USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Estadística para genética forense

Antía Vega Crego

2024/2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Estadística para genética forense

Antía Vega Crego

Junio, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa
Título: Estadística para genética forense
Breve descripción del contenido
El objetivo de este trabajo es estudiar los conceptos y métodos matemáticos empleados en el campo de la genética forense, en particular, en el estudio del parentesco.

Índice

Resumen

Los problemas de parentesco son un tema relevante dentro de la genética forense. En este trabajo se analizan los fundamentos matemáticos que permiten abordar estos problemas, en concreto el caso *standard trio* y el *standard duo*. Estos se exponen de manera detallada, acompañados de una presentación previa de conceptos de genética necesarios para entender la terminología de los problemas. Posteriormente se desarrollarán las nociones de probabilidad y estadística que permitirán modelizar dichas situaciones, así como resolverlas mediante distintos procedimientos de estimación y contrastes de hipótesis. Esta base matemática permitirá obtener conclusiones e interpretar los resultados genéticos de manera objetiva y coherente, garantizando su validez. Además, se presentan los paquetes *Familias* y *paramlink* del software **R**, con los que se podrán aplicar los métodos de resolución estudiados, resolviendo los ejemplos expuestos a lo largo del texto. Se trata de destacar la importancia de una buena base matemática a la hora de aplicarla en cualquier ámbito científico.

Abstract

Kinship analysis is a key area within forensic genetics. This paper explores the mathematical foundations necessary to address such cases, with a particular focus on the *standard trio* and *standard duo* scenarios. These cases are presented in detail, following an introduction to the essential genetic concepts required to understand the terminology. The work also develops the probabilistic and statistical notions needed to model these situations, enabling their resolution through various estimation techniques and hypothesis testing approaches. This rigorous mathematical framework supports objective and consistent interpretation of genetic results, ensuring their scientific validity. Additionally, the **R** software packages *Familias* and *paramlink* are introduced as practical tools for implementing the analytical methods discussed, with illustrative examples provided throughout. The overall aim is to emphasize the critical role of a strong mathematical foundation in the application of forensic genetics and other scientific disciplines.

Introducción

Las matemáticas son la base de todo estudio científico riguroso. Permiten no solo validar teorías y formalizar procesos complejos, sino también tomar decisiones fundamentadas; y pese a que, a menudo, su presencia pase desapercibida fuera del ámbito académico, se trata de una herramienta esencial en cualquier disciplina.

Aunque la genética forense pueda parecer un campo destacado únicamente en la biología, su aplicación práctica, sobre todo en el ámbito judicial donde el análisis de un perfil genético puede ser determinante para establecer relaciones de parentesco, exige una interpretación precisa de los datos genéticos. Cada conclusión, que a primera vista pueda parecer sencilla, como por ejemplo afirmar la probabilidad de paternidad o la coincidencia de dos perfiles genéticos, esconde una estructura matemática elaborada. Serán necesarias herramientas matemáticas como probabilidades, conceptos estadísticos y modelos matemáticos precisos, que permitan interpretar correctamente las pruebas biológicas analizadas.

Este trabajo pretende desarrollar y comprender en profundidad la base matemática bajo cada decisión aparentemente sencilla de los problemas de parentesco. No es suficiente el aplicar las fórmulas, es imprescindible comprender su procedencia, por qué se utilizan y las suposiciones que llevan implícitas; pues únicamente con un conocimiento en profundidad de la teoría subyacente será posible detectar errores, aplicar correctamente los procedimientos y garantizar que las conclusiones obtenidas sean justas y rigurosas; sobre todo en un campo como la genética forense donde estas conclusiones pueden tener consecuencias a gran alcance, como ocurre en los casos de paternidad.

Los matemáticos que trabajen en esta disciplina tienen la responsabilidad de verificar que las metodologías empleadas sean razonables, que los modelos se ajusten a la realidad y que todo este proceso esté respaldado por una base matemática lógica y coherente. El objetivo de este trabajo ha sido desarrollar los fundamentos de probabilidad y las herramientas estadísticas necesarias para la comprensión de los métodos empleados en la identificación de relaciones de parentesco en la genética forense. Se trata de crear una conexión entre la teoría matemática y su aplicación práctica, demostrando que solo desde un entendimiento riguroso de esta base matemática se

puede garantizar la fiabilidad de las conclusiones forenses.

El contenido se ha basado principalmente en el libro *Statistical DNA Forensics; Theory, Methods and Computation*, de Wing Kam Fung y Yue-Qing Hu [6], y se ha complementado con otras diversas fuentes que se pueden consultar en la bibliografía. La estructura de este trabajo se divide en cuatro capítulos. El primero de ellos es introductorio y recopila las nociones de genética y biología necesarias para poder seguir el desarrollo de la teoría del parentesco.

El segundo y tercer capítulos conforman la parte principal de este trabajo. En el segundo se desarrollan de manera rigurosa todos los conceptos de probabilidad y de estadística que serán necesarios a la hora de comprender los métodos de resolución empleados en los problemas de determinación del parentesco. Y en el tercero se explica la forma en la que se aplican estos conceptos matemáticos para resolver algunos de los problemas típicos de paternidad, como el caso *standard trio*, el *standard duo* o las probabilidades de exclusión; es decir, se relacionan las matemáticas con la genética, expresando los datos biológicos de manera que se puedan extraer conclusiones a partir de ellos. Se presentan además varios ejemplos que facilitarán la comprensión de los modelos teóricos y que se resolverán de forma manual inicialmente, y mediante el uso de las tecnologías en el último de los capítulos.

El cuarto y último capítulo presenta dos de los paquetes del software **R** más empleados en la genética forense. Estos son los paquetes *Familias* y *paramlink*, de los que se detallarán sus funciones e implementación, y que serán utilizados tanto para comprobar los resultados de los ejemplos resueltos a mano previamente, como para realizar cálculos más avanzados acerca de las probabilidades de exclusión.

Por último, recordar que con este trabajo se trata de proporcionar una visión básica pero rigurosa de la relación entre la genética y las matemáticas, intentando comprender en profundidad los conceptos que más se han detallado, y destacando la importancia de la base matemática a la hora de afrontar cualquier novedad en relación con la información genética.

Capítulo 1

Introducción a la genética

El objetivo de este primer capítulo es introducir los conceptos básicos del área de biología que serán necesarios en el desarrollo de este trabajo. Se presentan las nociones fundamentales sobre genética, necesarias para conocer la terminología empleada a lo largo del texto. Los contenidos están basados principalmente en los libros de Klug, Cummings y otros [7] y de Pierce [8].

Según se define en el Diccionario de la Lengua Española (DLE), la genética es una *parte de la biología que estudia los genes y la naturaleza y transmisión de los caracteres hereditarios.* Pero para entender esta definición debe especificarse lo que es un gen. Un *gen* es la unidad básica de información que "codifica" una característica heredada (como podrían ser el color de ojos o del pelo). A nivel molecular se definirá como una secuencia o fragmento de *ADN* (ácido desoxirribonucleico).

El ADN es un tipo de estructura molecular en la que está codificada la información genética; se trata de un conjunto de *nucleótidos* unidos entre sí que tiene forma de doble hélice y está compuesto por dos hebras de *bases nitrogenadas* complementarias. Las bases nitrogenadas son el compuesto más importante de los nucleótidos y, en el ADN, pueden ser de cuatro tipos distintos: adenina (A), timina (T), guanina (G) y citosina (C) y como se mencionaba previamente, se unirán de forma complementaria la timina de una hebra del ADN con la adenina de la otra hebra y la guanina con la citosina de la misma manera. Se puede ver la estructura de los nucleótidos en la figura 1.1. Cada uno recibe el nombre de la base que contiene.

Las distintas secuencias que puedan presentar los cuatro tipos de bases nitrogenadas será lo que codifique la información genética de cada ser vivo.

Los *cromosomas* son fragmentos de ADN en los que se organiza el material genético en los organismos. En las células humanas, que se componen de 23 pares de cromosomas, todos ellos a excepción de los sexuales son *homólogos*. Los cromosomas homólogos son aquellos que forman una

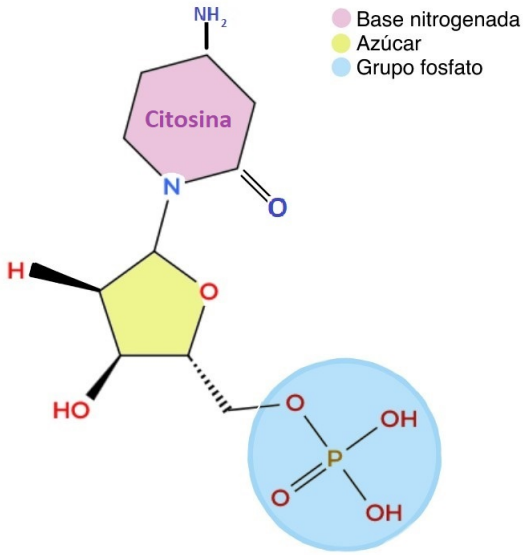

Figura 1.1: Estructura de un nucleótido con base nitrogenada citosina (C).

pareja por ser similares en tamaño y transportar información genética sobre el mismo conjunto de características hereditarias. Se trata de 22 pares de *autosomas* (cromosomas que no aportan información relativa al sexo del individuo) homólogos y un par formado por los cromosomas sexuales; en los que una de las copias del par de cromosomas procede de la madre, y la otra del padre, aunque se entremezclan generando así diferencias entre ellos.

La posición en el cromosoma de un gen particular se denomina *locus*; y a cada una de las posibles formas alternativas o secuencias que puedan existir para un gen se le denomina *alelo* (que se originan por las diferencias entre las dos copias de un par homólogo). Si un individuo posee dos alelos idénticos para un mismo gen, este se llamará *homocigótico* para ese gen; de no ser estos dos alelos idénticos, se dirá que el individuo es *heterocigótico* para ese gen.

Además, el *genotipo* de un individuo para un gen es el conjunto de alelos que este posee para ese gen concreto, con lo que el genotipo de un individuo será el conjunto de alelos que posea para su conjunto de genes; y al conjunto de posibles manifestaciones observables de una determinada característica es a lo que se le llama *fenotipo* (como puede ser el color de pelo).

Por ejemplo, Mendel experimentó con guisantes para estudiar varias de sus características, entre ellas el color de la semilla (amarillo o verde). El gen del color de la semilla consta solamente de dos alelos distintos a los que se denominará $\mathbf{Y}$ e $\mathbf{y}$, y se supone que el alelo $\mathbf{Y}$ es dominante sobre el $\mathbf{y}$. Entonces existirán tres posibles genotipos: $\mathbf{YY}$ (homocigoto dominante), $\mathbf{Yy}$ (heterocigoto) e $\mathbf{yy}$ (homocigoto recesivo), que expresarán dos posibles fenotipos: el fenotipo dominante, $\mathbf{Y}$, que se corresponde con el color amarillo y se obtiene con los genotipos $\mathbf{YY}$ e $\mathbf{Yy}$, y el fenotipo recesivo, $\mathbf{y}$, que se corresponde con el color verde y se obtiene con el genotipo $\mathbf{yy}$.

