTA_D"  "VWA"
[13] "D8S1179"  "TPOX"     "FGA"      "D19S433"  "D2S1338"  "D10S1248"
[19] "D1S1656"  "D22S1045" "D2S441"   "D12S391"  "SE33"     "D7S1517"
[25] "D3S1744"  "D2S1360"  "D6S474"   "D4S2366"  "D8S1132"  "D5S2500"
[31] "D21S2055" "D10S2325" "D17S906"  "APOAI1"   "D11S554"
> class(NorwegianFrequencies$FGA)
[1] "numeric"
```

El comando *names* nos permite ver los marcadores contenidos en esta base de datos. Para cada uno de ellos se incluyen únicamente sus alelos y las frecuencias. En este caso además los

objetos de la base de datos son de tipo numérico, pero si fuesen objetos ya creados de tipo FamiliasLocus, bastaría introducirlos como argumento y el objeto creado heredaría todas las propiedades del antiguo a no ser que se especifique lo contrario (incluyendo el nombre).

Para nuestro ejemplo vamos a cargar los datos directamente desde la base, de esta forma tendremos información sobre todos los posibles alelos de cada marcador, lo cual es muy útil por si queremos ampliar el modelo en algún momento. Una vez creados, de nuevo lo más operativo cuando hay varios marcadores es crear una lista con todos ellos.

```
> M1<-FamiliasLocus(NorwegianFrequencies$D3S1358)
> M2<-FamiliasLocus(NorwegianFrequencies$D3S1358)
> marcadores<-list(M1,M2)
```

4.1.3. Incluir genotipos

Ahora que tenemos el esquema de las relaciones y los marcadores definidos, debemos introducir los genotipos de cada individuo para cada marcador, que son al final los datos que tenemos para realizar nuestro contraste. Para ello simplemente vamos a construir una matriz de datos donde cada fila corresponda a un individuo y cada columna a un alelo (es decir, cada par de columnas se corresponden con un marcador).

```
> AF<-c(17,18,8,8)
> hijo<-c(17,17,8,8)
> matrizdatos<-rbind(AF,hijo)
> matrizdatos
     [,1] [,2] [,3] [,4]
AF     17   18    8    8
hijo   17   17    8    8
```

Es importante denotar el perfil correspondiente a cada individuo por el mismo identificador que hemos usado para construir las relaciones.

4.1.4. Cálculos: FamiliasPosterior

Utilizando los tres elementos que hemos construido, la lista de genealogías, la lista de marcadores y la matriz de genotipos, podemos invocar a la función **FamiliasPosterior**, que se encargará de realizar los cálculos necesarios para devolvernos la probabilidad a posteriori, la razón de verosimilitudes (global y por marcador) y la verosimilitud de cada hipótesis a la vista de los datos.

```
FamiliasPosterior(pedigrees,loci,datosmatrix,prior,ref=1,kinship=0,
simplifyMutations=FALSE)
```

- **pedigrees:** lista de genealogías (introducida en el apartado 4.1.1).
- **loci:** lista de marcadores (introducida en el apartado 4.1.2).

- **datosmatrix:** matriz con los genotipos de cada individuo para todos los marcadores (introducida en el apartado 4.1.3).
- **prior:** vector de probabilidades a priori de cada genealogía. Por defecto si no se indica se asume que todas las hipótesis son igualmente probables. Existe una función específica del paquete llamada **FamiliasPrior** que a partir de una lista de genealogías asigna probabilidades a priori para cada hipótesis. Permite introducir parámetros[2] de endogamia, promiscuidad, penalizar los esquemas de parentesco muy complejos (que involucren muchas generaciones) e incluso definir un número máximo de generaciones a considerar.
- **ref:** dentro de la lista de genealogías, indicador de la hipótesis que se utiliza de referencia para calcular las razones de verosimilitud. Por defecto es el primer perfil que ocupa la primera posición en el vector.
- **kinship:** valor del coeficiente de coancestralidad (ver sección 3.2.2). Por defecto se asume equilibrio de Hardy-Weinberg ($\theta = 0$).
- **simplifyMutations:** recurso para acelerar los cálculos en algunos casos con mutaciones a lo largo de varias generaciones.

Veámoslo en nuestro caso particular:

```
> resultado<-FamiliasPosterior(genealogias,marcadores,matrizdatos)
> resultado
$posterior
  noPadre     Padre
0.1843711 0.8156289

$prior
noPadre   Padre
    0.5     0.5

$LR
 noPadre    Padre
1.000000 4.423845

$LRperMarker
                              noPadre    Padre
NorwegianFrequencies$D3S1358        1 2.450403
NorwegianFrequencies$TPOX           1 1.805354

$likelihoods
     noPadre         Padre
0.0002229192 0.0009861600
```

[2]Como no será utilizado, remitimos al lector a [6] o a [5] para conocer estos argumentos en detalle.

```
$likelihoodsPerSystem
                                    noPadre       Padre
NorwegianFrequencies$D3S1358 0.002368082 0.005802755
NorwegianFrequencies$TPOX    0.094134933 0.169946848
```

El resultado es una lista que incluye los resultados que estábamos buscando. Si comparamos los resultados podemos ver que son análogos a los que hemos obtenido manualmente en la sección 3.1.2, en particular:

Prior y **Posterior** nos dan respectivamente las probabilidades a priori asumidas para cada hipótesis y las probabilidades a posteriori (vemos que para la hipótesis de paternidad, el valor coincide con el índice de Essen-Móller (3.25). Por su parte, **LR** es la razón de verosimilitud global y **LRperMarker** la de cada marcador, utilizando siempre como referencia la hipótesis de no relación. Como vemos coinciden con los resultados calculados manualmente: (3.15) y (3.20) para cada marcador y (3.22) la razón de verosimilitud global. Por último, **likelihoodsPerSystem** incluye las verosimilitudes de la hipótesis de paternidad (columna *Padre*) y no paternidad (columna *noPadre*) para cada uno de los marcadores por separado, que como vemos coinciden con los resultados (3.16) y (3.21). Las verosimilitudes globales de ambas hipótesis para los marcadores combinados nos las da la entrada **likelihoods** y coincide con el resultado obtenido en (3.24).

4.2. Ejemplos

En esta sección vamos a aprovechar el paquete Familias para presentar ejemplos un poco más complejos que requieren cálculos más laboriosos. En particular, se presentan ejemplos de algunas situaciones expuestas en el Capítulo 3.

4.2.1. IBD

Comenzamos con un caso relativo al ámbito de la genética forense al que se tuvo que enfrentar The College of American Pathologists en 2011 cuando fueron encontrados restos humanos en un bosque. La localización y el desgaste indicaban que los restos podían ser de una chica desaparecida cuya desaparición había sido denunciada. El objetivo era determinar si en efecto los restos ($Hueso$) pertenecían a esta persona desaparecida o si eran de un individuo desconocido. Los datos de referencia de los que se dispuso fueron los genotipos de la supuesta madre (AM) y su otra hija, la supuesta hermana (AS) tanto por parte de madre como de padre.

$$\begin{cases} H_1: \text{ Los restos pertenecen a la hija de } AM \text{ y hermana completa de } AS. \\ H_2: \text{ Los restos pertenecen a un individuo desconocido (no relacionado con } AM \text{ y } AS). \end{cases}$$

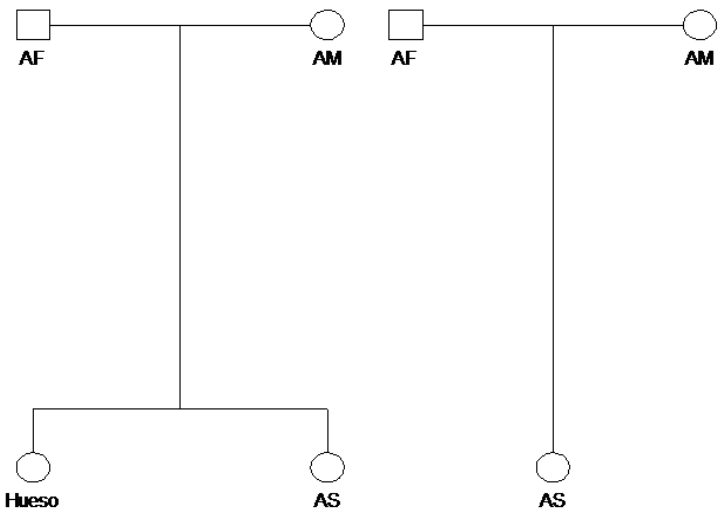

Figura 4.1: Ejemplo IBD (gráfico dibujado con el paquete Familias)

Los datos que vamos a estudiar son datos relativos al marcador $F13B$. Los genotipos obtenidos en las pruebas genéticas de los individuos involucrados son: $G_{AM} = 6/8$, $G_{AS} = 6/9$ y $G_{Hueso} = 6/9$. No vamos a considerar ningún artefacto (tales como dropouts, mutaciones o presencia de alelos silenciosos). Las frecuencias de los alelos involucrados son $p_6 = 0{,}086$, $p_8 = 0{,}152$ y $p_9 = 0{,}328$

```
> #IBD ejemplo: Hermana desaparecida

> p<-c("AM","Hueso","AS","AF")
> s<-c("female","female","female","male")

> ped1<-FamiliasPedigree(id=p,dadid=c(NA,"AF","AF",NA),
momid=c(NA,"AM","AM",NA),sex=s)
> ped2<-FamiliasPedigree(id=p,dadid=c(NA,NA,"AF",NA),
momid=c(NA,NA,"AM",NA),sex=s)

> par(mfrow=c(1,2))
> plot(ped1);plot(ped2)
Did not plot the following people: Hueso
> #La salida del comando es la Figura \ref{fig:ejIBD}

> genealogias<-list(notSister=ped2,isSister=ped1)
> genealogias
$notSister
$id
[1] "AM"    "Hueso" "AS"    "AF"

$findex
```

```
[1] 0 0 4 0

$mindex
[1] 0 0 1 0

$sex
[1] "female" "female" "female" "male"

attr(,"class")
[1] "FamiliasPedigree"

$isSister
$id
[1] "AM"    "Hueso" "AS"    "AF"

$findex
[1] 0 4 4 0

$mindex
[1] 0 1 1 0

$sex
[1] "female" "female" "female" "male"

attr(,"class")
[1] "FamiliasPedigree"

> #Marcador: alelos 6,7 y 8
> frecuencias<-c(0.086,0.152,0.328,1-(0.086+0.152+0.328))
> alelos<-c(6,8,9,"R") #R:alelo sobrante
> marcador<-FamiliasLocus(frecuencias,alelos)
> marcador
$locusname
[1] "frecuencias"

$alleles
    6     8     9     R
0.086 0.152 0.328 0.434

$femaleMutationType
[1] "No mutations"

$femaleMutationMatrix
  6 8 9 R
6 1 0 0 0
```

```
8 0 1 0 0
9 0 0 1 0
R 0 0 0 1

$maleMutationType
[1] "No␣mutations"