Finalmente, el *genoma* es el conjunto de toda la información hereditaria codificada en el ADN de un organismo. Por ejemplo, el genoma humano es el conjunto de todos los distintos genes que se encuentran presentes en los seres humanos.

En genética forense, se tratará con problemas en los que se verán involucrados distintos individuos y muestras de ADN. Por ejemplo, cuando se piden pruebas de parentesco entre padre e hijo en el caso de un juicio de paternidad, será el análisis de sus muestras de ADN el que permita, mediante técnicas estadísticas y de probabilidad, conocer su verdadera relación.

Se llamará *perfil de ADN* al genotipo del conjunto de *loci*[1] analizados. La técnica empleada para identificar individuos a partir de sus secuencias de ADN se denomina *obtención de perfiles o huellas de ADN*.
Los *marcadores genéticos* o marcadores de ADN son cualquier gen o fragmento corto de ADN cuya secuencia y ubicación son conocidas. En genética forense destacan dos tipos de marcadores:

- Los llamados microsatélites o *STRs*[2], que se obtienen mediante diversos procedimientos en el laboratorio, son fragmentos muy cortos de ADN que se repiten en tándem; fragmentos de entre dos y nueve pares de bases que se repiten entre siete y cuarenta veces. El número de repeticiones varía dependiendo del alelo, es decir, dependiendo de cada persona.

  Un ejemplo sencillo para entender esta definición aparece en el libro de Klug, Cummings y otros [7]. Se trata del STR en el locus D8S1179; formado por la secuencia de cuatro pares de bases nitrogenadas T C T A, repetido entre siete y veinte veces dependiendo del alelo.
  Se habla de pares de bases puesto que al componerse el ADN de dos hebras complementarias, siempre que se presente una de ellas, en este caso T C T A, en la otra hebra aparecerán las bases complementarias: A G A T. Así, un fragmento de ADN queda completamente determinado si se presenta únicamente una de sus hebras, pues la otra estará formada por las bases complementarias de la forma explicada previamente (A-T, C-G).
  Se conocen diecinueve posibles alelos del locus D8S1179 en la población. Un fragmento de una de las hebras de este locus se vería de la siguiente forma:

  T C T G T C T A T C T A T C T A T C T A T C T A T C T A T C T A T T C C

  Como se puede apreciar, en este ejemplo hay siete repeticiones de la secuencia T C T A.

- Los otros fragmentos de ADN más empleados en estos casos son los llamados polimorfismos de un solo nucleótido (*SNPs*[3]). La palabra *polimorfismo* se refiere a la cantidad de distintos alelos que puede presentar un gen y hace referencia a la variabilidad de este marcador genético. Los SNPs son lugares del genoma donde los distintos individuos de una especie

[1] *Loci* es el plural de locus.
[2] Del inglés, *short tandem repeats.*
[3] Del inglés, *single-nucleotide polymorphism.*

difieren en una única base nitrogenada.
Para ser un polimorfismo considerado SNP debe estar presente en al menos un 1 % de la población, es decir, que esa diferencia de una única base nitrogenada esté presente en al menos el 1 % de la población, en caso contrario se trataría de una variación genética rara. Hay SNPs que se asocian a otros; a este conjunto específico de SNPs (que están en alelos de loci fuertemente conectados) se le llama *haplotipo*. Además, estos distintos SNPs de un haplotipo están contenidos en un mismo cromosoma (o en una región de uno) y tienden a heredarse juntos.

Por ejemplo, siguiendo la ilustración de la definición de SNP del libro de Pierce [8], si se toman los cromosomas 1A, 1B y 1C, cada uno representando diferentes copias de un mismo cromosoma que se pueda encontrar en la población, entonces si la única diferencia entre ellos se produce en una única base nitrogenada se estará en presencia de un SNP (suponiendo que esta diferencia existe en al menos el 1 % de la población considerada). Se ilustra esto con los siguientes tres segmentos de ADN:

$$
\begin{aligned}
&1A: \ldots A\ C\ A\ C\ G\ C\ C \ldots T\ C\ G\ G\ G\ G\ T \ldots G\ T\ C\ G\ A\ C\ C \ldots \longrightarrow C\ G\ G \\
&1B: \ldots A\ C\ A\ C\ G\ C\ C \ldots T\ C\ G\ A\ G\ G\ T \ldots G\ T\ C\ A\ A\ C\ C \ldots \longrightarrow C\ A\ A \\
&1C: \ldots A\ C\ A\ T\ G\ C\ C \ldots T\ C\ G\ G\ G\ G\ T \ldots G\ T\ C\ A\ A\ C\ C \ldots \longrightarrow T\ G\ A
\end{aligned}
$$

Los tres SNPs en este caso se corresponderían con las tres columnas de bases destacadas en color rojo que, como se pude observar, se trata de un mismo fragmento de un locus donde los distintos individuos A, B y C difieren únicamente en una base (en el primer SNP el individuo C difiere por poseer una timina en lugar de una citosina, en el segundo se tiene en el individuo B una adenina en lugar de una guanina, y en el tercer SNP el individuo A presenta una guanina en lugar de una adenina).

Asimismo, se tendrían los siguientes haplotipos, que fueron definidos e indicados anteriormente:

A: C G G B: C A A C: T G A

Una *mutación* es un cambio en la información genética que se produce a la hora de la herencia, ya sea entre células o de los progenitores a la descendencia. Para entender mejor esto se ilustra con una adaptación de un ejemplo que aparece en el libro de Klug, Cummings y otros [7], donde se explica de manera sencilla este concepto.
Una mutación se daría por ejemplo cuando en el proceso de replicación se origina un cambio de base, esta es en concreto una mutación puntual. Para verlo mejor, si la secuencia original es de la forma G G G A G T G T A (se muestra solo una de las hebras) y la nueva secuencia mutada es de la forma G G G A T T G T A, esto implicará que se ha producido una mutación puntual por el hecho de aparecer en la nueva secuencia de ADN una timina en lugar de una guanina.

Capítulo 2

Probabilidad

En este segundo capítulo se introducen las herramientas de probabilidad necesarias para poder estudiar la estadística de la genética forense desde un punto de vista formal. Estos contenidos se desarrollan principalmente a partir de las obras de Vélez Ibarrola [12] y Evett y Weir [4]. Además, se entrará en el área de la inferencia estadística para tratar con los contrastes de hipótesis, para lo que se han tomado como guía los libros de Vélez y García [13] y de Rohatgi y Ehsanes Saleh [11], y, como apoyo, el libro de Fisz [5].

Los experimentos de interés serán aquellos de los cuales se conocen todos los posibles resultados, pero no se puede predecir uno concreto para cada realización. Estos serán los llamados *experimentos aleatorios*, que se pueden repetir en condiciones idénticas de forma indefinida.
Cuando se trata con un experimento aleatorio, lo primero que se necesita es conocer su *espacio muestral*, que es el conjunto de todos sus posibles resultados y que se denota por Ω. Cualquier subconjunto $A \subset \Omega$ recibirá el nombre de *suceso*.

Para poder definir la probabilidad será necesario introducir el concepto de $\sigma-$*álgebra*, que proporciona la estructura necesaria para poder definir la aplicación correctamente.

Definición 2.1. Dado un espacio muestral Ω, se llamará $\sigma-$álgebra de Ω a aquella familia de subconjuntos $\mathcal{A} \subset \Omega$ que verifique las siguientes condiciones:

1. El conjunto vacío, pertenece a la $\sigma-$álgebra, $\mathcal{A}$. ($\emptyset \in \mathcal{A}$).

2. Para cualquier suceso A de la $\sigma-$álgebra $\mathcal{A}$ se verifica que su complementario, A^c, también está contenido en ella. ($\forall A \in \mathcal{A},\ A^c \in \mathcal{A}$).

3. Dada una familia numerable de sucesos, $\{A_n\}_{n\in\mathbb{N}}$, $A_n \in \mathcal{A}$, se tiene que su unión pertenece también a la $\sigma-$álgebra: $\bigcup_n A_n \in \mathcal{A}$.

Como consecuencia inmediata se tiene que $\Omega \in \mathcal{A}$ y además, por las leyes de De Morgan[1], si se tiene $\{A_n\}_{n\in\mathbb{N}}$, con $A_n \in \mathcal{A}$; entonces $\cap_n A_n \in \mathcal{A}$.
En un espacio muestral finito, la $\sigma-$álgebra considerada será la familia de conjuntos $\mathcal{P}(\Omega)$.
Con esto, se tiene que $(\Omega, \mathcal{A})$ es un espacio medible o probabilizable; es decir, se podrá asociar a cada suceso de $\mathcal{A}$ una probabilidad, que será una medida de la incerteza de que, al realizar el experimento, ocurra un suceso $A \in \mathcal{A}$.

A lo largo de la historia de las matemáticas destacan dos maneras de evaluar la incertidumbre de un suceso, esto es, de calcular la posibilidad de que ocurra un suceso $A \in \mathcal{A}$:

- La regla *de Laplace* se basa en un experimento donde el espacio muestral es finito y se supone equiprobabilidad de las soluciones; entonces la probabilidad de que ocurra el suceso, A, se calcula como:

$$\text{Probabilidad de } A = \frac{\text{Número de casos favorables al suceso } A}{\text{Número de casos posibles}}.$$

  Pero este método implica la suposición de equiprobabilidad de posibles resultados y solo se puede aplicar si el espacio muestral es finito.

  Ejemplo 2.2. En genética forense, un posible espacio de probabilidad sería el mismo del experimento de Mendel con el color de los guisantes mencionado anteriormente. Si se toma como espacio muestral, Ω, el conjunto formado por las cuatro combinaciones de alelos posibles para la descendencia (teniendo en cuenta que solo se tienen dos alelos **Y** e **y**): **YY**, **Yy**, **yY** e **yy**. Cabe destacar que el genotipo **Yy** y el **yY** producen el mismo fenotipo. Se tendría que el conjunto de sucesos, $\mathcal{A}$, podría ser por ejemplo el de fenotipos, esto es, el color de la semilla: amarillo o verde. El color amarillo se da siempre que aparezca el alelo **Y** en el genotipo y el verde únicamente con el genotipo **yy**. Así, las probabilidades asociadas calculadas mediante la regla de Laplace vendrían dadas por 1/4 para el suceso "ser de color verde", esto es la probabilidad del genotipo **yy** (un caso favorable de entre cuatro posibles, pues hay cuatro posibles genotipos), y 3/4 para el suceso "ser de color amarillo", es decir, el número de genotipos que producen una semilla de color amarillo (el **YY**, el **yY** y el **Yy**) entre el número total de genotipos (cuatro).

- Existe un segundo enfoque para medir la incertidumbre de un suceso, que es más intuitivo y se podría denominar *frecuentista*. Este se basa en las frecuencias de ocurrencia de cada suceso al realizar el mismo experimento un amplio número de veces, N. Se calcularía entonces la frecuencia del suceso $A \in \mathcal{A}$ como:

$$\text{Frecuencia relativa de } A = \frac{\text{Número de veces que ha ocurrido el suceso } A}{\text{Número de veces que se ha realizado el experimento}}.$$

[1] $\mathrm{P}((A \cup B)^c) = \mathrm{P}(A^c \cap B^c)$; $\mathrm{P}((A \cap B)^c) = \mathrm{P}(A^c \cup B^c)$.

Usando la ley de estabilidad de las frecuencias se dice que, cuando el número de veces que se realiza el experimento, N, tiende a infinito, el número en el que se estabiliza la frecuencia relativa será lo que se entiende como la probabilidad del suceso:

$$\text{Probabilidad de } A = \lim_{N\to\infty} \frac{n}{N},$$

siendo n el número de veces que ha ocurrido el suceso $\mathcal{A}$ en las N repeticiones del experimento.

Con esto cabe destacar que en algunos casos sería necesario realizar numerosas repeticiones del experimento para poder alcanzar esta estabilización de las frecuencias, lo que no resulta práctico y podría suponer un elevado coste, tanto económico como temporal.

En 1933 el matemático Andréi Kolmogórov proporcionó una definición formal de probabilidad, que permitirá realizar los cálculos de probabilidades en todos los casos.

Definición 2.3 (**Kolmogórov**)**.** Dados un espacio muestral y una $\sigma-$álgebra, $(\Omega, \mathcal{A})$, se define una *probabilidad* como una aplicación $\mathrm{P} : \mathcal{A} \longrightarrow [0,1]$ que verifica los siguientes axiomas:

1. Dada cualquier colección numerable de sucesos $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$, que sean disjuntos entre sí ($A_n \cap A_m = \emptyset,\ \forall\ n \neq m$), se tiene que

$$\mathrm{P}\left(\bigcup_n A_n\right) = \sum_n \mathrm{P}(A_n).$$

2. $\mathrm{P}(\Omega) = 1$.

Se llama *espacio de probabilidad* al conjunto $(\Omega, \mathcal{A}, \mathrm{P})$.

2.1. Probabilidad condicionada

Ejemplo 2.4. Se tiene que dos familias distintas han denunciado la desaparición de su hijo y que la policía ha encontrado un niño perdido. Los padres de la familia 1 tienen grupos sanguíneos A y 0, y los de la familia 2 tienen los grupos A y AB. Se consideran un suceso y su complementario:

$$\begin{cases} C: & \text{“el niño pertenece a la familia 1”} \\ C^c: & \text{“el niño pertenece a la familia 2”.} \end{cases}$$

Con estos datos, y sin más información, se puede decir que la probabilidad inicial (o a priori) de que el niño pertenezca a cada una de las dos familias es de un medio:

$$\mathrm{P}(C) = \frac{1}{2} = \frac{\text{Número de casos favorables (pertenencia a la familia 1)}}{\text{Número de casos posibles (pertenencia a la familia 1, pertenencia a la familia 2)}}.$$

$$\mathrm{P}(C^c) = 1 - \mathrm{P}(C) = \frac{1}{2}.$$

Tras analizar el grupo sanguíneo del niño se encuentra que es de tipo AB y se define un nuevo suceso,

$$D : \text{“ el grupo sanguíneo del niño es AB”}.$$

Sabiendo que es imposible obtener como herencia el grupo AB siendo los de los progenitores el A y el 0, se tiene entonces que la probabilidad de que el niño sea hijo de la familia 1 es ahora cero:

$$\mathrm{P}(C \mid D) = 0.$$

Así, la probabilidad del suceso complementario (la pertenencia del niño a la familia 2) es 1, pues de los grupos A y AB sí es posible obtener descendencia del grupo AB. Con esto se aprecia que la información obtenida tras el análisis sanguíneo del niño, es decir, el hecho de ocurrir el suceso D, ha condicionado la probabilidad del suceso C.

Este ejemplo permite ilustrar el hecho de que la probabilidad asociada a un suceso se pueda ver afectada por el acontecimiento de otro suceso del mismo espacio muestral.

Definición 2.5. Sea $(\Omega, \mathcal{A}, \mathrm{P})$ un espacio de probabilidad, y sea $B \in \mathcal{A}$ un suceso con $\mathrm{P}(B) > 0$. Para cada suceso $A \in \mathcal{A}$ se define la *probabilidad de A condicionada por B* como:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

De esto se deduce de manera inmediata que:

$$\mathrm{P}(A \cap B) = \mathrm{P}(B)\mathrm{P}(A \mid B). \tag{2.1}$$

A continuación se exponen los tres resultados más significativos en los que la probabilidad condicionada desempeña un papel fundamental.

En el primero de estos tres resultados se generaliza la fórmula (2.1) para un conjunto numerable de sucesos.

Proposición 2.6 (**Regla del producto**). *Sean $A_1, \ldots, A_n$ sucesos de un espacio de probabilidad $(\Omega, \mathcal{A}, P)$, y cumpliéndose $P(A_1 \cap \ldots \cap A_{n-1}) > 0$; entonces se tiene que*

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \ldots \cap A_{n-1}).$$

Demostración. El resultado se obtiene aplicando reiteradamente la igualdad (2.1); en efecto:

$$\mathrm{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathrm{P}\left(\bigcap_{i=1}^{n-1} A_i\right) \mathrm{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) =$$

$$= \mathrm{P}\left(\bigcap_{i=1}^{n-2} A_i\right) \mathrm{P}\left(A_{n-1} \mid \bigcap_{i=1}^{n-2} A_i\right) \mathrm{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right).$$

Y continuando iterativamente se llega a la expresión deseada. □

El segundo resultado expone una manera de expresar la probabilidad de un suceso empleando las probabilidades condicionadas.

Teorema 2.7 (**Ley de las probabilidades totales**). *Dado un espacio de probabilidad* $(\Omega, \mathcal{A}, P)$*, si se tiene* $\{A_n\} \subset \mathcal{A}$ *una partición numerable del espacio muestral* Ω *tal que* $P(A_n) > 0 \quad$ *para* $n = 1, 2, \ldots$*; entonces para cualquier suceso del espacio muestral,* B*, se puede expresar su probabilidad como:*

$$P(B) = \sum_{n\in\mathbb{N}} P(A_n)P(B \mid A_n). \tag{2.2}$$

Demostración. Por ser $\{A_n\}$ una partición de Ω se tiene que los sucesos A_n son disjuntos entre sí, y que $\cup_{n\in\mathbb{N}} A_n = \Omega$. Con esto, el segundo axioma de la definición de probabilidad de Kolmogórov y la igualdad (2.1):

$$\mathrm{P}(B) = \mathrm{P}\left(\bigcup_{n\in\mathbb{N}} (B \cap A_n)\right) = \sum_{n\in\mathbb{N}} \mathrm{P}(B \cap A_n) = \sum_{n\in\mathbb{N}} \mathrm{P}(A_n)\mathrm{P}(B \mid A_n).$$

□

Por último, el tercero de los resultados permite expresar la probabilidad *a posteriori* del suceso A_i, es decir, su probabilidad asignada tras ocurrir el suceso B; empleando la probabilidad *a priori* del suceso A_i (sin estar condicionado por ningún otro suceso) y las probabilidades condicionadas del suceso B por los A_i.

Teorema 2.8 (**Regla de Bayes**). *Si se tienen* A, B *dos sucesos en un espacio de probabilidad* $(\Omega, \mathcal{A}, P)$ *con* $P(B) > 0$*, entonces*

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}.$$

Aplicando ahora este resultado a una partición $\{A_i\} \subset \mathcal{A}$ *del espacio muestral* Ω*; cumpliendo* $P(A_i) > 0, \forall i \in \{1, \ldots, n\}$*; y empleando la ley de las probabilidades totales* (2.2)*, se cumple que:*

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{n\in\mathbb{N}} P(A_n)P(B \mid A_n)}, \quad i = 1, \ldots, n.$$

Demostración. Basta ver que intercambiando los sucesos A y B en la definición de probabilidad condicionada (Definición 2.5) se puede reescribir la igualdad (2.1) como:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B \mid A), \tag{2.3}$$

ya que $A \cap B = B \cap A$. De donde, igualando (2.1) y (2.3),

$$\mathrm{P}(A \mid B)\mathrm{P}(B) = \mathrm{P}(B \mid A)\mathrm{P}(A).$$

Obteniéndose así la fórmula de la regla de Bayes. □

Definición 2.9. Dado un espacio de probabilidad, $(\Omega, \mathcal{A}, \mathrm{P})$ y sea $A \in \mathcal{A}$ un suceso de la $\sigma-$álgebra; se define la *odds (ventaja a favor) de* A como:

$$O(A) = \frac{\mathrm{P}(A)}{\mathrm{P}(A^c)} = \frac{\mathrm{P}(A)}{1 - \mathrm{P}(A)}.$$

Reordenando esta fórmula, se pueden calcular probabilidades a partir de la odds:

$$\mathrm{P}(A) = \frac{O(A)}{1 + O(A)}.$$

Ejemplo 2.10. En un partido de baloncesto, sea A el suceso "que gane el equipo local", con probabilidad $\mathrm{P}(A) = 2/3$; entonces "que gane el equipo visitante", tendrá una probabilidad $\mathrm{P}(A^c) = 1/3$. Se tiene entonces que la odds de A es: $O(A) = \dfrac{2/3}{1/3} = 2$. Es decir, la ventaja a favor del equipo local es 2 a 1.
Si en cambio se considera el lanzamiento de un dado, donde la probabilidad del suceso B : "obtener un 3" es $\mathrm{P}(B) = 1/6$, y la probabilidad de "no obtener un 3" es $\mathrm{P}(B^c) = 5/6$, se tiene que la odds de B es: $O(B) = (1/6)/(5/6) = 1/5$. Es decir, la ventaja a favor de "obtener un 3" es de 1 a 5.

2.2. Independencia

Cuando en un espacio de probabilidad, dados dos sucesos $A, B \in \mathcal{A}$, el hecho de que ocurra el suceso A, no provoca ningún cambio en la probabilidad del suceso B, entonces la probabilidad de B coincidirá con su probabilidad condicionada por el suceso A:

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(B \cap A)}{\mathrm{P}(A)} = \mathrm{P}(B).$$

Esto significará que $\mathrm{P}(B \cap A) = \mathrm{P}(B) \cdot \mathrm{P}(A)$ y en este caso se dirá que los dos sucesos son independientes. Seguidamente se define este concepto de manera formal para una colección numerable de sucesos:

Definición 2.11. Dado un espacio de probabilidad, $(\Omega, \mathcal{A}, \mathrm{P})$, y dada una colección de sucesos $A_1, \ldots, A_n \in \mathcal{A}$; se dice que son *independientes* si y solo si:

$$\mathrm{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathrm{P}(A_1) \cdot \ldots \cdot \mathrm{P}(A_n).$$

Además, se dirá que son *mutuamente independientes* si:

$$\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i) \cdot \mathrm{P}(A_j), \text{ con } i \neq j, \ \forall\, i, j = 1, \ldots, n.$$

En el ámbito de la genética, la independencia es importante a la hora de analizar los distintos marcadores. Será habitual tratar con poblaciones en *Equilibrio de Hardy-Weinberg* (EHW); esta ley indica que los alelos de un locus particular del genotipo estudiado son mutuamente independientes, lo que simplifica mucho los cálculos. Esta ley se puede escribir de la siguiente manera.