$maleMutationMatrix
  6 8 9 R
6 1 0 0 0
8 0 1 0 0
9 0 0 1 0
R 0 0 0 1

$simpleMutationMatrices
[1] TRUE

$Stabilization
[1] "NONE"

attr(,"class")
[1] "FamiliasLocus"

> #Genotipos:
> AM<-c(6,8)
> Hueso<-c(6,9)
> AS<-c(6,9)
> matrizdatos<-rbind(AM,Hueso,AS)
> matrizdatos
      [,1] [,2]
AM       6    8
Hueso    6    9
AS       6    9

> #Calculos verosimilitud y probabilidades:
> resultado<-FamiliasPosterior(genealogias,marcador,matrizdatos)
> resultado
$posterior
notSister  isSister
0.1452463 0.8547537

$prior
notSister  isSister
      0.5       0.5
```

```
$LR
notSister  isSister
 1.000000  5.884855

$LRperMarker
            notSister isSister
frecuencias         1 5.884855

$likelihoods
   notSister     isSister
0.0002418901 0.0014234885

$likelihoodsPerSystem
               notSister    isSister
frecuencias 0.0002418901 0.001423489
```

Por lo tanto los datos que hemos obtenido son 5.8849 veces más probables si los restos pertenecen a la familia considerada (AM y AS) que si pertenecen a un individuo desconocido no relacionado con estas personas.

4.2.2. IBD, parentesco complejo y endogamia

En este ejemplo, de nuevo del ámbito de la criminalística, vamos a resolver un problema que presenta a la vez una posible situación de endogamia y más de dos hipótesis, una de las cuales involucra cálculos IBD.

Se analiza un marcador para tres individuos: una niña (CH), la madre indiscutida ($Madre$) y el padre indiscutido de esta última, a la vez abuelo del niño (AF). El objetivo es determinar si el (AF) es el padre de la niña. Sin embargo, se sospecha también del hermano de la madre (AU[3]), al que no se ha logrado realizar la prueba de ADN, pero que se sabe que es hermano completo, tanto por parte de padre como de madre, de AM.

$$\begin{cases} H_1: \ AF \text{ es el padre de la niña.} \\ H_2: \ AU \text{ es el padre de la niña.} \\ H_3: \ \text{El padre de la niña es un individuo desconocido no relacionado con } AF \text{ y } AU. \end{cases}$$

La información con la que contamos es la siguientes: $G_{AF} = 1/2$, $G_{AM} = 1/3$ y $G_{CH} = 1/2$. Las probabilidades alélicas en la población a la que pertenecen son $p_1 = p_2 = p_3 = 0{,}05$, con lo que vemos que son alelos bastante minoritarios.

[3]Del inglés *Alleged Uncle* (supuesto tío).

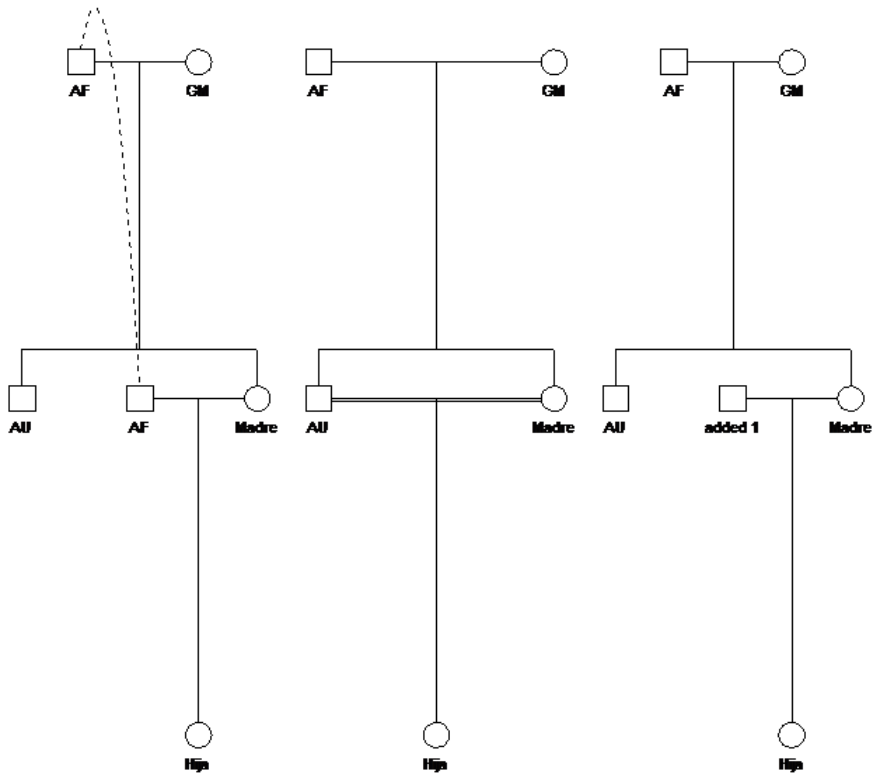

Figura 4.2: Ejemplo de parentesco complejo y endogamia

Hay que tener en cuenta que para definir parentescos como ser hermanos por ambas partes utilizando Familias es necesario crear a ambos progenitores aunque de alguno no poseamos información que pueda ser utilizada en el caso, como puede ser la madre de AM y abuela de CH en este caso, que identificaremos por GM[4]. Notemos también que como no contamos con información sobre AU, los cálculos para contrastar H_2 se basan en la técnica IBD.

Como es habitual en estos casos, la razón de verosimilitudes se calcula contrastando las hipótesis frente a la última, que considera que el padre es un individuo desconocido, pues es la forma más imparcial de compararlas. Es decir, obtendremos:

$$LR_i = \frac{Pr(datos \mid H_i)}{Pr(datos \mid H_3)}, i \in \{1, 2\}\,. \tag{4.1}$$

```
> #EJEMPLO: Parentesco complejo y endogamia

> p<-c("AF","GM","AU","Madre","Hija")
> s<-c("male","female","male","female","female")

> #AF es el padre:
> ped1<-FamiliasPedigree(id=p,c(NA,NA,"AF","AF","AF"),
```

[4]Del inglés *Grandmother* (abuela).

```
c(NA,NA,"GM","GM","Madre"),sex=s)

> #AU es el padre:
> ped2<-FamiliasPedigree(id=p,c(NA,NA,"AF","AF","AU"),
c(NA,NA,"GM","GM","Madre"),sex=s)

> #Un desconocido no relacionado con AF o AU es el padre
> ped3<-FamiliasPedigree(id=p,c(NA,NA,"AF","AF",NA),
c(NA,NA,"GM","GM","Madre"),sex=s)

> par(mfrow=c(1,3))
> plot(ped1);plot(ped2);plot(ped3)
> #Representacion en la Figura \ref{fig:ejParentComplej}

> genealogias<-list(Desconocido=ped3,esAF=ped1,esAU=ped2)
> genealogias
$Desconocido
$id
[1] "AF"    "GM"    "AU"    "Madre" "Hija"

$findex
[1] 0 0 1 1 0

$mindex
[1] 0 0 2 2 4

$sex
[1] "male"   "female" "male"   "female" "female"

attr(,"class")
[1] "FamiliasPedigree"

$esAF
$id
[1] "AF"    "GM"    "AU"    "Madre" "Hija"

$findex
[1] 0 0 1 1 1

$mindex
[1] 0 0 2 2 4

$sex
[1] "male"   "female" "male"   "female" "female"
```

```
attr(,"class")
[1] "FamiliasPedigree"

$esAU
$id
[1] "AF"    "GM"    "AU"    "Madre" "Hija"

$findex
[1] 0 0 1 1 3

$mindex
[1] 0 0 2 2 4

$sex
[1] "male"   "female" "male"   "female" "female"

attr(,"class")
[1] "FamiliasPedigree"

> #Marcador: alelos 1,2 y 3
> frecuencias<-c(0.05,0.05,0.05,1-(0.05+0.05+0.05))
> alelos<-c(1,2,3,"R") #R:alelo sobrante
> marcador<-FamiliasLocus(frecuencias,alelos)
> marcador
$locusname
[1] "frecuencias"

$alleles
   1    2    3    R
0.05 0.05 0.05 0.85

$femaleMutationType
[1] "No␣mutations"

$femaleMutationMatrix
  1 2 3 R
1 1 0 0 0
2 0 1 0 0
3 0 0 1 0
R 0 0 0 1

$maleMutationType
[1] "No␣mutations"

$maleMutationMatrix
```

```
  1 2 3 R
1 1 0 0 0
2 0 1 0 0
3 0 0 1 0
R 0 0 0 1

$simpleMutationMatrices
[1] TRUE

$Stabilization
[1] "NONE"

attr(,"class")
[1] "FamiliasLocus"

> #Genotipos:
> AF<-c(1,2)
> Madre<-c(1,3)
> Hija<-c(1,2)
> matrizdatos<-rbind(AF,Madre,Hija)
> matrizdatos
      [,1] [,2]
AF       1    2
Madre    1    3
Hija     1    2

> #Calculos verosimilitud y probabilidades:
> resultado<-FamiliasPosterior(genealogias,marcador,matrizdatos)
> resultado
$posterior
Desconocido         esAF         esAU
 0.06153846   0.61538462   0.32307692

$prior
Desconocido         esAF         esAU
  0.3333333    0.3333333    0.3333333

$LR
Desconocido         esAF         esAU
       1.00        10.00         5.25

$LRperMarker
            Desconocido esAF esAU
frecuencias           1   10 5.25
```

```
$likelihoods
 Desconocido          esAF          esAU
3.125000e-06 3.125000e-05 1.640625e-05