Proposición 2.12 (**Ley de Hardy-Weinberg**). *Considerando una población suficientemente grande, en la que los emparejamientos son aleatorios y que no se vea afectada por mutaciones, migración ni selección natural, es decir, que no esté sujeta a ninguna fuerza evolutiva, entonces:*

1. *Las probabilidades alélicas de la población no varían con el tiempo.*
2. *Considerando dos alelos de un locus, **Y**e **y**, entonces las probabilidades genotípicas se estabilizan después de una generación de emparejamiento aleatorio en las proporciones p_1^2 para el genotipo **YY**, $2p_1p_2$ para el genotipo **Yy**, y p_2^2 para el **yy**; donde p_1 es la probabilidad del alelo **Y** y p_2 la del alelo **y**.*

Demostración. Se consideran emparejamientos aleatorios dentro de una población suficientemente grande. Se tienen en un locus dos alelos, $\mathbf{Y}$ e $\mathbf{y}$ (es decir, en una de las copias del par homólogo se tiene el alelo $\mathbf{Y}$ y en la otra copia, en el mismo lugar, se tiene el alelo $\mathbf{y}$), con lo que se sabe que, en la población, los tres posibles genotipos serán: $\mathbf{YY}$, $\mathbf{Yy}$ e $\mathbf{yy}$; y se supone que sus probabilidades en dicha población son $\mathrm{P}(\mathbf{YY}) = p_{11}$, $\mathrm{P}(\mathbf{Yy}) = p_{12}$ y $\mathrm{P}(\mathbf{yy}) = p_{22}$ respectivamente. Entonces las probabilidades de los alelos $\mathbf{Y}$ e $\mathbf{y}$ serán, en la población, respectivamente:

$$\begin{aligned} p_1 = \mathrm{P}(\mathbf{Y}) &= \mathrm{P}(\mathbf{Y} \mid \mathbf{YY}) \cdot \mathrm{P}(\mathbf{YY}) + \mathrm{P}(\mathbf{Y} \mid \mathbf{Yy}) \cdot \mathrm{P}(\mathbf{Yy}) + \mathrm{P}(\mathbf{Y} \mid \mathbf{yy}) \cdot \mathrm{P}(\mathbf{yy}) = \\ &= 1 \cdot p_{11} + (1/2) \cdot p_{12} + 0 \cdot p_{22} = p_{11} + p_{12}/2, \end{aligned} \tag{2.4}$$

$$\begin{aligned} p_2 = \mathrm{P}(\mathbf{y}) &= \mathrm{P}(\mathbf{y} \mid \mathbf{yy}) \cdot \mathrm{P}(\mathbf{yy}) + \mathrm{P}(\mathbf{y} \mid \mathbf{Yy}) \cdot \mathrm{P}(\mathbf{Yy}) + \mathrm{P}(\mathbf{y} \mid \mathbf{YY}) \cdot \mathrm{P}(\mathbf{YY}) = \\ &= 1 \cdot p_{22} + (1/2) \cdot p_{12} + 0 \cdot p_{11} = p_{22} + p_{12}/2. \end{aligned} \tag{2.5}$$

En la siguiente tabla se muestran las probabilidades asociadas a cada pareja de genotipos, siendo M el genotipo de la madre y P el genotipo del padre.

	P		
M	**YY**	**Yy**	**yy**
YY	p_{11}^2	$p_{11}p_{12}$	$p_{11}p_{22}$
Yy	$p_{12}p_{11}$	p_{12}^2	$p_{12}p_{22}$
yy	$p_{22}p_{11}$	$p_{22}p_{12}$	p_{22}^2

Tabla 2.1: Probabilidades para un emparejamiento aleatorio en una población suficientemente grande.

Además, se pueden ver en la tabla 2.2 los posibles genotipos para un descendiente de estas combinaciones madre-padre (D: descendencia, M-P: progenitores (madre-padre)).

	M-P								
D	**YY YY**	**YY Yy**	**YY yy**	**Yy YY**	**Yy Yy**	**Yy yy**	**yy YY**	**yy Yy**	**yy yy**
YY	1	1/2	0	1/2	1/4	0	0	0	0
Yy	0	1/2	1	1/2	1/2	1/2	1	1/2	0
yy	0	0	0	0	1/4	1/2	0	1/2	1

Tabla 2.2: Genotipos para la descendencia.

Sean P^*_{11}, P^*_{12} y P^*_{22} las probabilidades asociadas a los genotipos **YY**, **Yy** e **yy** de la segunda generación respectivamente. De las tablas 2.1 y 2.2 se pueden obtener estas probabilidades a partir de las probabilidades de los genotipos de los progenitores. Por ejemplo:

$$P^*_{12} = (1/2)\cdot p_{11}p_{12} + 1\cdot p_{11}p_{22} + (1/2)\cdot p_{12}p_{11} + (1/2)\cdot p_{12}^2 + (1/2)\cdot p_{12}p_{22} + 1\cdot p_{22}p_{11} + (1/2)\cdot p_{22}p_{12}.$$

Es decir, $P^*_{12} = 2p_1p_2$.
Análogamente, se obtiene que $P^*_{11} = p_1^2$ y que $P^*_{22} = p_2^2$. Con lo que las probabilidades genotípicas de la descendencia quedan completamente determinadas por las probabilidades alélicas de los progenitores.
Ahora, mediante las ecuaciones (2.4) y (2.5) se pueden expresar las probabilidades alélicas en la segunda generación como:

$$p_1^* = P^*_{11} + P^*_{12}/2 = p_1^2 + p_1p_2 = p_1, \quad p_2^* = 1 - p_1^* = 1 - p_1 = p_2.$$

Pues cabe recordar que $p_1 + p_2 = 1$ por haber solo dos alelos. Y con esto se tiene que:

$$P^*_{11} = p_1^{*2}, \quad P^*_{12} = 2p_1^*p_2^* \quad \text{y} \quad P^*_{22} = p_2^{*2},$$

lo que demuestra la ley de Hardy-Weinberg. □

2.3. Variables aleatorias

2.3.1. Caso unidimensional

En genética forense son importantes y de uso frecuente los contrastes de hipótesis; pero para poder comprenderlos será necesario introducir primeramente algunos conceptos acerca de las variables aleatorias.

Definición 2.13. Dado un espacio de probabilidad, $(\Omega, \mathcal{A}, \mathrm{P})$, se llama *variable aleatoria* a una función $X : \Omega \longrightarrow \mathbb{R}$ que verifique:

$$X^{-1}(B) \in \mathcal{A},\ \forall B \in \mathcal{B},\ \text{con } X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in B\} \subset \Omega.$$

Donde $\mathcal{B}(\mathbb{R})$ es la $\sigma-$*álgebra de Borel de* $\mathbb{R}$, que está formada por todos los intervalos de $\mathbb{R}$ que se pueden construir como unión o intersección de intervalos de la forma $(-\infty, x]$ con $x \in \mathbb{R}$.

Para poder manejar las probabilidades de las variables aleatorias se emplean las denominadas funciones de distribución.

Definición 2.14. Dado un espacio de probabilidad, $(\Omega, \mathcal{A}, \mathrm{P})$ y una variable aleatoria, X, definida en este espacio; se denomina *función de distribución* de la variable aleatoria X a la siguiente función, $F : \mathbb{R} \longrightarrow [0, 1]$, que asocia a cada $x \in \mathbb{R}$ un valor $F(x) \in [0, 1]$:

$$F(x) = \mathrm{P}\{\omega \in \Omega : X(\omega) \leq x\}.$$

Para simplificar la notación se empleará $\mathrm{P}(X \leq x) = \mathrm{P}\{\omega \in \Omega : X(\omega) \leq x\}$. Además, dependiendo de las características de esta función de distribución, se diferencian dos tipos de variables aleatorias:

1. Si existe un conjunto finito o infinito numerable, $D \subset \mathbb{R}$, tal que $\mathrm{P}\{\omega \in \Omega : X(\omega) \in D\} = \mathrm{P}(X \in D) = 1$; entonces la variable aleatoria X se denomina *discreta*. En ese caso la función que asigna

   $$p(x) = \mathrm{P}\{\omega \in \Omega : X(\omega) = x\} = \mathrm{P}(X = x),\ \forall x \in D$$

   se llama *función de masa de probabilidad* de X.
   Nótese que $p(x) \geq 0,\ \forall x \in D$ y $\sum_{x \in D} p(x) = 1$. Además, se verifica que:

   $$F(x) = \sum_{x_i \leq x} p(x_i),\ \text{con } x_i \in D,\ \text{y } x \in \mathbb{R}.$$

2. Si existe una función $f : \mathbb{R} \longrightarrow \mathbb{R}$ no negativa, verificando:

   $$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t = \mathrm{P}(X \leq x),\ \forall x \in \mathbb{R},$$

   entonces la variable aleatoria X se denomina *continua*, y la función f se llama *función de densidad* de X. Además se verifica que:

   $$\int_{-\infty}^{\infty} f(t)\mathrm{d}t = 1.$$

A continuación se presenta una medida de la localización de los valores que toma la variable aleatoria, dentro de su rango de valores posibles.

Definición 2.15. Dada una variable aleatoria, X, en un espacio de probabilidad $(\Omega, \mathcal{A}, \mathrm{P})$, con función de distribución F se define su *esperanza matemática, o media,* $E[X]$, de la siguiente manera:

- Cuando la variable aleatoria es discreta y supuesto que $\sum_k |x_k|\mathrm{P}(X = x_k) < \infty$:

$$E[X] = \sum_k x_k \cdot p(x_k), \text{ siendo } \{x_1, x_2, \ldots, x_k, \ldots\} \text{ los valores posibles de } X.$$

- Cuando la variable aleatoria es continua y supuesto que $\int_{-\infty}^{\infty} |x| f(x)\mathrm{d}x < \infty$:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x, \text{ siendo } f \text{ la función de densidad de } X.$$

Ejemplo 2.16. Un caso sencillo de variable aleatoria unidimensional es aquel en el que la variable toma únicamente dos valores: $D = \{0, 1\}$ (fracaso o éxito), y cuya función de probabilidad está definida como:

$$p(x) = \begin{cases} \theta, & \text{si } x = 1 \text{ (probabilidad de éxito)}, \\ 1 - \theta, & \text{si } x = 0 \text{ (probabilidad de fracaso)}, \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Se dice que esta variable aleatoria, X, sigue una distribución Bernoulli de parámetro $\theta \in [0, 1]$, $Bernoulli(\theta)$, entonces su función de probabilidad será de la forma:

$$p(x) = \theta^x \cdot (1 - \theta)^{1-x}, \text{ con } x \in \{0, 1\}.$$

Además, su función de distribución viene dada por:

$$F(x) = \begin{cases} 0, & \text{si } x < 0, \\ 1 - \theta, & \text{si } 0 \leq x < 1, \\ 1, & \text{si } x \geq 1. \end{cases}$$

2.3.2. Caso multidimensional

En alguna ocasión se querrán estudiar varias variables de forma conjunta; por ejemplo, cuando se desea realizar un estudio sobre el "peso" y la "estatura" en una población. Surge entonces, con naturalidad, el concepto de variable aleatoria multidimensional o *vector aleatorio*.

Definición 2.17. Sea $(\Omega, \mathcal{A}, \mathrm{P})$ un espacio de probabilidad; un *vector aleatorio* es un $X = (X_1, \ldots, X_n) : \Omega \longrightarrow \mathbb{R}^n$ que verifique:

$$X^{-1}(B) \in \mathcal{A}, \forall B \in \mathcal{B}^n, \text{ con } X^{-1}(B) = \{\omega \in \Omega | X(\omega) = (X_1(\omega), \ldots, X_n(\omega)) \in B\} \subset \Omega.$$

Donde $\mathcal{B}^n$ es la $\sigma-$álgebra de Borel en $\mathbb{R}^n$, que se define de forma análoga al caso unidimensional como la $\sigma-$álgebra formada por las uniones e intersecciones de los intervalos de $\mathbb{R}^n$ de la forma $(-\infty, x]$ con $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, es decir, los conjuntos de la forma $(-\infty, x_1] \times \ldots \times (-\infty, x_n]$.

A continuación se presenta la definición de función de distribución de un vector aleatorio:

Definición 2.18. Dado un vector aleatorio de $\mathbb{R}^n$, $X = (X_1, \ldots, X_n)$, su *función de distribución conjunta* será la aplicación $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ que verifique que $F(x) = \mathrm{P}(X \leq x)$, siendo $x \in \mathbb{R}^n$; es decir:

$$F(x_1, \ldots, x_n) = \mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) \ \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Si se tiene un vector aleatorio $(X_1, \ldots, X_n)$, entonces para cada $i = 1, \ldots, n$, X_i es una variable aleatoria con función de distribución *marginal* F_i, para todo $i \in \{1, \ldots, n\}$. Donde

$$F_i(x_i) = \lim_{x_j \to \infty} F(x_1, \ldots, x_i, \ldots, x_n) \ \forall j = 1, \ldots, n \text{ con } j \neq i.$$

Esto es, $F_i(x_i) = \lim\limits_{\substack{x_j \to \infty \\ j \neq i}} \mathrm{P}(X_1 \leq x_1, \ldots, X_i \leq x_i, \ldots, X_n \leq x_n) = \mathrm{P}(X_i \leq x_i) \ \forall i = 1, \ldots, n.$

La independencia de variables aleatorias resultará de gran interés, al igual que ocurre con la independencia de sucesos.

Definición 2.19. Sea $(X_1, \ldots, X_n)$ un vector aleatorio y sean F la función de distribución conjunta y $\{F_k\}_{k=1}^n$ las distribuciones marginales respectivas de $X_1, \ldots, X_n$.
Se dice que $X_1, \ldots, X_n$ son *mutuamente independientes* si, y solo si:

$$F(x_1, \ldots, x_n) = \prod_{k=1}^{n} F_k(x_k), \ \forall \ (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Ejemplo 2.20. Se tiene un vector aleatorio $(X_1, \ldots, X_n)$ donde $X_1, \ldots, X_n$ son variables aleatorias independientes con distribución $Bernoulli(\theta)$. Es decir, la función de distribución de cada X_i es:

$$F_i(x_i) = \begin{cases} 0, & \text{si } x_i < 0, \\ 1 - \theta, & \text{si } 0 \leq x_i < 1, \\ 1, & \text{si } x_i \geq 1. \end{cases}$$

Entonces :

$$F(x) = \prod_{i=1}^{n} F_i(x_i) = \prod_{i=1}^{n} (1 - \theta)^{1 - x_i}.$$

Además, su función de masa de probabilidad será:

$$\mathrm{P}(X = x) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i}.$$

2.4. Estimación

Cuando se quiere conocer una característica acerca de un conjunto de individuos como, por ejemplo, el "colesterol" en la población adulta de Galicia, o la proporción de personas que consiguen empleo al terminar los estudios universitarios en Galicia, no será posible recabar los datos

de todos de los individuos implicados. Por ejemplo, encuestar en un corto período de tiempo a todas las personas que han terminado estudios universitarios en Galicia no es factible. Este conjunto de individuos sobre el que se quiere estudiar una característica se denomina *población*. La solución a este problema será analizar un subconjunto de la población, llamado *muestra*, de modo que la información contenida en ella permita "inferir" información sobre la población.

De forma general, se tiene una variable aleatoria X, que describe la población que se está estudiando, y cuya función de distribución es $F(X;\theta)$, que es una función conocida pero dependiente de un parámetro desconocido $\theta \in \Theta$ que se quiere estimar; donde $\Theta \in \mathbb{R}$ es lo que se llama *espacio paramétrico*. Por ejemplo, si se quiere conocer el "colesterol" en la población adulta de Galicia, la variable aleatoria X será el "colesterol", y el parámetro desconocido θ que se quiere estimar será la media de "colesterol" en la población adulta de Galicia.

Sean $X_1, \ldots, X_n$ n variables aleatorias independientes e idénticamente distribuidas, es decir, que tienen todas la misma distribución F de X (la variable que se está estudiando), que representan las n realizaciones del experimento aleatorio (la muestra aleatoria). Y sea $f(x_1, \ldots, x_n; \theta)$, con $(x_1, \ldots, x_n) \in \mathbb{R}^n$ su función de masa de probabilidad o función de densidad conjunta. Entonces, cada variable aleatoria X_i tendrá función de densidad, o función de masa de probabilidad, $f(x_i, \theta)$. El vector aleatorio $(X_1, \ldots, X_n)$ es lo que se conoce como *muestra aleatoria simple* de tamaño n de X. El número n es el llamado *tamaño muestral*, y el resultado $(x_1, \ldots, x_n)$ obtenido tras realizar el experimento se denomina una *realización muestral*.

Ejemplo 2.21. Se quiere estudiar la proporción del alumnado de la Universidad de Santiago de Compostela que estudia en la biblioteca en la época de exámenes finales.
El experimento consiste en preguntar a cada individuo si estudia en la biblioteca en la época de exámenes finales o no. Con lo que se tiene la variable aleatoria:

$$X = \begin{cases} 1, & \text{si la respuesta del individuo es "sí",} \\ 0, & \text{si la respuesta del individuo es "no".} \end{cases}$$

Se define entonces una muestra aleatoria de tamaño n, $(X_1, \ldots, X_n)$, donde cada $F_i(x_i)$ es una $Bernoulli(\theta)$, es decir, las funciones de distribución marginales dependen de un parámetro $\theta \in [0, 1]$. Se tiene entonces $f(x_i;\theta) = p(x_i) = \theta^{x_i} \cdot (1-\theta)^{1-x_i}$. Y el parámetro θ representa la proporción de individuos de la población que estudian en la biblioteca en época de exámenes finales y es el parámetro que se quiere *estimar* (se calcula de forma aproximada a partir de la información obtenida de la muestra).

Definición 2.22. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple con función de distribución $F(X_1, \ldots, X_n; \theta)$. Se define el *estimador*, $\hat{\theta}$ de θ, como una función de la muestra que permite conocer el valor aproximado del parámetro θ desconocido de la población.
Se denominará *estimación* de θ al valor numérico de $\hat{\theta}$ obtenido para una realización muestral.

Para el ejemplo anterior, un posible estimador para el parámetro θ sería la proporción de alumnos de la muestra que estudian en la biblioteca en la época de exámenes finales; que se calcula como

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\text{número de alumnos de la muestra que estudian en la biblioteca}}{\text{tamaño de la muestra}}.$$

2.4.1. Estimación por máxima verosimilitud

Se considera ahora la siguiente situación: se quiere averiguar la proporción de individuos con ojos azules de una población. Si se extrae una muestra de diez individuos de esta población y se encuentra que uno de ellos tiene los ojos azules, es evidente que la proporción en la población no puede ser cero, pues en la muestra hay una proporción de 0,1 de individuos con los ojos azules. Pero si ahora hubiese que decidir si la proporción vale 0,2 o 0,8 en la población, habría que hacer un análisis un poco más en profundidad. Si valiese 0,8 significaría que hay una probabilidad del 80 % de que cada individuo de la población tenga los ojos azules; en el otro caso, 0,2, implicaría que hay una probabilidad del 20 % de que cada individuo de la población tenga los ojos azules. Parece entonces más verosímil, ante el resultado muestral obtenido, que el valor de la proporción en la población sea de 0,2, pues es lo más "cercano" a 0,1.

Este ejemplo permite introducir el concepto de verosimilitud, y esta misma idea es la empleada en el llamado *principio de máxima verosimilitud*.

Definición 2.23. Dada $(X_1, \ldots, X_n)$ una muestra aleatoria simple con distribución $F(X_1, \ldots, X_n; \theta)$, con $\theta \in \Theta$, y fijada la realización muestral $x_1, \ldots, x_n$, se llama *función de verosimilitud* a $f(x_1, \ldots, x_n; \theta)$, considerada como función del parámetro θ.

Dada una realización muestral $(x_1, \ldots, x_n)$ y dado $\theta \in \Theta$, la función de verosimilitud se calcula como:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Como se puede deducir del ejemplo anterior, se buscará el valor del parámetro θ que maximice esta función de verosimilitud.

Definición 2.24. Se define el *estimador de máxima verosimilitud* de θ como aquel $\hat{\theta}(X_1, \ldots, X_n)$ que maximiza la función de verosimilitud; es decir, que satisface la ecuación:

$$f(x_1, \ldots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} f(x_1, \ldots, x_n; \theta) = \max_{\theta \in \Theta} \prod_{i=1}^{n} f(x_i; \theta).$$

Dado que la función logaritmo es monótona creciente estrictamente, se trabajará con el logaritmo de la función de verosimilitud, lo que permitirá estudiar el máximo de una suma en

lugar del máximo de un producto, lo que resultará más sencillo. Con esto se tendrá que:

$$\log\left[\max_{\theta\in\Theta} f(x_1,\ldots,x_n;\theta)\right] = \max_{\theta\in\Theta} \log[f(x_1,\ldots,x_n;\theta)] = \\ = \max_{\theta\in\Theta} \log\left(\prod_{i=1}^{n} f(x_i;\theta)\right) = \max_{\theta\in\Theta} \sum_{i=1}^{n} \log f(x_i;\theta).$$

2.4.2. Contrastes de hipótesis

Para introducir los contrastes de hipótesis se presenta un ejemplo ilustrativo que sirve de ayuda para comprender los conceptos básicos de los mismos.

Ejemplo 2.25. Supóngase que ahora se quiere comprobar si la proporción de alumnos de la Universidad de Santiago de Compostela que estudian en la biblioteca es mayor que el 50 % es decir, si ha aumentado y es necesario aumentar el número de puestos disponibles, o si, por el contrario, sigue siendo menor o igual que el 50 % :

$$\begin{cases} H_0: & \theta \leq 0{,}5 \\ H_a: & \theta > 0{,}5. \end{cases}$$

En este caso se tienen entonces dos hipótesis, la *hipótesis alternativa*, H_a, que se quiere comprobar, esta hipótesis sería "la proporción es mayor al 50 %"; y su complementaria, la *hipótesis nula*, H_0, que es el dato conocido previamente y que se acepta mientras no se demuestre que ha ocurrido un cambio: "la proporción es menor o igual al 50 %".