$likelihoodsPerSystem
            Desconocido      esAF          esAU
frecuencias   3.125e-06 3.125e-05 1.640625e-05
```

Como vemos, que AF sea el padre hace los datos diez veces más probables que si lo es un desconocido, lo que supone una evidencia que favorece H_1 frente a H_3. Análogamente, si AU es el padre los datos 5,25 veces más probables que si lo es un desconocido. Basándonos solo en este marcador en principio se refuerza la sospecha de existencia de endogamia. Por supuesto, esta información es insuficiente para considerar concluyentes los resultados. Una buena manera de continuar sería analizar más marcadores y tratar de obtener más información sobre el genotipo de AU.

4.2.3. Identificación de víctimas en desastres: DVI

Una aplicación muy destacada de todas las herramientas presentadas a lo largo del trabajo es la identificación de víctimas en desastres DVI[5], como pudo ser la identificación de restos tras el huracán Katrina o tras los atentados del en Nueva York del 11 de septiembre de 2001.

Lo normal en este tipo de situaciones es contar con una lista de desaparecidos y una lista de muestras halladas. La manera habitual de modelizar estas situaciones[6] es analizar las muestras no identificadas (que pueden ser o no de individuos diferentes) y elaborar una lista $X_1, X_2, ..., X_n$ de perfiles genéticos de individuos no identificados, a las que vamos a referirnos como **desaparecidos**. Por otra parte vamos a contar con una lista de individuos a los que pueden pertenecer dichas muestras. Para analizar si hay coincidencia, se elabora un perfil genético para cada uno de los candidatos. A este conjunto de perfiles $X'_1, X'_2, ..., X'_k$ le llamaremos **individuos de referencia** o **familias**. El motivo es que los perfiles se construyen a través de datos de referencia recogidos antemortem, que pueden ser desde muestras antiguas de desaparecidos (casos de desaparición), muestras de los sospechosos (casos criminales) o datos de referencia de personas emparentadas con el individuo en cuestión (identificación en catástrofes). Al conjunto de genotipos utilizados para construir el perfil genético de un desaparecido le llamaremos **familia**.

En resumen, vamos a contar con dos bases de datos en nuestro problema:

- No identificados: $NI = X_1, X_2, ..., X_n$,
- Desaparecidos: $D = X'_1, X'_2, ..., X'_k$.

[5]Del inglés *Disaster Victim Identification* (identificación de víctimas en desastres).

[6]Con cambios mínimos este planteamiento se puede aplicar a investigaciones criminales. Basta sustituir la lista de muestras de víctimas del desastre por una lista de muestras de la escena del crimen y la de desaparecidos (es decir, de perfiles de referencia) por la de sospechosos a los que puede pertener.

El objetivo es encontrar la asignación más probable. En general, la comparación en estos casos se basa en un proceso iterativo que puede ser muy costoso. En general suelen usarse paquetes más potentes que implementan atajos para ahorrar coste computacional, pero vamos a plantear un ejemplo y resolverlo con Familias para ilustrar que, al final, los cálculos subyacentes están basados en el contenido presentado en el capítulo 2.

Vamos a investigar un caso sencillo. El hundimiento de un pequeño bote en un río cercano ha dejado un número desconocido de víctimas. Del escenario del accidente se han extraído y analizado cinco muestras de ADN: R_1, R_2, R_3, R_4, R_5. Los genotipos obtenidos para el marcador que vamos a estudiar son $(12/12, 12/12, 13/14, 15/15, 13/13)$ respectivamente. Se ha denunciado la desaparición, presumiblemente en el naufragio, de tres personas cada una de una familia. Para determinar sus perfiles genéticos se ha recopilado la siguiente información: para la familia F_1 se conoce el genotipo del padre $G_{AF}^{F_1} = 12/12$. Para la familia F_2 se cuenta con una muestra antemortem de la desaparecida de la que se ha obtenido el genotipo 13/14. Por último, para la tercera familia se ha podido analizar el ADN del hermano de la desaparecida, y se ha obtenido $G_{BR}^{F_3} = 15/15$.

Las frecuencias en la población de los alelos involucrados son: $p_{12} = 0{,}1, p_{13} = 0{,}2, p_{14} = 0{,}05$ y $p_{15=0,2}$.

Para buscar la mejor asignación, vamos a trabajar con probabilidades a posteriori admitiendo equiprobabilidad para todas las asignaciones. Es decir, para la familia F_1:

$$\begin{cases} H_1: \ R_1 \text{ pertenece a } F_1. \\ H_2: \ R_2 \text{ pertenece a } F_1. \\ H_3: \ R_3 \text{ pertenece a } F_1. \\ H_4: \ R_4 \text{ pertenece a } F_1. \\ H_5: \ R_5 \text{ pertenece a } F_1. \\ H_6: \ \text{El individuo perteneciente a } F_1 \text{ no ha sido hallado.} \end{cases}$$

$$Pr(H_1) = Pr(H_2) = Pr(H_3) = Pr(H_4) = Pr(H_5) = Pr(H_6) = \frac{1}{6} \tag{4.2}$$

De nuevo se van a contrastar las hipótesis frente a la última, la posibilidad de que los restos de la persona desaparecida correspondiente a esta familia no hayan sido encontrados.

```
> #FAMILIA 1
> library(Familias)

> v<-c("R1","R2","R3","R4","R5")
> #x es un individuo no relacionado que no es ninguno de los anteriores.
```

```
> #Marcador:
> #Definimos el marcador descrito en el enunciado:
> frecuencias<-c(0.1,0.2,0.05,0.2,1-(0.1+0.2+0.05+0.2))
> alelos<-c(12,13,14,15,"a")
> M<-rbind(frecuencias,alelos)
> marcador<-FamiliasLocus(frecuencias,alelos)

> #Genotipos:
> R1<-c(12,12)
> R2<-c(12,12)
> R3<-c(13,14)
> R4<-c(15,15)
> R5<-c(13,13)
> AF<-c(12,12)
> matrizdatos<-rbind(R1,R2,R3,R4,R5,AF)

> sex=c("male","female")
> dadid=c(NA,"AF")
> momid=c(NA,NA)

> noPertenece=1:length(v)
> siPertenece=1:length(v)
> Result=cbind(noPertenece,siPertenece)

> for (i in 1:length(v)){
+ w=1:length(v)
+ matdat<-matrizdatos
+ id=c("AF",v[i])
+ ped1<-FamiliasPedigree(id,dadid,momid,sex)
+ ped2<-FamiliasPedigree(id,dadid=c(NA,NA),momid=c(NA,NA),sex)
+ linajes<-list(noPertenece=ped2,siPertenece=ped1)
+ w<-w[-i]
+ matdat<-matrizdatos[-w,]
+ resultado<-FamiliasPosterior(linajes,marcador,matdat)
+ Result[i,]<-resultado$LR
+ }
> Result=rbind(Result,c(1,1))

> #Probabilidades a posteriori:
> LRi<-Result[,"siPertenece"]
> Post1<-LRi/sum(LRi)

> #FAMILIA 2:
> #Lo he modelizado como si tuviese una familia con padre homocigotos
```

```
> para cada alelo del individuo.

> #Genotipos:
> AF<-c(13,13)
> MO<-c(14,14)
> matrizdatos<-rbind(R1,R2,R3,R4,R5,AF,MO)

> sex=c("male","female","female")
> dadid=c(NA,NA,"AF")
> momid=c(NA,NA,"MO")

> noPertenece=1:length(v)
> siPertenece=1:length(v)
> Result=cbind(noPertenece,siPertenece)

> for (i in 1:length(v)){
+ w=1:length(v)
+ matdat<-matrizdatos
+ id=c("AF","MO",v[i])
+ ped1<-FamiliasPedigree(id,dadid,momid,sex)
+ ped2<-FamiliasPedigree(id,dadid=c(NA,NA,NA),momid=c(NA,NA,NA),sex)
+ linajes<-list(noPertenece=ped2,siPertenece=ped1)
+ w<-w[-i]
+ matdat<-matrizdatos[-w,]
+ resultado<-FamiliasPosterior(linajes,marcador,matdat)
+ Result[i,]<-resultado$LR
+ }
> Result=rbind(Result,c(1,1))

> #Probabilidades a posteriori:
> LRi<-Result[,"siPertenece"]
> Post2<-LRi/sum(LRi)

> #FAMILIA 3

> #Genotipos:
> BR<-c(15,15)
> matrizdatos<-rbind(R1,R2,R3,R4,R5,BR)

> sex=c("male","female","male","female")
> dadid=c(NA,NA,"AF","AF")
> momid=c(NA,NA,"AM","AM")

> noPertenece=1:length(v)
> siPertenece=1:length(v)
```

```
> Result=cbind(noPertenece,siPertenece)

> for (i in 1:length(v)){
+ w=1:length(v)
+ matdat<-matrizdatos
+ id=c("AF","AM","BR",v[i])
+ ped1<-FamiliasPedigree(id,dadid,momid,sex)
+ ped2<-FamiliasPedigree(id,dadid=c(NA,NA,"AF",NA),momid=c(NA,NA,"AM",NA),sex)
+ linajes<-list(noPertenece=ped2,siPertenece=ped1)
+ w<-w[-i]
+ matdat<-matrizdatos[-w,]
+ resultado<-FamiliasPosterior(linajes,marcador,matdat)
+ Result[i,]<-resultado$LR
+ }
> Result=rbind(Result,c(1,1))

> #Probabilidades a posteriori:
> LRi<-Result[,"siPertenece"]
> Post3<-LRi/sum(LRi)

> #Resultado final:
> Post1
[1] 0.47619048 0.47619048 0.00000000 0.00000000 0.00000000 0.04761905
> Post2
[1] 0.00000000 0.00000000 0.98039216 0.00000000 0.00000000 0.01960784
> Post3
[1] 0.02272727 0.02272727 0.02272727 0.81818182 0.02272727 0.09090909
```

Veamos como interpretar estos resultados. Cada una de las filas anteriores representa las probabilidades a posteriori de cada hipótesis para las familias F_1, F_2 y F_3 respectivamente.

Para la primera familia está claro que los restos R_3, R_4 y R_5 pueden descartarse como procedentes del individuo desaparecido que reclaman, al igual que la hipótesis de que los restos no hayan sido hallados. Lógicamente, R_1 y R_2 presentan la misma probabilidad porque tienen el mismo genotipo. A la segunda familia parece bastante claro que hemos de asignarle la muestra R_3. La tercera familia puede ser la que de lugar a más confusión, aunque parece que la muestra que coincide más con el perfil buscado es la muestra R_4.

Una conclusión razonable basándonos en la información para este marcador, sería que las muestras R_1 y R_2 proceden de la misma víctima, que pertenece a la familia F_1. La muestra R_3 correspondería a la familia F_2 y la muestra R_4, a la familia F_3. Si investigando las circunstancias no parece probable que existan más víctimas, cabe sospechar que la muestra R_5 proceda de la misma fuente que R_4, pero se haya genotipado erróneamente (no es raro que en este tipo de casos las muestras estén degradadas).

Capítulo 5

Fundamentos teóricos del modelo

5.1. Introducción: inferencia bayesiana

Vamos a presentar la base conceptual y las distribuciones necesarias para realizar inferencia sobre un parámetro ϕ a partir de una muestra[1]. La base de la inferencia bayesiana se basa en el Teorema de Bayes tal y como se menciona en la sección (2.2.3)

5.1.1. Distribución a priori conjugada (conjugate prior)

Se sabe que existe cierta variable en la población con distribución paramétrica conocida, pero de la que se ignora el parámetro que denotaremos por ϕ. En este contexto, se han extraído de la población varias observaciones, que consideraremos datos y llamaremos $\boldsymbol{x}$. Nuestro objetivo es determinar la verosimilitud del parámetro, es decir, la probabilidad de cada valor del parámetro ϕ teniendo en cuenta los datos $\boldsymbol{x}$. Los datos con los que contaremos serán: la función de verosimilitud $\phi \longrightarrow Pr(\boldsymbol{x}|\phi)$, que tiene una forma fija conocida que se deduce de la forma en la que se recogen los datos, la probabilidad a priori del parámetro $Pr(\phi)$ y la muestra $\boldsymbol{x}$ extraída de la población.

Recordamos la expresión (2.10):

$$Pr(\phi|\boldsymbol{x}) = \frac{Pr(\boldsymbol{x}|\phi)\ Pr(\phi)}{\int_\phi Pr(\boldsymbol{x}|\phi)\ Pr(\phi)\ d\phi}. \tag{5.1}$$

Los cálculos para obtener la probabilidad anterior dependen en gran medida de la probabilidad a priori asumida para ϕ, ya que esta determina la expresión algebraica de $Pr(\boldsymbol{x}|\phi)Pr(\phi)$ y, como consecuencia, puede dificultar o facilitar el cálculo de la integral (teniendo incluso que llegar a acudir a métodos numéricos). Si para la probabilidad a priori escogida se cumple que $Pr(\boldsymbol{x}|\phi)Pr(\phi)$ tiene la misma forma algebraica que $Pr(\phi)$, entonces se dice que es la **probabilidad a priori conjugada** de la distribución de $Pr(\boldsymbol{x}|\phi)$.

Nota 5.1. *La probabilidad a priori adoptada para el parámetro tendrá a su vez sus respectivos parámetros, que para evitar confusión se conocen como* ***hiperparámetros****.*

[1] El contenido de esta primera parte está basado principalmente en las siguientes fuentes: [7], [2], [22], [20], [19], [23] y [21].

5.1.2. Distribuciones

Distribución multinomial

La distribución multinomial es un tipo de distribución discreta que generaliza la distribución categórica.

Dado un experimento aleatorio que puede presentar $\{1, 2, ...k\}$ resultados diferentes, cada uno de ellos con probabilidad p_i de forma que $\sum_{i=1}^{k} p_i = 1$, se dice que el resultado de este experimento sigue una distribución categórica de parámetro $\boldsymbol{p}$, con $\boldsymbol{p} = (p_1, ...p_k)$.

Si en la situación anterior repetimos el experimento n veces, entonces el conjunto de resultados se puede expresar a través de un vector $\boldsymbol{C} = (C_1, ...C_k)$, donde cada C_i representa la cantidad de resultados de tipo i obtenidos en los n intentos. Se dice entonces que el vector de resultados $\boldsymbol{C}$ sigue una distribución binomial de parámetros n y $\boldsymbol{p}$.

$$\boldsymbol{C} \sim Multinom(n, \boldsymbol{p})$$

La función de probabilidad de esta distribución es:

$$Pr(\boldsymbol{C} = \boldsymbol{x}) = \begin{cases} \frac{n!}{x_1!\cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} & si \quad \sum_{i=1}^{k} x_i = n, \\ \\ 0 & \text{en caso contrario.} \end{cases} \tag{5.2}$$

Teniendo en cuenta la expresión de la función gamma para números naturales

$$\Gamma(z) = (z-1)!, z \in \mathbb{N}, \tag{5.3}$$

podemos reescribir la función de densidad utilizando la siguiente igualdad:

$$Pr(\boldsymbol{C} = \boldsymbol{x}) = \frac{n!}{x_1!\cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} = \frac{\Gamma(\sum_{i=1}^{k} x_i + 1)}{\prod_{i=1}^{k} \Gamma(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i}. \tag{5.4}$$

Ejemplo 5.2. Para resolver un caso de determinación de parentesco se estudia un marcador M que puede presentar tres alelos distintos A, B, C todos ellos equiprobables a priori. Como parte del caso hemos secuenciado el ADN para este marcador de dos individuos involucrados obteniendo los genotipo: $A/B, \quad C/C$. Calculemos la probabilidad de obtener esta muestra:

$k = 3$, los tres alelos son equiprobables y $n = 4$ luego $\boldsymbol{C} \sim Multinom\left(4, \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)\right)$.

A la vista de los genotipos observados: $\boldsymbol{x} = (1, 1, 2)$.

Por tanto, utilizando la ecuación (5.4) obtenemos:

$$Pr(\boldsymbol{C} = \boldsymbol{x}) = \frac{4!}{1!1!2!} \left(\frac{1}{3}\right)^1 \left(\frac{1}{3}\right)^1 \left(\frac{1}{3}\right)^2 = \frac{2}{3^4}. \tag{5.5}$$

Distribución Dirichlet

La distribución Dirichlet (también llamada distribución beta multivariante) es una distribución continua multivariante determinada por un parámetro $\boldsymbol{\lambda}$. Este parámetro es un vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_k)$, $k \geq 2$ y $\lambda_i > 0$ para todo $i \in \{1, ..., k\}$.

El soporte de una variable Dirichlet es el conjunto de los vectores de dimensión k tales que $x_i > 0$ para todo $i \in \{1, ..., k\}$ y $\sum_{i=1}^{k} x_i = 1$. A la vista de estas condiciones, queda claro que es posible tomar como distribución a priori de un vector de probabilidades k-dimensional $\boldsymbol{p}$ una distribución Dirichlet con $\boldsymbol{\lambda}$ su hiperparámetro.

$$\boldsymbol{p} \sim \text{ Dirichlet}(\boldsymbol{\lambda}) \tag{5.6}$$

La función de densidad para esta variable, que denotaremos por $\pi(\boldsymbol{p})$, es:

$$\pi(\boldsymbol{p}) \equiv \pi(\boldsymbol{p}|\boldsymbol{\lambda}) = \frac{1}{B(\boldsymbol{\lambda})} \prod_{i=1}^{k} p_i^{\lambda_i - 1}, \tag{5.7}$$

donde $B(\boldsymbol{\lambda})$ es la función beta $B(\boldsymbol{\lambda}) = \frac{\prod_{i=1}^{k} \Gamma(\lambda_i)}{\Gamma(\sum_{i=1}^{k} \lambda_i)}$.

La distribución marginal de cada p_i en estas circunstancias es $p_i \sim Beta(\lambda_i, \Lambda - \lambda_i)$, donde $\Lambda = \sum_{i=1}^{k} \lambda_i$. Por lo tanto su media y su varianza son respectivamente:

$$E(p_i) = \frac{\lambda_i}{\Lambda}, \tag{5.8}$$

$$Var(p_i) = \frac{\tilde{\lambda_i}(1 - \tilde{\lambda_i})}{\Lambda + 1}, \qquad \tilde{\lambda_i} = \frac{\lambda_i}{\Lambda}. \tag{5.9}$$

5.1.3. Probabilidad a posteriori

Supongamos ahora que estamos estudiando un marcador no autosómico con k alelos posibles. Denotamos por p_i la probabilidad de cada uno de esos alelos, y asumimos que la distribución a priori de $\boldsymbol{p}$ es una Dirichlet de hiperparámetro $\boldsymbol{\lambda}$ (5.6). Supongamos ahora que hemos obtenido el genotipo relativo a este marcador para una serie de individuos, y denotamos por $\boldsymbol{c}$ al vector contador para esta muestra. Estudiemos la probabilidad a posteriori de $\boldsymbol{p}$ una vez observados estos datos particularizando la expresión (2.9):

$$Pr(\boldsymbol{p}|\boldsymbol{c}) = \frac{Pr(\boldsymbol{p} \cap \boldsymbol{c})}{Pr(\boldsymbol{c})} = \frac{Pr(\boldsymbol{c}|\boldsymbol{p})Pr(\boldsymbol{p})}{Pr(\boldsymbol{c})}. \tag{5.10}$$

Conforme a lo que hemos visto en las secciones previas, podemos asumir:

- $Pr(\boldsymbol{c}|\boldsymbol{p})$ es la función de probabilidad de una $Multinomial(n, \boldsymbol{p})$.
- $Pr(\boldsymbol{p}) \equiv \pi(\boldsymbol{p})$ es la función de densidad de una $Dir(\boldsymbol{\lambda})$.

Por tanto considerando (5.2) y (5.7):

$$Pr(\boldsymbol{c}|\boldsymbol{p})Pr(\boldsymbol{p}) \equiv Pr(\boldsymbol{c}|\boldsymbol{p})\pi(\boldsymbol{p}) = \frac{n!}{\prod_{i=1}^{n} c_i!} \prod_{i=1}^{k} p_i^{c_i} \cdot \frac{1}{B(\boldsymbol{\lambda})} \prod_{i=1}^{k} p_i^{\lambda_i - 1}. \tag{5.11}$$