La hipótesis nula goza de "presunción de inocencia", pues es considerada cierta salvo que se demuestre lo contrario, es decir, H_0 se acepta, y se trata de determinar si existen evidencias suficientes a partir de la muestra como para poder rechazarla. En ese caso se habrá probado la hipótesis alternativa y se dirá que existen pruebas significativas a favor de H_a. Esto se explica más formalmente en la siguiente definición.

Definición 2.26. Una hipótesis paramétrica es una afirmación acerca del parámetro desconocido θ, por ejemplo, se denota por hipótesis nula: $H_0 : \theta \in \Theta_0$, donde $\Theta_0 \subset \Theta$. La hipótesis alternativa será $H_a : \theta \in \Theta - \Theta_0$.
Si el conjunto Θ_0, o respectivamente $\Theta - \Theta_0$, contiene un punto únicamente, se dice que la hipótesis H_0, o respectivamente H_a, es una *hipótesis simple*. En caso contrario, se dirá que es una *hipótesis compuesta*.

Los contrastes de hipótesis conllevan un problema de decisión que se representa mediante la Tabla 2.3. Se llama *Error de tipo I* al error cometido cuando se rechaza la hipótesis nula siendo esta cierta; y *Error de tipo II* al cometido por aceptar la hipótesis nula siendo esta falsa.

	Decisión	
Realidad	Aceptar	Rechazar
H_0 es cierta	Correcto	Error tipo I
H_0 es falsa	Error tipo II	Correcto

Tabla 2.3: Problema de decisión.

Definición 2.27. La probabilidad de cometer un error de tipo I al realizar un contraste de hipótesis se denota *nivel de significación* y viene dada por:

$$\alpha = \mathrm{P}(\text{Rechazar } H_0 | H_0 \text{ es cierta}).$$

La probabilidad de detectar que una hipótesis es falsa se denomina *potencia*, β, y se tiene por lo tanto la probabilidad de cometer un error de tipo II:

$$\beta = \mathrm{P}(\text{Rechazar } H_0 | H_0 \text{ es falsa}) = 1 - \mathrm{P}(\text{Aceptar } H_0 | H_0 \text{ es falsa}).$$

Con esto, se tienen dos errores distintos que minimizar. Si se quiere asegurar que el nivel de significación α sea cero, basta con aceptar siempre la hipótesis nula H_0, pero esto implicaría en muchas ocasiones cometer un error de tipo II considerable; es decir, en numerosas situaciones se aceptaría la hipótesis nula siendo esta falsa.

Una forma de minimizar ambos errores es fijar un nivel de significación α y seleccionar un criterio para resolver el contraste de forma que proporcione la mayor potencia posible, β, pues esto implica el menor error de tipo II posible.

En este proceso de un contraste de hipótesis, se debe tener lo que se conoce como un *estadístico de contraste*, que va a ser función de la muestra aleatoria simple y que debe reflejar si los datos muestrales son más compatibles con la hipótesis nula o con la alternativa.

En una primera fase, si se fija el nivel de significación, α, se puede dividir el espacio muestral en dos regiones disjuntas, denominadas *región de aceptación*, C, y *región de rechazo*, C^c:

- Si el valor obtenido de aplicar el estadístico de contraste a la realización muestral $(x_1, \ldots, x_n)$ pertenece a la región de rechazo, C, entonces se rechaza la hipótesis nula H_0. Esto significará que los datos proporcionan evidencias suficientes como para demostrar la hipótesis alternativa H_a. Al haber sido fijado previamente el nivel de significación α se debe construir esta región C de forma que su probabilidad bajo H_0, esto es, su probabilidad suponiendo que se verifica la hipótesis nula, sea como mucho α.
- Si el valor obtenido al aplicar el estadístico de contraste a la realización muestral pertenece a la región de aceptación C^c, entonces se acepta la hipótesis nula, H_0. Esta región es la complementaria de la región de rechazo y cabe destacar que el hecho de aceptar H_0 no

significa que se haya demostrado su veracidad, tan solo significa que no se han encontrado pruebas suficientes para demostrar la hipótesis alternativa H_a, por lo que se acepta H_0.

Los contrastes de este tipo donde el espacio muestral se divide en dos regiones son los denominados *test no aleatorizados*.

En los *test aleatorizados* se tiene en cuenta la potencia del contraste de hipótesis, y tras obtenerse la realización muestral no se toma directamente la decisión de aceptar o rechazar la hipótesis nula, sino que se emplea un mecanismo aleatorio para tomar esa decisión. Este nuevo mecanismo surge del objetivo de que el nivel de significación, α, coincida con el supremo de la función de potencia bajo la hipótesis nula, $\sup_{\theta\in\Theta_0} \beta(\theta)$, es decir, $\sup_{\theta\in\Theta_0} \beta(\theta) = \alpha$, con la finalidad de encontrar un equilibrio en el que tanto el error de tipo I como el error de tipo II sean lo más pequeños posible.

Definición 2.28. Sea χ el espacio muestral formado por todos los posibles valores para la muestra aleatoria simple $(X_1, \ldots, X_n)$, $\chi \subset \mathbb{R}^n$, y sea $x = (x_1, \ldots, x_n)$ una realización muestral. Se define la *función crítica* asociada a un test como una función medible,

$$\varphi : \chi \longrightarrow [0, 1]$$

es decir, una función que verifique que: $\forall B \in \mathcal{B}[0, 1]$, $\varphi^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$.

Con esto, en los test aleatorizados, tras la observación de la realización muestral $(x_1, \ldots, x_n)$, se debe efectuar un sorteo con probabilidad $\varphi(x_1, \ldots, x_n)$ de rechazar H_0 y probabilidad $1 - \varphi(x_1, \ldots, x_n)$ de aceptarla.

Nótese que, en el caso anterior, es decir, para un test no aleatorizado, esta función crítica solo tomará los valores 0 (si el resultado muestral se encuentra en la región de aceptación) o 1 (si el resultado está en la región de rechazo). Es decir:

$$\varphi(x_1, \ldots, x_n) = \begin{cases} 1, & \text{si } (x_1, \ldots, x_n) \in C, \\ 0, & \text{si } (x_1, \ldots, x_n) \in C^c. \end{cases}$$

Se trata de la "función indicadora" de la región crítica.

A continuación se define la *función de potencia*.

Definición 2.29. Sea θ el parámetro del que depende la distribución poblacional sobre el que se quiere realizar un contraste. Sea $H_0 : \theta \in \Theta_0$ la hipótesis nula, y sea $H_a : \theta \in \Theta - \Theta_0$ la hipótesis alternativa. Se define la función de potencia de un test como la función:

$$\begin{aligned} \beta : \Theta &\longrightarrow [0, 1] \\ \theta &\longrightarrow \beta(\theta) = \mathrm{P}(\text{Rechazar } H_0) = E_\theta[\varphi(X_1, \ldots, X_n)] \end{aligned}$$

En el caso discreto, cuando la variable aleatoria X es discreta y tiene función de masa de probabilidad $f(x;\theta)$, esta función de potencia adopta la siguiente forma:

$$\beta(\theta) = \sum_{x\in\chi} \varphi(x) f(x;\theta),$$

donde $x = (x_1, \ldots, x_n)$ y $f(x;\theta)$ es la probabilidad de que ocurra x bajo θ.
En el caso continuo, cuando la variable aleatoria X es continua, la función de potencia se escribe como sigue:

$$\beta(\theta) = \int_{\chi} \varphi(x) f(x;\theta) \mathrm{d}x,$$

donde, de nuevo, $x = (x_1, \ldots, x_n)$ y $f(x;\theta)$ es el valor de la función de densidad dado θ.

En el caso de los test no aleatorizados, la función de potencia es la probabilidad asociada a la región de rechazo, $\mathrm{P}_\theta(C)$, pues recordando que, para los test no aleatorizados la función crítica se define como:

$$\varphi(x_1, \ldots, x_n) = \begin{cases} 1, & \text{si } (x_1, \ldots, x_n) \in C, \\ 0, & \text{si } (x_1, \ldots, x_n) \in C^c, \end{cases}$$

Entonces la función de potencia será:

- En el caso discreto:

$$\beta(\theta) = \begin{cases} \sum_{x\in\chi} 1 \cdot f(x;\theta), & \text{si } (x_1, \ldots, x_n) \in C, \\ \sum_{x\in\chi} 0 \cdot f(x;\theta), & \text{si } (x_1, \ldots, x_n) \in C^c, \end{cases}$$

  Esto es, vale 0 en caso de que la realización muestral pertenezca a la región de aceptación, y toma el valor de la función de densidad en el caso en el que la realización muestral se encuentre en la región crítica, es decir, la probabilidad asociada a la región de rechazo.

- En el caso continuo, de forma análoga, la función de potencia toma el valor 0 en el caso en que la realización muestral se encuentra en la región de aceptación y toma como valor la función de densidad en el caso en que la realización muestral está en la región crítica; es decir, de nuevo se trata de la probabilidad asociada a la región de rechazo.

Definición 2.30. Sea $\theta \in \Theta$ el parámetro del que depende la distribución teórica y acerca del cual se realiza el contraste de hipótesis:

- Se dice que un test es de nivel de significación α si se cumple que: $\beta(\theta) \leq \alpha \quad \forall\theta \in \Theta_0$.
- Se denomina *tamaño del test* al valor numérico: $\sup_{\theta\in\Theta_0} \beta(\theta)$.

2.4.3. Test de razón de verosimilitudes

La idea de este test se basa en estimar los parámetros por el método de máxima verosimilitud, comparando los resultados de la función de verosimilitud cuando $\theta \in \Theta_0$ con los resultados obtenidos cuando $\theta \in \Theta$.

Definición 2.31. Dado un contraste de hipótesis $H_0 : \theta \in \Theta_0$ frente a $H_a : \theta \in \Theta - \Theta_0$; se define la *razón de verosimilitudes* como el cociente:

$$\lambda(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} f(x_1, \ldots, x_n; \theta)}{\sup_{\theta \in \Theta} f(x_1, \ldots, x_n; \theta)}.$$

Como $\Theta_0 \subset \Theta$, $0 \leq \lambda(x_1, \ldots, x_n) \leq 1$, y si la hipótesis nula fuese cierta se esperaría que el cociente tenga un valor próximo a 1. Si por el contrario H_0 fuese falsa se observaría un valor del cociente más próximo a cero, lo que indicaría una mayor discrepancia entre "lo observado" (hipótesis alternativa) y "lo esperado", esto es, lo que se da por cierto (la hipótesis nula).

Definición 2.32. Para un contraste de hipótesis de la forma: $H_0 : \theta \in \Theta_0$ frente a $H_a : \theta \in \Theta - \Theta_0$, se define el *test de razón de verosimilitudes* como:

$$\varphi(x_1, \ldots, x_n) = \begin{cases} 1, & \text{si } \lambda(x_1, \ldots, x_n) < c, \\ \gamma, & \text{si } \lambda(x_1, \ldots, x_n) = c, \\ 0, & \text{si } \lambda(x_1, \ldots, x_n) > c. \end{cases}$$

Donde $c \in (0, 1)$ y $\gamma \in [0, 1]$ se seleccionan de forma que el test tenga tamaño α.