Comprobemos que la distribución Dirichlet es la distribución a priori conjugada de la multinomial. Para ello basta ver que la probabilidad a posteriori de $\boldsymbol{p}$ sigue una distribución de tipo Dirichlet (como la distribución Dirichlet es continua, la notación debe cambiar para hablar de su función de densidad $Pr(\boldsymbol{c}|\boldsymbol{p}) \equiv \pi(\boldsymbol{c}|\boldsymbol{p})$):

$$\pi(\boldsymbol{p}|\boldsymbol{c}) \propto_p \frac{n!}{\prod_{i=1}^{n} c_i!} \prod_{i=1}^{k} p_i^{c_i} \cdot \frac{1}{B(\boldsymbol{\lambda})} \prod_{i=1}^{k} p_i^{\lambda_i - 1} \propto_p \prod_{i=1}^{k} p_i^{c_i + \lambda_i - 1}, \tag{5.12}$$

$\propto_p$ indica proporcionalidad como función de $\boldsymbol{p}$.

Por tanto en efecto la probabilidad a posteriori sigue una distribución de tipo Dirichlet, en particular si condicionamos la variable $\boldsymbol{p} \sim Dirichlet(\boldsymbol{\lambda})$ al vector de observaciones $\boldsymbol{c}$, el resultado es una variable:

$$\boldsymbol{p}|\boldsymbol{c} \sim Dirichlet(\boldsymbol{\lambda} + \boldsymbol{c}) \tag{5.13}$$

5.2. Modelo teórico para la inferencia sobre linajes

En este capítulo vamos a explicar como construir un modelo que permita utilizar los datos sobre el genoma de un grupo de individuos para determinar la relación existente entre ellos. El contenido se ha construido a partir de [7] y [8].

Los datos de partida ($\mathcal{D}$) serán los perfiles genéticos relativos a uno o varios marcadores de un grupo de sujetos involucrados en el problema, que serán obtenidos mediante técnicas de secuenciación. Ante esta situación existen dos planteamientos posibles: construir una genealogía a partir de los datos o considerar un conjunto predeterminado de posibles genealogías y utilizar los datos para decidir. Nos centraremos en este segundo enfoque. Por tanto, contaremos también con una colección $\{\mathcal{P}_1, ..., \mathcal{P}_N\}$ de posibles linajes relacionando al conjunto de individuos de interés. La definición de distintos linajes descansa en la suposición de que los fundadores en cada caso (los individuos de la genealogía cuyos padres no están incluídos en la genealogía) no están relacionados. Este planteamiento presenta dos ventajas: en primer lugar se restringen los posibles escenarios a un grupo conocido y en segundo lugar cada posibilidad se formula en términos de una de esas posibles genealogías. Por último, por supuesto será necesario conocer la probabilidad a priori de cada posible linaje: $\{\rho_1, ..., \rho_N\}$.

Contando con los elementos anteriores, podemos entonces aplicar el Teorema de Bayes (2.2.3) y obtenemos la siguiente expresión para la verosimilitud del linaje conocidos los datos:

$$Pr(\mathcal{P}_j|\mathcal{D}) = \frac{p_j Pr(\mathcal{D}|\mathcal{P}_j)}{\sum_{k=1}^{N} p_k Pr(\mathcal{D}|\mathcal{P}_k)}. \tag{5.14}$$

Será necesario entonces encontrar la manera de calcular $Pr(\mathcal{D}|\mathcal{P})$ para cada posible linaje[2]. La forma de lograrlo es construir un modelo. En este caso dicho modelo constará de tres partes e involucrará los siguientes elementos (ver la Figura 5.1):

[2]Para desarrollar la teoría se prescindirá de subíndices y se trabajará sobre una genealogía genérica $\mathcal{P}$ a fin de aligerar la notación.

- $\mathcal{P}$: linaje que vamos a considerar, esto es, un posible esquema de las relaciones de los individuos. El modelo que vamos a ver admite en la genealogía individuos para los que solo un progenitor está contemplado dentro del linaje, lo que supone un avance con respecto a planteamientos previos.

- $\mathcal{D}$ (ya definido)

- $\mathcal{G}$: es el genotipo escalonado de los individuos a los que se ha realizado la prueba de ADN. Por genotipo escalonado entendemos el genotipo en el que no solo se especifican los dos alelos del individuo sino también su procedencia (es decir, si es materno o paterno). En general esta información es difícil de obtener experimentalmente, pero encontrar un modelo que funcione en este caso servirá en particular si el genotipo no está escalonado.

- $\mathcal{F}$: denota el genotipo escalonado[3] y ordenado de los fundadores del linaje. Cuando solo un progenitor está incluido en $\mathcal{P}$, se considera fundador al progenitor externo y puede suceder que solo se incluya en $\mathcal{F}$ un alelo del mismo (el que se transmita a la descendencia, por ejemplo ver (*) en la Figura 5.1).

Nota 5.3. *Cuando contamos con m marcadores, se utilizará la notación $\mathcal{D}_i, \mathcal{G}_i, \mathcal{F}_i$ para denotar el correspondiente subconjunto de datos relativo al marcador i-ésimo.*

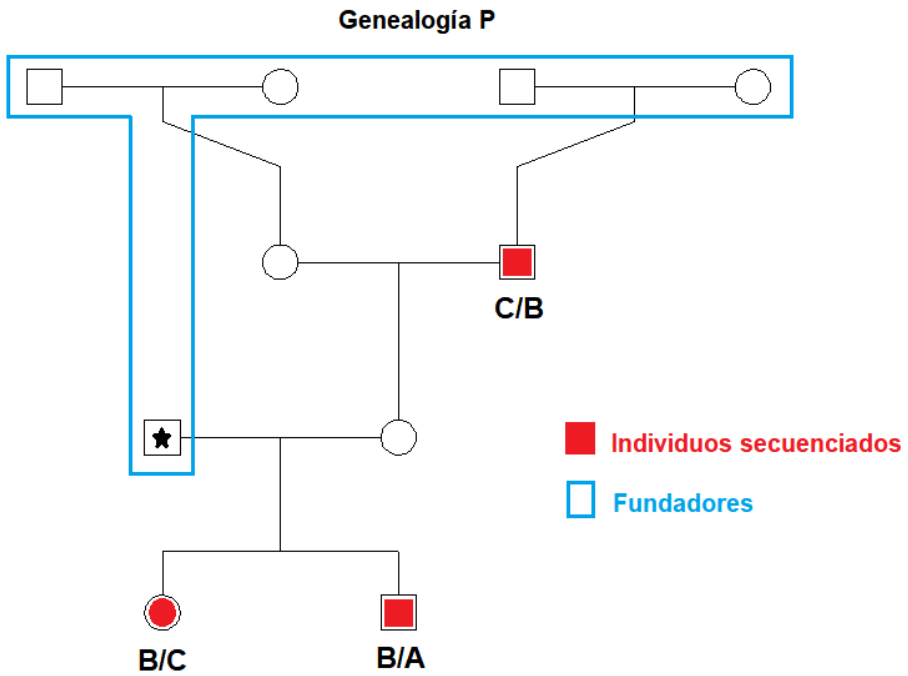

Figura 5.1: Ejemplo de posible linaje

[3]El genotipo escalonado para los fundadores implica que a cada alelo de un fundador se le asigna una procedencia al azar.

Descomponemos la probabilidad objetivo en función de los términos anteriores:

$$Pr(\mathcal{D}|\mathcal{P}) = \sum_{\mathcal{G}} \sum_{\mathcal{F}} Pr(\mathcal{D}, \mathcal{F}, \mathcal{G}|\mathcal{P}). \tag{5.15}$$

Cada sumando se decompone utilizando el Teorema de Factorización (2.6):

$$Pr(\mathcal{D}, \mathcal{F}, \mathcal{G}|\mathcal{P}) = Pr(\mathcal{D}|\mathcal{F}, \mathcal{G}, \mathcal{P}) \cdot Pr(\mathcal{F}|\mathcal{G}, \mathcal{P}) \cdot Pr(\mathcal{G}|\mathcal{P}). \tag{5.16}$$

Asumimos las siguientes simplificaciones:

- Dados los genotipos obtenidos para las personas testadas $\mathcal{D}$, vamos a suponer que toda la información que pueda influir en los datos observados está recogida en $\mathcal{G}$ (es decir, $\mathcal{F}$ y $\mathcal{P}$ influyen en los datos observados solo a través de su influencia sobre $\mathcal{G}$).

$$Pr(\mathcal{D}|\mathcal{F}, \mathcal{G}, \mathcal{P}) = Pr(\mathcal{D}|\mathcal{G}). \tag{5.17}$$

- Vamos a suponer que el linaje considerado no influye en los genotipos de los fundadores.

$$Pr(\mathcal{F}|\mathcal{P}) = Pr(\mathcal{F}). \tag{5.18}$$

Introduciendo los supuestos anteriores en la expresión (5.15):

$$Pr(\mathcal{D}|\mathcal{P}) = \sum_{\mathcal{G}} \sum_{\mathcal{F}} Pr(\mathcal{D}|\mathcal{G})\ Pr(\mathcal{G}|\mathcal{F}, \mathcal{P})\ Pr(\mathcal{F}). \tag{5.19}$$

El modelo se dividirá entonces en tres partes:

- **Modelo a nivel poblacional $\mathbf{Pr(\mathcal{F})}$**: busca especificar la probabilidad de cada posible genotipo de los fundadores.

- **Modelo a nivel genealógico $\mathbf{Pr(\mathcal{G} \mid \mathcal{F}, \mathcal{P})}$**: dado un genotipo para los fundadores y una genealogía, permite especificar la probabilidad de un genotipo particular para las personas testadas.

- **Modelo a nivel observacional $\mathbf{Pr(\mathcal{D} \mid \mathcal{G})}$**: especificado el genotipo de los individuos testados, permite calcular la probabilidad de obtener los datos $\mathcal{D}$ al realizar las pruebas de ADN.

Vamos a limitarnos a explicar la construcción de cada parte del modelo poniendo énfasis en los aspectos conceptuales. Por ello, situaciones complejas como la existencia de mutaciones, dropouts, dependencia entre marcadores, etc no serán tenidos en cuenta.

5.2.1. Modelos a nivel poblacional

Recordemos que el objetivo es calcular $Pr(\mathcal{F})$.

Vamos a considerar un único marcador autosómico que puede presentar k alelos diferentes, cuya probabilidad en la población viene dada por el vector $\boldsymbol{p} = (p_1, p_2, ..., p_k)$ (cuyos valores deben sumar 1). Se asumirá además equilibrio de Hardy-Weinberg (ver la sección 2.3). Empezaremos hablando de la incertidumbre en las probabilidades poblacionales de cada alelo. Veremos como tener en cuenta esta incertidumbre a la hora de calcular la probabilidad del genotipo de los fundadores. Por último veremos como implementar en nuestro modelo la estratificación de la población.

Incertidumbre en las frecuencias

El procedimiento habitual para estimar la probabilidad de cada alelo en la población es tomar la frecuencia relativa de dicho alelo en una base de datos. Es decir, el marcador involucrado debe estar estudiado, y por tanto se contará con un base de datos con N observaciones del alelo (es decir, observaciones correspondientes a $N/2$ individuos) y un vector contador, es decir, un vector $\boldsymbol{C} = (C_1, C_2, ..., C_v)$ donde $C_i = \#\{\text{observaciones del alelo } i \text{ en la base de datos}\}$ y $\{1, 2, ..., v\}$ el conjunto de alelos observados. Es decir, la probabilidad de cada alelo se estima con un enfoque frecuentista, utilizando la regla de Laplace:

$$\boldsymbol{p} = \frac{\boldsymbol{C}}{N}. \tag{5.20}$$

Este enfoque tiene varios inconvenientes, por un lado no hay manera de tener en cuenta la incertidumbre del parámetro. Por otro, y quizás más importante, si un alelo no ha sido observado a la hora de confeccionar la base de datos se le asigna automáticamente una probabilidad cero, lo que imposibilita cualquier cálculo. Veremos a continuación un enfoque que corrige ambos defectos, en particular asigna a cada *posible* alelo un escalar positivo (muy pequeño si no ha sido observado en la base de datos, pero distinto de cero).

Nota 5.4. *A lo largo de este capítulo nos hemos referido a la probabilidad de encontrar un tipo de alelo específico de forma aleatoria en la población como probabilidad poblacional o probabilidad del alelo en la población. Aunque esa sería la denominación correcta en términos de rigurosidad, debido a que tradicionalmente se estima usando las frecuencias relativas del alelo en la base de datos, está muy consolidado en la literatura el término frecuencias alélicas para referirse a la probabilidad de los alelos. En lo sucesivo se adoptará esta nomenclatura que aunque no sea la más adecuada, es la más extendida.*

Vamos a calcular las frecuencias alélicas empleando ahora un enfoque bayesiano. Esta vez tendremos en cuenta los k alelos *posibles* desde un punto de vista biológico, hayan sido observados o no al construir la base de datos. Es decir el vector $\boldsymbol{C} = (C_1, C_2, ..., C_k)$ de observaciones puede tener componentes que sean cero.

Ejemplo 5.5. Supongamos que estamos considerando un marcador STR. En la base de datos hemos observamos los siguientes alelos: 12, 13 y 15 repeticiones respectivamente. Empleando un enfoque frecuentista, únicamente podríamos realizar operaciones para estos tres alelos. Sin embargo, desde el punto de vista biológico es razonable suponer la existencia también de un alelo con 14 copias, e incluso otros alelos con copias parciales, siendo también coherente suponer una probabilidad a priori menor para estos últimos. Este nuevo enfoque nos permite incluir estos alelos no observados, e incluso asignar una probabilidad menor a los alelos con copias parciales que al alelo 14, pese a que en ambos casos no han sido observados.

Para lograrlo, vamos a utilizar la distribución a posteriori de $\boldsymbol{p}$, pues queremos tener en cuenta las observaciones recogidas en la base de datos. Así, recuperamos la expresión (5.13):

$$\boldsymbol{p}|\boldsymbol{c} \equiv Dirichlet(\boldsymbol{\lambda} + \boldsymbol{c}), \tag{5.21}$$

y notamos que entonces cada $p_i|\boldsymbol{C}$ tiene distribución:

$$p_i|\boldsymbol{C} \sim Beta(\lambda_i + C_i, \Lambda + N - \lambda_i - C_i), \tag{5.22}$$

por lo que de acuerdo con la ecuación (5.9), su varianza será:

$$Var(p_i|\boldsymbol{C}) = \frac{\frac{\lambda_i+C_i}{\Lambda+N}\left(1 - \frac{\lambda_i+C_i}{\Lambda+N}\right)}{\Lambda + N + 1} = \frac{(\lambda_i + C_i)(\Lambda + N - \lambda_i - C_i)}{(\Lambda + N)^2(\Lambda + N + 1)}. \tag{5.23}$$

Por lo tanto, a medida que aumenta la cantidad de observaciones recogidas en la base de datos (es decir, N) menor será la varianza y con más precisión podremos estimar el parámetro $\boldsymbol{p}$.

Estimación de $\boldsymbol{p}$

Para obtener una estimación puntual del vector de frecuencias que tenga en cuenta la información de la muestra observada, se toma la esperanza de su distribución a posteriori (5.8):

$$\hat{\boldsymbol{p}} = E(\boldsymbol{p} \mid \boldsymbol{C}) = \frac{\lambda + \boldsymbol{C}}{\Lambda + N}. \tag{5.24}$$

Existe una analogía entre el papel de los λ_i y los contadores C_i en la expresión anterior. Debido a esta analogía, es común referirse a λ como vector de pseudocontadores, a pesar de que sus elementos ni siquiera tienen por qué ser naturales. Pueden establecerse distintos valores para distintos alelos, por ejemplo en el ejemplo 5.5 se asignaría un valor menor a los alelos con copias parciales. Como se puede ver en la expresión anterior, queda solventado además el problema para operar con alelos no observados, pues ahora se les asigna siempre una probabilidad positiva.

Probabilidad del genotipo de los fundadores

El fin último de esta sección es obtener la probabilidad del genotipo de los fundadores, $Pr(\mathcal{F})$. Para conseguirlo, la primera opción y la más evidente es calcular la probabilidad utilizando como vector de frecuencias la estimación que hemos obtenido en el apartado anterior (5.24):

$$Pr(\mathcal{F}) \approx Pr(\mathcal{F}|\widehat{\boldsymbol{p}}). \tag{5.25}$$

Sin embargo, con este enfoque no se tiene en cuenta la incertidumbre relativa al parámetro. El resultado siempre diferirá del real, si se necesita mucha precisión esto puede ser un problema.

Por eso, la opción preferente es recurrir a la integración sobre el parámetro en cuestión, pues de esta forma estamos utilizando la densidad de probabilidad $\pi(\phi)$ que condensa nuestro conocimiento del parámetro.

Es decir, recordando como se calculan las probabilidades totales para el caso de una distribución continua en la sección 2.2.3, y teniendo en cuenta que ahora $\boldsymbol{p}$ es el parámetro y el genotipo ordenado de los fundadores $\mathcal{F} = (f_1, f_2, ..., f_F)$ es la muestra extraída de la población:

$$Pr(\mathcal{F}) = \int_{\boldsymbol{p}} Pr(\mathcal{F}|\boldsymbol{p})\pi(\boldsymbol{p})d\boldsymbol{p}. \tag{5.26}$$

La información contenida en $\mathcal{F}$ puede condensarse en un nuevo vector de observaciones o vector contador $\boldsymbol{c} = (c_1, ..., c_k)$, donde c_i cuenta los alelos de tipo i que aparecen en $\mathcal{F}$. La suma de alelos observados es $\sum_{i=1}^{k} c_i = n$ la cantidad total de alelos incluidos en $\mathcal{F}$. Entonces:

$$Pr(\mathcal{F}|\boldsymbol{p}) = \prod_{i=1}^{k} p_i^{c_i}, \tag{5.27}$$

sustituyendo en (5.26), si suponemos que el parámetro $\boldsymbol{p}$ tiene una distribución Dirichlet de parámetro $\boldsymbol{\alpha}$, obtenemos:

$$Pr(\mathcal{F}) = \int_{\boldsymbol{p}} \prod_{i=1}^{k} p_i^{c_i}\pi(\boldsymbol{p})d\boldsymbol{p} = \frac{B(\boldsymbol{\alpha}+\boldsymbol{c})}{B(\boldsymbol{\alpha})} \tag{5.28}$$

Demostración.

$$Pr(\mathcal{F}) = \int_{\boldsymbol{p}} \prod_{i=1}^{k} p_i^{c_i}\pi(\boldsymbol{p})d\boldsymbol{p} = \int_{\boldsymbol{p}} \prod_{i=1}^{k} p_i^{c_i}\frac{1}{B(\alpha)}\prod_{i=1}^{k} p_i^{\alpha_i-1}d\boldsymbol{p} = \frac{1}{B(\alpha)}\int_{\boldsymbol{p}} \prod_{i=1}^{k} p_i^{\alpha_i+c_i-1}, \tag{5.29}$$

como la función de probabilidad de una distribución $Beta(\boldsymbol{\alpha}+\boldsymbol{c})$ viene dada por:

$$\frac{1}{B(\boldsymbol{\alpha}+\boldsymbol{c})}\prod_{i=1}^{k} {p_i}^{\alpha_i+c_i-1}d\boldsymbol{p}. \tag{5.30}$$

Si integramos la expresión anterior sobre el dominio de $\boldsymbol{p}$, como se trata de la función de probabilidad de una distribución, el resultado será la unidad. Por tanto podemos despejar como sigue:

$$B(\boldsymbol{\alpha}+\boldsymbol{c}) = \int_{\boldsymbol{p}} \prod_{i=1}^{k} {p_i}^{\alpha_i+c_i-1}d\boldsymbol{p}, \tag{5.31}$$

y sustituyendo en (5.29) obtenemos el resultado enunciado. □

Pero de (5.13) sabemos que en realidad en nuestro caso $\boldsymbol{\alpha} = \boldsymbol{\lambda} + \boldsymbol{C}$ ya que la función de densidad se obtiene como función de densidad a posteriori de $\boldsymbol{p}$ a la vista de la base de datos: $\boldsymbol{p}|\boldsymbol{C}$. Entonces sustituyendo en la expresión anterior:

$$Pr(\mathcal{F}) = \frac{B(\lambda+\boldsymbol{C}+\boldsymbol{c})}{B(\lambda+\boldsymbol{C})}. \tag{5.32}$$

Podemos reescribir la fórmula anterior utilizando la definición de función beta (5.1.2) de la siguiente forma: si j denota la posición en la secuencia ordenada de genotipos, a_j denota el tipo de alelo observado en j y b_j la cantidad de alelos de ese tipo observados hasta la posición $j-1$ incluida.

$$\begin{aligned} Pr(\mathcal{F}) &= \frac{B(\lambda+\boldsymbol{C}+\boldsymbol{c})}{B(\lambda+\boldsymbol{C})} = \frac{\prod_{i=1}^{k}\Gamma(\lambda_i+C_i+c_i)}{\prod_{i=1}^{k}\Gamma(\lambda_i+C_i)} \cdot \frac{\Gamma(\Lambda+N)}{\Gamma(\Lambda+N+n)} = \\ &= \frac{\prod_{i=1}^{k}[(\lambda_i+C_i)(\lambda_i+C_i+1)\cdots(\lambda_i+C_i+c_i-1)]}{(\Lambda+N)(\Lambda+N+1)\cdots(\Lambda+N+n-1)} = \prod_{j=1}^{n}\frac{\lambda_{a_j}+C_{a_j}+b_j}{\Lambda+N+j-1}. \end{aligned} \tag{5.33}$$

Reinterpretando la probabilidad anterior en base a esta nueva expresión, se puede entender como el producto de la probabilidad de observar en $\mathcal{F}$ el primer alelo condicionado a la información contenida en la base de datos sobre el mismo, por la probabilidad de observar el segundo alelo condicionado a la información contenida en la base de datos sobre él *y el hecho de haber observado el primer alelo* y así sucesivamente. Es decir, si denotamos a la base de datos por *db* estamos calculando:

$$\mathcal{F} = a_1, a_2, a_3, ..., a_n$$

$$Pr(\mathcal{F}) \equiv Pr(\mathcal{F}|db) = Pr(a_1|db) \cdot Pr(a_2|a_1, db) \cdot \ \cdots \ \cdot Pr(a_n|a_1, ..., a_{n-1}, db). \tag{5.34}$$

Otra forma de verlo es que basta con ir actualizando la base de datos secuencialmente de forma que se tenga en cuenta toda la información disponible y utilizar las frecuencias relativas para estimar las probabilidades:

1. Para calcular la probabilidad de observar a_1 solo contamos con la información recogida en la base de datos (el vector de observaciones $\boldsymbol{C} = (C_{a_1}, ..., C_{a_k})$ y el tamaño N) y con la distribución a priori dada por el hiperparámetro λ. Por lo tanto estimaremos esta probabilidad utilizando la esperanza condicionada conforme a (5.24):

   $$Pr(a_1|db) = E(a_1|\boldsymbol{C}) = \frac{\lambda_{a_1}+C_{a_1}}{\Lambda+N}. \tag{5.35}$$

   Actualizamos la base de datos con esta observación: db' cuenta ahora con $\Lambda+N+1$ observaciones, y su vector de contadores $\boldsymbol{C'}$ está dado por:

   $$C'_{a_i} = \begin{cases} C_{a_i} & si \quad i = 1, \\ \\ C_{a_i}+1 & si \quad i \neq 1. \end{cases} \tag{5.36}$$

2. Calculamos ahora la probabilidad de la segunda observación a_2

   $$Pr(a_2|db, a_1) = Pr(a_2|db') = E(a_2|\boldsymbol{C'}) = \begin{cases} \frac{\lambda_{a_2}+C_{a_2}}{\Lambda+N+1} & si \quad a_1 = a_2, \\ \\ \frac{\lambda_{a_2}+C_{a_2}+1}{\Lambda+N+1} & si \quad a_1 \neq a_2. \end{cases} \tag{5.37}$$