En el caso en que ambas hipótesis, nula y alternativa, sean simples: $H_0 : \theta = \theta_0$; $H_a : \theta = \theta_1$, la razón de verosimilitudes se reduce a:

$$\lambda(x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n; \theta_0)}{\text{máx}[f(x_1, \ldots, x_n; \theta_0), f(x_1, \ldots, x_n; \theta_1)]},$$

con lo que el hecho de comparar $\lambda(x_1, \ldots, x_n) < c$ (con $c < 1$) equivale a hacer la comparación:

$$\frac{f(x_1, \ldots, x_n; \theta_1)}{f(x_1, \ldots, x_n; \theta_0)} > k = \frac{1}{c}.$$

Con esto quedan presentados los conceptos de probabilidad y estadística que serán aplicados en el capítulo siguiente. Los problemas de parentesco que se van a tratar en el próximo capítulo se resolverán mediante una aplicación a la genética forense de los resultados sobre probabilidad y estadística vistos en este capítulo. Los contrastes de hipótesis que se harán se basan en el test de razón de verosimilitudes, y se verá que las conclusiones se realizan directamente a partir de la razón de verosimilitudes.

Capítulo 3

Parentesco

En este capítulo se analizarán diferentes tipos de problemas relacionados con el parentesco en genética forense. Primero se estudiará el conocido como *standard trio*, en el que se trata de averiguar si un hombre, del que se posee su genotipo, es el padre biológico de un niño, con base en los datos aportados por el genotipo materno y el del propio niño.

Seguidamente se estudiará el caso en el que se desconoce el genotipo materno. Además, se presentará una forma de excluir de la posible paternidad a un hombre elegido al azar en la población. Se emplearán los recursos y conceptos presentados previamente, siguiendo principalmente la obra de Fung y Hu [6], el libro de Egeland, Kling y Mostad [3] y el de Evett y Weir [4].

En este capítulo se trabajará siempre bajo la hipótesis de una población en *equilibrio de Hardy-Weinberg*. Como se explicó previamente, esta ley implica que los alelos de un locus son mutuamente independientes. Además, considerando dos alelos de un locus, Y_i e Y_j, cuyas probabilidades en la población son p_i y p_j respectivamente, entonces según la ley de Hardy-Weinberg las probabilidades genotípicas (de aquellos genotipos formados por estos dos alelos) son: p_i^2 para el genotipo Y_iY_i, $2p_ip_j$ para el genotipo Y_iY_j y p_j^2 para el genotipo Y_jY_j.

3.1. Coincidencia de dos "muestras" (perfiles de ADN)

Este es uno de los casos más sencillos para introducir las técnicas de resolución de los problemas de parentesco. Se supone una población en *equilibrio de Hardy-Weinberg* (EHW). Se tiene el perfil de ADN de un individuo concreto de la población, S, cuyo genotipo[1] para el marcador genético que se está estudiando es $G_S = Y_iY_j$, donde Y_i e Y_j son dos alelos diferentes de este

[1]Con "genotipo" se está denotando el genotipo del perfil de ADN disponible, que en este caso está formado por un solo marcador genético, no el genotipo completo del individuo.

marcador (si $i \neq j$); y se analiza una muestra de ADN, E, cuyo genotipo para este marcador es $G_E = Y_iY_j$. Se quiere conocer si el perfil de ADN obtenido de la muestra, E, procede del sujeto S; o si, por el contrario, pertenece a otra persona diferente, no relacionada, de la población considerada.

Se plantea, por lo tanto, el siguiente contraste de hipótesis:

$$\begin{cases} H_0: & \text{Los genotipos completos de } S \text{ y } E \text{ coinciden} \\ H_a: & \text{Los genotipos completos de } S \text{ y } E \text{ no coinciden.} \end{cases}$$

Con este contraste se quiere comprobar si el perfil de ADN analizado, E, no coincide con el del sujeto S de la población, es decir, si, a pesar de coincidir los genotipos para el marcador concreto que se estudia, no se trata de la misma persona.

Por la ley de Hardy-Weinberg, las probabilidades $\mathrm{P}(G_S = Y_iY_j)$ y $\mathrm{P}(G_E = Y_iY_j)$ serán $2p_ip_j$, si $i \neq j$, o p_i^2, si $i = j$. Siendo p_i la probabilidad del alelo Y_i en la población, y p_j la del alelo Y_j.

La probabilidad del genotipo observado para el sujeto S, en la población, es independiente del hecho de que el perfil de ADN, E, coincida con S (H_0), o no (H_a); por lo tanto:

$$\mathrm{P}(G_S = Y_iY_j \mid H_0) = \mathrm{P}(G_S = Y_iY_j \mid H_a) = \mathrm{P}(G_S = Y_iY_j) = \begin{cases} 2p_ip_j, & \text{si } i \neq j, \\ p_i^2, & \text{si } i = j. \end{cases}.$$

Además, en el caso en el que se verifica la hipótesis alternativa, H_a, cuando la persona de la muestra analizada, E, no es el sujeto, S, se tiene que la probabilidad de ambos genotipos es independiente. Y como también, el genotipo G_E es independiente del hecho de que se verifique la hipótesis nula o la alternativa, se tiene que:

$$\mathrm{P}(G_E = Y_iY_j \mid G_S = Y_iY_j, H_a) = \mathrm{P}(G_E = Y_iY_j \mid H_a) = \mathrm{P}(G_E = Y_iY_j)$$

Y como se verifica la ley de Hardy-Weinberg, esta probabilidad viene dada por:

$$\mathrm{P}(G_E = Y_iY_j \mid G_S = Y_iY_j, H_a) = \begin{cases} p_i^2, & \text{si } i = j, \\ 2p_ip_j, & \text{si } i \neq j, \end{cases}$$

donde p_i es la probabilidad en la población del alelo Y_i y p_j es la del alelo Y_j.

En cambio, cuando se verifica la hipótesis nula, H_0, se está suponiendo que la persona propietaria de la muestra de ADN, E, es el sujeto S, con lo que:

$$\mathrm{P}(G_E = Y_iY_j \mid G_S = Y_iY_j, H_0) = 1.$$

Una vez destacados estos hechos, se presenta la razón de verosimilitudes (*LR*, *Likelihood Ratio*) en este problema. En este caso, las funciones de verosimilitud son funciones de masa de

probabilidad, esto es, se trabajará con cocientes de probabilidades. Asimismo, se trata de un contraste de dos hipótesis simples, por lo que se tendrá la ecuación:

$$LR = \frac{\mathrm{P}(G_E = Y_iY_j, G_S = Y_iY_j \mid H_0)}{\mathrm{P}(G_E = Y_iY_j, G_S = Y_iY_j \mid H_a)}.$$

Donde la función de verosimilitud del numerador es la probabilidad de obtener los perfiles de ADN analizados, bajo la hipótesis nula H_0, y la función de verosimilitud del denominador se corresponde con la probabilidad de obtener dichos perfiles bajo la hipótesis alternativa H_a.

Y empleando las condiciones descritas anteriormente y la regla del producto, se tiene que la razón de verosimilitudes es:

$$LR = \frac{\mathrm{P}(G_E = Y_iY_j \mid G_S = Y_iY_j, H_0)}{\mathrm{P}(G_E = Y_iY_j \mid G_S = Y_iY_j, H_a)} \cdot \frac{\mathrm{P}(G_S = Y_iY_j \mid H_0)}{\mathrm{P}(G_S = Y_iY_j \mid H_a)} = \frac{\mathrm{P}(G_E = Y_iY_j \mid G_S = Y_iY_j, H_0)}{\mathrm{P}(G_E = Y_iY_j \mid H_a)}.$$

Finalmente, como la hipótesis H_0 dice que la muestra pertenece al sujeto S, se tiene que el numerador en esta ecuación es igual a 1 y la probabilidad del denominador viene definida por la ley de Hardy-Weinberg (como se ha explicado previamente), con lo que la razón de verosimilitudes quedaría:

$$LR = \begin{cases} 1/p_i^2, & \text{si } i = j \\ 1/2p_ip_j, & \text{si } i \neq j. \end{cases}$$

Este valor da lugar al siguiente test de razón de verosimilitudes:

$$\varphi(x) = \begin{cases} 1, & \text{si } LR < c, \\ \gamma, & \text{si } LR = c, \\ 0, & \text{si } LR > c. \end{cases}$$

que es un test aleatorizado donde $c \in (0, 1)$ y $\gamma \in [0, 1]$ se seleccionan de forma que el test tenga "tamaño α". El método habitual para resolver este contraste de hipótesis se basa en analizar directamente la razón de verosimilitudes.

Si LR es menor que c ($c < 1$), se tiene que el perfil obtenido será más favorable a la hipótesis alternativa; es decir, se cumplirá la condición $LR < c$ del test de razón de verosimilitudes por lo que será más probable el rechazo de la hipótesis nula, o expresándolo de otra forma, existirán pruebas a favor de la hipótesis alternativa H_a. Si por el contrario, LR resulta mayor que uno (es decir, mayor que c), se deducirá que el perfil cumple la condición $LR > c$ del test de razón de verosimilitudes, con lo que será más favorable la hipótesis nula H_0. Se observa que si, por ejemplo, $LR = 3$, lo que se está diciendo es que el perfil estudiado es 3 veces más probable bajo la hipótesis nula que bajo la alternativa.

En este caso se ha trabajado con un solo marcador genético para poder exponer la metodología empleada de manera sencilla, pero lo habitual será analizar varios marcadores genéticos de forma simultánea para un perfil de ADN. Esto se debe a la necesidad de obtener unas probabilidades muy elevadas (resultados "casi seguros"), de forma que las decisiones tomadas en los casos judiciales, como puede ser la coincidencia de una muestra de ADN de un crimen con el genotipo de un sospechoso, o la decisión de que un hombre es realmente el padre biológico de un niño, sean 'lo más fiables posible". Se trata de sentencias de gran importancia y con consecuencias que pueden dar lugar a situaciones graves, que deben estar apoyadas por unos resultados concluyentes.

Hay que destacar que se trabaja con la hipótesis de una población en *equilibrio de Hardy-Weinberg*. Esto permite calcular la razón de verosimilitudes del perfil general, esto es, del conjunto de marcadores, como el producto de las razones de verosimilitud de cada uno de los marcadores. Es decir, se buscan n marcadores genéticos independientes con sus respectivas razones de verosimilitud, $\{LR_i\}_{i=1}^n$ (calculadas como se ha explicado previamente), de modo que la razón de verosimilitudes del perfil de ADN dado por esos n marcadores es:

$$LR_T = \prod_{i=1}^{n} LR_i.$$

3.2. Standard trio

El problema de *standard trio* se encuentra de forma común en las pruebas de paternidad. Se trata del caso en el que se tienen como datos los perfiles de ADN del hijo, H, de la madre, M, y del supuesto padre, SP, y se quiere comprobar si este hombre al que se ha denominado "supuesto padre" es realmente el padre biológico del niño. Se supone que estas tres personas pertenecen a una población en *equilibrio de Hardy-Weinberg* y que la madre, el supuesto padre, y el padre biológico del niño no están relacionados biológicamente entre ellos.