3. El proceso continuaría iterativamente hasta completar las n observaciones.

Estratificación de la población. Subpoblaciones

Como ya hemos mencionado en la sección (3.2.2), la suposición de que los genotipos de los fundadores son obtenidos de la población de forma aleatoria es una simplificación que a menudo se aparta de la realidad. En la práctica, existen mucho factores que influyen en las probabilidades de que dos individuos tengan descendencia juntos. El hecho de que haya mayor tendencia a reproducirse dentro de un subconjunto determinado de la población puede alterar las probabilidades de observar los distintos alelos. Existen varios enfoques para lidiar con esta situación:

El enfoque más sencillo es asumir que la secuencia de alelos observados en $\mathcal{F}$ proviene de una población desconocida, es decir, para la que no tenemos base de datos. Bastaría con aplicar la expresión (5.33) asumiendo $N = 0$ y $\boldsymbol{C} = \mathbf{0}$. Sin embargo es habitual que una subpoblación comparta hasta cierto punto similaridades con otras conocidas, y esta información se desperdicia utilizando este mecanismo.

Otra opción es suponer que la subpoblación de interés es desconocida, pero que es similar a otras conocidas. Así si por ejemplo se conocen tres poblaciones A, B y C a las que se parece, podemos utilizar las bases de datos y los hiperparámetros de estas poblaciones para construir el hiperparámetro de la población de interés:

$$\lambda = \omega_A(\lambda_A + C_A) + \omega_B(\lambda_B + C_B) + \omega_C(\lambda_C + C_C), \tag{5.38}$$

donde ω_A se conoce como peso de la población A, y es el porcentaje de dicha población que esperamos encontrar de forma aleatoria en nuestra subpoblación de interés.

La tercera opción es, si se puede, asumir que la subpoblación de la que proceden los fundadores se parece mucho a una población conocida, lo suficiente como para partir de una estimación de las probabilidades alélicas basada en las probabilidades de esta última. Denotamos por $\widehat{\boldsymbol{q}}$ a este estimador. En estas circunstancias nos interesa medir el grado de separación o encapsulamiento de nuestra subpoblación. Para ello vamos a introducir un nuevo parámetro en la distribución a priori de las frecuencias alélicas en la subpoblación:

$$\lambda = \left(\frac{1}{\theta} - 1\right)\widehat{\boldsymbol{q}}, \qquad \theta \in (0, 1). \tag{5.39}$$

Cuanto mayor sea el parámetro θ, más próximo a 0 será el coeficiente que acompaña a $\widehat{\boldsymbol{q}}$, lo que significa que la subpoblación está más separada de la población de partida (sus individuos presentan mayor tendencia a relacionarse entre sí). Por tanto el coeficiente θ es un medidor del grado de separación entre el grupo de interés y el de referencia. Entonces sustituyendo en (5.33), teniendo en cuenta que $\Lambda = \sum_{i=1}^{n} \lambda_i = \frac{1}{\theta} - 1$, obtenemos:

$$Pr(\mathcal{F}) = \prod_{j=1}^{n} \frac{\left(\frac{1}{\theta} - 1\right) + b_j}{\left(\frac{1}{\theta} - 1\right) + j - 1} = \prod_{j=1}^{n} \frac{b_j\theta + (1-\theta)\,\widehat{\boldsymbol{q}}}{1 + (j-2)\,\theta}. \tag{5.40}$$

Y hemos obtenido la expresión correspondiente a la corrección theta (sección 3.2.2). Por tanto existe una conexión entre tener en cuenta la incertidumbre en las frecuencias alélicas al calcular $Pr(\mathcal{F})$ y la corrección theta.

Se puede explorar las consecuencias de este último enfoque, es decir, de la conexión entre la probabilidad a priori del vector de frecuencias y la corrección theta.

Por ejemplo, si se dispone de una base de datos pequeña es común calcular $Pr(\mathcal{F})$ utilizando una distribución Dirichlet con un parámetro de la forma $\lambda = \Lambda\widehat{\boldsymbol{q}}$. Se podría aprovechar la conexión calculando en su lugar la probabilidad bajo una corrección theta (esta opción está implementada en muchos tipo de software) de parámetro $\theta = \frac{1}{\Lambda+1}$ ya que utilizando la expresión (5.39):

$$\frac{1}{\theta} - 1 = \Lambda \Rightarrow \theta = \frac{1}{\Lambda + 1}. \tag{5.41}$$

En el sentido contrario, si lo que se desea es calcular la verosimilitud de los datos originales considerando un valor de θ no nulo, teniendo en cuenta que la influencia de θ sobre los datos se produce solo a través de $\mathcal{F}$:

$$Pr(\mathcal{D}|\mathcal{G}, \theta = \theta_0) = \sum_{\mathcal{F}} Pr(\mathcal{D}|\mathcal{F}, \mathcal{G}) Pr(\mathcal{F}|\theta = \theta_0). \tag{5.42}$$

Pero calcular $Pr(\mathcal{F}|\theta = \theta_0)$ es equivalente a calcular la probabilidad de $\mathcal{F}$ utilizando un probabilidad a priori dada por $\lambda = \left(\frac{1}{\theta_0} - 1\right)\widehat{\boldsymbol{q}}$ como acabamos de ver, entonces calculamos la probabilidad total (5.26):

$$Pr(\mathcal{F}|\theta = \theta_0) = \int_{\boldsymbol{p}} Pr(\mathcal{F}|\boldsymbol{p})\pi(\boldsymbol{p})d\boldsymbol{p}. \tag{5.43}$$

En vez de calcular la integral se pueden simular R vectores de frecuencias $\{\boldsymbol{p_1}, \boldsymbol{p_2}, ..., \boldsymbol{p_R}\}$ a partir de la distribución Dirichlet de parámetro $\lambda = \left(\frac{1}{\theta_0} - 1\right)\widehat{\boldsymbol{q}}$ y aproximar el resultado por el siguiente promedio:

$$Pr(\mathcal{F}|\theta = \theta_0) \approx \frac{1}{R}\sum_{j=1}^{R} Pr(\mathcal{F}|\boldsymbol{p_j}). \tag{5.44}$$

Teniendo en cuenta esta expresión y revirtiendo la descomposición hecha en (5.42), obtenemos una aproximación del resultado para la que no necesitamos más que cálculos sencillos:

$$Pr(\mathcal{D}|\mathcal{G}, \theta = \theta_0) \approx \frac{1}{R}\sum_{j=1}^{R} Pr(\mathcal{D}|\boldsymbol{p_j}, \mathcal{G}). \tag{5.45}$$

Nota 5.6. *En algunos casos el grupo de fundadores proviene de varias subpoblaciones. Si se conoce quien procede de cada una, se puede replicar el procedimiento anterior para cada grupo y después multiplicar las frecuencias obtenidas. A veces no se conoce exactamente quien pertenece a cada subpoblación, pero todavía podemos usar el razonamiento previo si contamos con las posibles asignaciones y las probabilidades a priori de cada una, sumando sobre todas las posibles combinaciones al calcular la probabilidad de cada genotipo $\mathcal{F}$ para los fundadores.*

Si estamos estudiando m marcadores independientes, debemos considerarlos de forma conjunta al calcular las probabilidades para cada posible genotipo de los fundadores, pues al introducir la incertidumbre del parámetro en los cálculos se induce dependencia entre ellos.

Es decir, si para el genotipo $\mathcal{F} = (f_1, ..., f_F)$, donde $f_i = (f_{i1}, ..., f_{im})$ es el conjunto de alelos en la posición i-ésima para cada marcador, y denotamos por $p(j, a)$ la probabilidad de observar el alelo a en el marcador j, obtendríamos:

$$Pr(\mathcal{F}) = \prod_{i=1}^{F} \prod_{j=1}^{m} p(j, f_{ij}). \tag{5.46}$$

5.2.2. Modelos a nivel de genealogía

Dada una genealogía $\mathcal{P}$, prefijado un genotipo para los fundadores $\mathcal{F}$ y conocido como obtener $Pr(\mathcal{F})$, esta sección se centrará en qué se puede decir ahora sobre el genotipo escalonado de las personas testadas $\mathcal{G}$. Es decir, el objetivo es determinar $Pr(\mathcal{G}|\mathcal{F})$.

Para calcular esta probabilidad habrá que sumar sobre todos los posibles esquemas de herencia, es decir, sobre todas las posibles formas en que los alelos de los fundadores se pueden trasmitir a lo largo de las generaciones hasta los individuos genotipados. Para tratar esta información, vamos en primer lugar a a establecer la cantidad de relaciones de paternidad/maternidad que hay en el linaje que estamos considerando $\{1, ..., R\}$. Ahora se define la siguiente variable sobre cada una de ellas: V_r para $r \in \{1, ..., R\}$, donde V_r indica si el alelo transmitido a la descendencia procede del cromosoma paterno o del cromosoma materno del progenitor. Es decir, nuevamente cobra relevancia el origen de cada alelo, una información difícil de obtener en la práctica. Entonces:

$$Pr(\mathcal{G}|\mathcal{F}) = \sum_{V_1,...V_R} Pr(\mathcal{G}, V_1, ..., V_R|\mathcal{F}) = \sum_{V_1,...V_R} Pr(\mathcal{G}|V_1, ..., V_R, \mathcal{F}) \prod_{r=1}^{R} Pr(V_r). \tag{5.47}$$

Ahora bien, si estamos considerando un único marcador autosómico (marcador i), podemos definir la variable de la siguiente manera:

$$V_r = \begin{cases} 0 \quad si \quad \text{se transmite el alelo materno del progenitor,} \\ \\ 1 \quad si \quad \text{se transmite el alelo paterno del progenitor,} \end{cases} \tag{5.48}$$

De esta forma, asumiendo equiprobabilidad a priori para ambas posibilidades en cada una de las relaciones $r \in \{1, ...R\}$

$$Pr(V_r = 0) = Pr(V_r = 1) = \frac{1}{2}, \tag{5.49}$$

y sustituyendo en (5.47), obtenemos la siguiente expresión simplificada para los genotipos de las personas testadas relativos al marcador i estudiado:

$$Pr(\mathcal{G}_i|\mathcal{F}_i) = \frac{1}{2^R} \sum_{V_1,...V_R} Pr(\mathcal{G}_i, V_1, ..., V_R|\mathcal{F}_i). \tag{5.50}$$

Si se considera el caso sin mutaciones, entonces la probabilidad condicionada anterior se calcula de forma muy sencilla. Será cero si el esquema de la herencia no es coherente con $\mathcal{G}$ (es

decir, si no hace llegar a los individuos cuyo ADN ha sido secuenciado los alelos presentes en su genotipo) y la unidad si lo es:

$$Pr(\mathcal{G}_i, V_1, ..., V_R|\mathcal{F}_i) = \begin{cases} 1 \quad si \quad \text{el reparto es coherente con } \mathcal{G}, \\ \\ 0 \quad si \quad \text{el reparto no es coherente con } \mathcal{G}. \end{cases} \tag{5.51}$$

Cuando la situación requiera el estudio de $\{1, ..., m\}$ marcadores independientes, la probabilidad global buscada puede calcularse como el producto de las probabilidades anteriores:

$$Pr(\mathcal{G}|\mathcal{F}) = \prod_{i=1}^{m} Pr(\mathcal{G}_i|\mathcal{F}_i). \tag{5.52}$$

Ejemplo 5.7. Recuperamos la situación presentada en la sección anterior, luego hemos fijado una genealogía $\mathcal{P}$. Supongamos ahora que contamos con un genotipo específico para los progenitores $\mathcal{F}$. Además que los genotipos de los de los individuos analizados coinciden con el resultado de las pruebas de ADN, luego $\mathcal{G}$ es conocido. Vamos a tratar de calcular su probabilidad condicionada a $\mathcal{F}$ y $\mathcal{P}$. Supongamos que el orden de los alelos dentro del genotipo para un marcador indica ahora su procedencia, el primero sería el alelo paterno y el segundo el alelo materno. Evidentemente, el orden de los alelos de los fundadores se fija arbitrariamente. La Figura 5.2 ilustra la situación del problema:

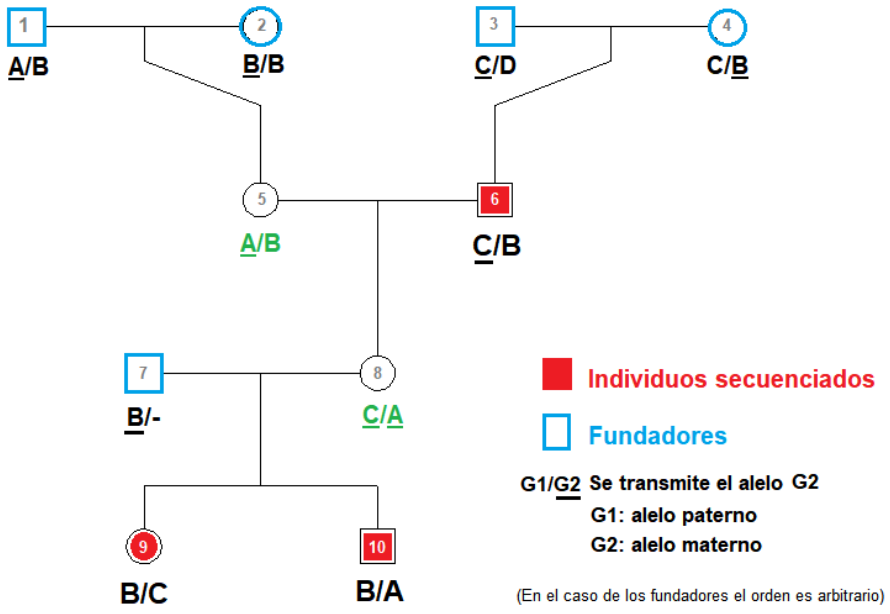

Figura 5.2: Ejemplo. Modelo a nivel de genealogía

- Definimos el número de relaciones de maternidad/paternidad: $R = 10$

  La numeración que se ha asignado está recogida en la Tabla 5.7. Las relaciones se denotan $i - j$ donde i es progenitor (padre/madre) de j.

- Definimos la variable indicativa del origen del alelo heredado en cada relación. Solo existen dos posibles esquemas de herencia compatibles con la situación descrita: el primero ($V_r, r \in \{1, ..., 10\}$) está recogido en la Figura 5.2. El segundo ($V'_r, r \in \{1, ..., 10\}$) es exactamente igual excepto porque 5 hereda el alelo materno de 2. Reflejamos las dos posibilidades en la tabla siguiente:

r	1	2	3	4	5	6	7	8	9	10
Relación	$1-5$	$2-5$	$3-6$	$4-6$	$5-8$	$6-8$	$7-9$	$7-10$	$8-9$	$8-10$
V_r	1	1	1	0	1	1	1	1	1	0
V'_r	1	0	1	0	1	1	1	1	1	0

  Por tanto para este marcador, aplicando la expresión (5.47) obtenemos:

$$Pr(\mathcal{G}|\mathcal{F}) = \frac{1}{2^{10}}\left[Pr(\mathcal{G}, V_1, ..., V_{10}|\mathcal{F}) + Pr(\mathcal{G}, V'_1, ..., V'_{10}|\mathcal{F})\right] = \frac{2}{2^{10}} = \frac{1}{2^9}. \tag{5.53}$$

5.2.3. Modelo a nivel observacional

Ahora calcularemos la probabilidad de obtener los datos $\mathcal{D}$ sabiendo el verdadero genotipo escalonado $\mathcal{G}$ de los individuos a los que hemos realizado análisis de ADN, es decir, $Pr(\mathcal{D} \mid \mathcal{G})$.

En principio cabría suponer que una prueba muestra tan solo los alelos que ya existen en el individuo, es decir:

$$Pr(\mathcal{D} \mid \mathcal{G}) = \begin{cases} 1 \quad si \quad \mathcal{D} \text{ coincide con } \mathcal{G}, \\ \\ 0 \quad si \quad \mathcal{D} \text{ no coincide con } \mathcal{G}. \end{cases} \tag{5.54}$$

Si tomamos el subconjunto $\{G_1, ..., G_J\}$ de $\mathcal{G}$ que hacen la probabilidad anterior 1 podemos sustituir en la ecuación (5.15):

$$Pr(\mathcal{D} \mid \mathcal{P}) = \sum_{\mathcal{G}} \sum_{\mathcal{F}} Pr(\mathcal{D} \mid \mathcal{G}) Pr(\mathcal{G} \mid \mathcal{F}), \mathcal{P}) Pr(\mathcal{F}) = \sum_{j=1}^{J} \sum_{\mathcal{F}} Pr(G_j \mid \mathcal{F}, \mathcal{P}) Pr(\mathcal{F}). \tag{5.55}$$

Sin embargo a la hora de secuenciar el ADN pueden producirse numerosos errores que ya hemos mencionado en capítulos previos: un alelo puede no ser detectado por el tipo de prueba escogida (alelo silencioso), fallar al amplificarse (dropout), puede haber una contaminación de la muestra (dropin) o un error en la secuenciación producida tanto por el método utilizado como por un error humano del investigador. Cualquiera de los casos anteriores puede desembocar en que los datos obtenidos $\mathcal{D}$ y el genotipo $\mathcal{G}$ no coincidan. En ese caso los cálculos de probabilidades requieren de la introducción de parámetros específicos para cada posible fuente de error. Hemos visto algunos ejemplos de como implementar estos parámetros en el capítulo 3.

Cabe destacar que los cálculos realizados para construir el modelo pueden plantearse, en cualquiera de las tres partes, también para genotipos no escalonados de $\mathcal{F}$ y $\mathcal{G}$. La única diferencia es que habrá un número inferior de cálculos pues hay situaciones que son diferentes teniendo en cuenta la procedencia de cada alelo, pero indistinguibles sin esta información.

Con esto cerramos el capítulo, pues ya hemos visto el planteamiento básico del modelo teórico. El próximo paso sería estudiar cómo estimar los parámetros involucrados en el modelo a partir de los datos. Finalmente, para tomar una decisión a partir de la información disponible, habría que aplicar resultados de teoría de la decisión.

Bibliografía

[1] Atkinson, E.; Mester, C.; Schaid, D.; Sinnwell, J. and Therneau, T., *Package 'kinship2'*, Human heredity 78(2), Karger Publishers, 91-93, 2014

[2] Byron, H., *Bayesian inference, Cran. R-project, LaplacesDemon Package*, Farmington, CT: Statisticat, LLC, 2013

[3] Dupuy, B. M.; Kling, D. and Stenersen, M., *Frequency data for 35 autosomal STR markers in a Norwegian, an East African, an East Asian and Middle Asian population and simulation of adequate database size*, Forensic Science International: Genetics Supplement Series 4(1), Elsevier, e378-e379, 2013

[4] Egeland, T., *Familias: Book, R version and Courses*, 2017, URL: `https://familias.name/book.html`

[5] Egeland, T.; Mevåg, B,; Mostad, P. and Stenersen, M., *Beyond traditional paternity and identification cases. Selecting the most probable pedigree*, Forensic Science International 110, 2000

[6] Egeland, T.; Mostad, P. and Simonsson, I., *Familias: Probabilities for Pedigrees Given DNA Data*, R package version 2.4. 2016, URL: `https://CRAN.R-project.org/package=Familias`

[7] Egeland, T.; Kling, D. and Mostad, P., *Relationship inference with familias and R: statistical methods in forensic genetics*, Academic Press, 2015

[8] Elston, R.C. and Stewart, J., *A general model for the genetic analysis of pedigree data*, Human heredity 21(6), Karger Publishers, 523-542, 1971

[9] Fung, W.K. and Hu, Y-Q., *Statistical DNA forensics: theory, methods and computation*, John Wiley & Sons, 2008

[10] Khlestkina, E. and Salina, E., *SNP markers: Methods of analysis, ways of development, and comparison on an example of common wheat*, Genetika 42, 725-36, 2006, DOI:10.1134/S1022795406060019

[11] Montes Suay, F., *Introducción a la Probabilidad*, Universitat de València. Departament d'Estadìstica i Investigació Operativa, 2007

[12] The Editors of Encyclopaedia Britannica, *Single nucleotide polymorphism*, Encyclopædia Britannica, inc., 2019 URL: https://www.britannica.com/science/single-nucleotide-polymorphism

[13] Tilanus, M.G.J., *Short tandem repeat markers in diagnostics: what's in a repeat?*, Leukemia 20(8), Nature Publishing Group, 1353-1355, 2006

[14] Vélez Ibarrola, R., *Cálculo de Probabilidades 2*, EDIA1 Ediciones Académicas, 2004

[15] Vigeland, M.D., *forrel: Forensic Pedigree Analysis and Relatedness Inference*, R package version 1.0.1, 2020, URL: https://CRAN.R-project.org/package=forrel

[16] Vigeland, M.D., *pedprobr: Probability Computations on Pedigrees*, R package version 0.3, 2020, URL: https://CRAN.R-project.org/package=pedprobr

[17] Vigeland, M.D., *pedtools: Creating and Working with Pedigrees and Marker Data*, 2020, URL:https://CRAN.R-project.org/package=pedtools

[18] Vigeland, M.D., *ribd: Pedigree-based Relatedness Coefficients*, R package version 1.1.0, 2020, URL: https://CRAN.R-project.org/package=ribd

[19] Walck C., *Hand-book on statistical distributions for experimentalists*, University of Stockholm 10, 2007

[20] Wikipedia contributors, *Bayesian inference - Wikipedia, The Free Encyclopedia*, 2020, URL: https://en.wikipedia.org/w/index.php?title=Bayesian_inference&oldid=965272883

[21] Wikipedia contributors. *Beta function - Wikipedia, The Free Encyclopedia*, 2020, URL: https://en.wikipedia.org/w/index.php?title=Beta_function&oldid=965363012

[22] Wikipedia contributors, *Conjugate prior - Wikipedia, The Free Encyclopedia*, 2020, URL: https://en.wikipedia.org/w/index.php?title=Conjugate_prior&oldid=966332810

[23] Wikipedia contributors, *Dirichlet distribution - Wikipedia, The Free Encyclopedia*, 2020, URL: https://en.wikipedia.org/w/index.php?title=Dirichlet_distribution&oldid=965052757

[24] Wikipedia contributors. *DNA profiling - Wikipedia, The Free Encyclopedia*, 2020, URL: https://en.wikipedia.org/w/index.php?title=DNA_profiling&oldid=964843229

[25] Wikipedia. *Genoma - Wikipedia, La enciclopedia libre*, 2020, URL: https://es.wikipedia.org/w/index.php?title=Genoma&oldid=127043119

[26] Wikipedia. *HBB - Wikipedia, La enciclopedia libre*, 2019, URL: https://es.wikipedia.org/w/index.php?title=HBB&oldid=120757015

[27] Wikipedia, *Polimorfismo de nucleótido único - Wikipedia, La enciclopedia libre*, 2020, URL: `https://es.wikipedia.org/w/index.php?title=Polimorfismo_de_nucle%C3%B3tido_%C3%BAnico&oldid=127366954`