El contraste de hipótesis que ilustra este problema sería el que sigue:

$$\begin{cases} H_0: & \text{El supuesto padre, } SP, \text{ es el padre biológico del niño, } H, \\ H_a: & \text{El supuesto padre, } SP, \text{ no es el padre biológico del niño, } H. \end{cases}$$

Nótese que la hipótesis alternativa H_a podría reescribirse como que el padre biológico es un hombre de la población que no es SP; además, la maternidad no se pone en duda.

Viéndolo de otra manera, el contraste de hipótesis es el siguiente:

$$\begin{cases} H_0: & \text{El genotipo del padre biológico coincide con el del supuesto padre , } SP, \\ H_a: & \text{El genotipo del padre biológico no coincide con el del supuesto padre, } SP. \end{cases}$$

3.2.1. Índice de paternidad

De manera general, se llamará G_{SP}, G_M y G_H a los genotipos del supuesto padre, la madre y el hijo respectivamente, que constituyen los datos de los que se dispone. Por ejemplo, si se tuviera un solo marcador con dos alelos: $Y_1,\ Y_2$, entonces G_H podría ser $Y_1Y_1,\ Y_1Y_2$ o Y_2Y_2.

Entonces la razón de verosimilitudes, LR, representa el denominado *índice de paternidad* (IP) que sería:

$$LR = IP = \frac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)},$$

puesto que, como se indicó previamente, los datos disponibles son los genotipos de la madre, del niño y del supuesto padre. En los problemas de paternidad se empleará siempre la notación "índice de paternidad", pero se debe tener presente que se trata de la razón de verosimilitudes.

En esta fórmula, $\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)$ es la probabilidad de que, cuando SP es el padre biológico del niño, se obtengan en la población los genotipos G_H para el niño, G_M para la madre y G_{SP} para el padre. (Estos tres genotipos deberán cumplir algunas características, por ejemplo, el genotipo de la madre debe compartir al menos un alelo con el del hijo, pues la maternidad nunca se pone en duda, y el genotipo del supuesto padre, al ser considerado bajo la hipótesis nula como el padre biológico, deberá también compartir al menos uno de sus alelos con el genotipo del niño).

Por otra parte, $\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)$ representa la probabilidad de que se obtengan en la población los genotipos G_H para el niño, G_M para la madre (que, de nuevo, debe tener al menos uno de sus alelos iguales a los del niño) y el genotipo G_{SP}, que en este caso no estará relacionado con los otros dos por suponerse bajo la hipótesis alternativa que el genotipo de SP no es el del padre biológico.

Teniendo en consideración que tanto la probabilidad de obtener en la población el genotipo de la madre, G_M, como la de obtener el genotipo del supuesto padre, G_{SP}, no varían bajo ninguna de las hipótesis del contraste, es decir, la probabilidad de obtener el genotipo de la madre o del supuesto padre es independiente del hecho de que este último sea el padre biológico del niño o no; y que además son independientes entre ellos, se tiene que:

$$\mathrm{P}(G_M \mid G_{SP}, H_0) = \mathrm{P}(G_M \mid G_{SP}, H_a) = \mathrm{P}(G_M), \quad \mathrm{P}(G_{SP} \mid H_0) = \mathrm{P}(G_{SP} \mid H_a) = \mathrm{P}(G_{SP}),$$

donde $\mathrm{P}(G_M \mid G_{SP}, H_0)$ es la probabilidad de obtener el genotipo de la madre, G_M, en la población sabiendo que SP es el padre biológico. $\mathrm{P}(G_M \mid G_{SP}, H_a)$ es la probabilidad de obtener el genotipo de la madre en la población, sabiendo que SP no es el padre biológico del niño y $\mathrm{P}(G_M)$ es la probabilidad de obtener el genotipo de la madre en la población.
De la misma forma, $\mathrm{P}(G_{SP} \mid H_0)$ es la probabilidad de obtener el genotipo del supuesto padre, G_{SP}, en la población, sabiendo que es el padre biológico del niño, $\mathrm{P}(G_{SP} \mid H_a)$ es la probabilidad

de obtener el genotipo G_{SP}, sabiendo que SP no es el padre biológico del niño y $\mathrm{P}(G_{SP})$ es la probabilidad de obtener el genotipo G_{SP} en la población.

Además, bajo la hipótesis alternativa H_a, se tiene que SP no es el padre biológico del niño, por lo que los genotipos del niño y de SP serán independientes:

$$\mathrm{P}(G_H \mid G_M, G_{SP}, H_a) = \mathrm{P}(G_H \mid G_M, H_a).$$

En este caso, $\mathrm{P}(G_H \mid G_M, G_{SP}, H_a)$ representa la probabilidad de que el niño tenga el genotipo G_H, sabiendo que el de la madre es G_M (y que debe compartir al menos uno de sus alelos con ella), y sabiendo que el genotipo del supuesto padre es G_{SP}, que bajo la hipótesis alternativa no es su padre biológico. Entonces, esta probabilidad es igual a $\mathrm{P}(G_H \mid G_M, H_a)$, la probabilidad de que el niño tenga el genotipo G_H sabiendo que su madre tiene genotipo G_M y que SP no es su padre biológico (el genotipo de H es independiente del genotipo de SP).

De esta misma forma, $\mathrm{P}(G_H \mid G_M, G_{SP}, H_0)$ representa la probabilidad de que el genotipo del niño sea G_H sabiendo que el de la madre es G_M (y que debe compartir al menos uno de sus alelos con ella) y que el genotipo del supuesto padre es G_{SP}, que bajo la hipótesis nula se considera el padre biológico y por lo tanto el niño deberá compartir al menos uno de sus alelos con él.

A la vista de estos hechos y empleando la regla del producto, se puede reescribir el índice de paternidad como:

$$IP = \frac{\mathrm{P}(G_H \mid G_M, G_{SP}, H_0)}{\mathrm{P}(G_H \mid G_M, G_{SP}, H_a)} \cdot \frac{\mathrm{P}(G_M \mid G_{SP}, H_0)}{\mathrm{P}(G_M \mid G_{SP}, H_a)} \cdot \frac{\mathrm{P}(G_{SP} \mid H_0)}{\mathrm{P}(G_{SP} \mid H_a)} = \frac{\mathrm{P}(G_H \mid G_M, G_{SP}, H_0)}{\mathrm{P}(G_H \mid G_M, H_a)}.$$

A continuación se presenta un caso particular de esta situación en el que se analiza un solo marcador genético de un locus concreto, con al menos cuatro alelos diferentes (Y_i, Y_j, Y_k y Y_l), y en el que se considera que los genotipos del hijo, la madre y el supuesto padre son como sigue:

$$G_H = Y_iY_j, \quad G_M = Y_iY_k, \quad G_{SP} = Y_jY_l.$$

En esta situación, el numerador del índice de paternidad es sencillo de calcular, pues bajo la hipótesis nula hay una probabilidad de $1/2$ de que la madre, M, transmita su alelo Y_i al hijo, H; y de la misma forma, hay una probabilidad de $1/2$ de que, bajo la hipótesis nula H_0, el supuesto padre SP (que bajo H_0 se supone que es el padre biológico) transmita su alelo Y_j al niño, H. Así, el numerador de IP quedaría

$$\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_k,\ G_{SP} = Y_jY_l,\ H_0) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Para calcular el denominador, bajo la hipótesis alternativa, en el caso de la madre se tiene el mismo resultado: la probabilidad de que el niño H herede el alelo Y_i de su madre es de $1/2$.

Ahora bien, dado que bajo esta hipótesis, H_a, el supuesto padre no es el padre biológico del niño, la probabilidad de que este tenga un alelo Y_j vendrá dada por la probabilidad de este alelo en la población según la ley de Hardy-Weinberg, es decir, p_j. Con esto, la probabilidad que aparece en el denominador del índice de paternidad, IP, resultará:

$$\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_k,\ H_a) = \frac{1}{2} \cdot p_j.$$

Y por lo tanto, el índice de paternidad es:

$$IP = \frac{\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_k,\ G_{SP} = Y_jY_l,\ H_0)}{\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_k,\ H_a)} = \frac{1/4}{p_j/2} = \frac{1}{2p_j},$$

en el caso en el que los tres sujetos sean heterocigóticos.

En las tablas siguientes se exponen los distintos índices de paternidad teniendo en cuenta las posibles combinaciones alélicas para la madre y el supuesto padre dependiendo de si el hijo es homocigótico o heterocigótico; que se calculan de manera análoga. Como ejemplo se presenta de manera detallada otro de los cálculos de estas tablas: el caso en el que tanto el hijo como la madre y el supuesto padre son heterocigóticos, con genotipos $G_H = Y_iY_j$, $G_M = Y_iY_j$ y $G_{SP} = Y_iY_k$. Cabe recordar que la maternidad nunca se pone en duda.

Para calcular el numerador del índice de paternidad en este caso,

$$\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_j,\ G_{SP} = Y_iY_k,\ H_0),$$

hay que tener en cuenta que bajo la hipótesis nula (SP es el padre biológico del niño), se tiene una probabilidad de 1/2 de que el niño herede el alelo Y_i de SP, y una probabilidad de 1/2 de que herede el alelo Y_j de su madre. Con esto,

$$\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_j,\ G_{SP} = Y_iY_k,\ H_0) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Para el denominador, $\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_j,\ H_a)$, se tiene una probabilidad de 1/2 de que el niño herede el alelo Y_i de la madre, con lo que la probabilidad de que tenga el alelo Y_j, teniendo en cuenta que bajo H_a el supuesto padre no es su padre biológico, viene dada por p_j. Por otra parte, hay una probabilidad de 1/2 de que el alelo que herede de su madre sea el Y_j, con lo que la probabilidad de que tenga el alelo Y_i será de p_i. En resumen, el denominador es:

$$\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_j,\ H_a) = \frac{1}{2} \cdot p_j + \frac{1}{2} \cdot p_i = \frac{p_i + p_j}{2}.$$

Y gracias a estos cálculos, el índice de paternidad para el caso $G_H = Y_iY_j$, $G_M = Y_iY_j$ y $G_{SP} = Y_iY_k$ es:

$$IP = \frac{\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_j,\ G_{SP} = Y_iY_k,\ H_0)}{\mathrm{P}(G_H = Y_iY_j \mid G_M = Y_iY_j,\ H_a)} = \frac{1/4}{(p_i + p_j)/2} = \frac{1}{2(p_i + p_j)}.$$

Los casos restantes se pueden ver de manera semejante, y a continuación se presentan las tablas resumen de todos estos índices de paternidad.

G_H	Y_iY_i				
G_M	Y_iY_i		Y_iY_j		
G_{SP}	Y_iY_i	Y_iY_j	Y_iY_i	Y_iY_j	Y_iY_k
IP	$1/p_i$	$1/(2p_i)$	$1/p_i$	$1/(2p_i)$	$1/(2p_i)$

Tabla 3.1: Índices de paternidad para el caso H homocigótico.

G_H	Y_iY_j									
G_M	Y_iY_i			Y_iY_j			Y_iY_k			
G_{SP}	Y_iY_j	Y_jY_j	Y_jY_k	Y_iY_i	Y_iY_j	Y_iY_k	Y_iY_j	Y_jY_j	Y_jY_k	Y_jY_l
IP	$\frac{1}{2p_j}$	$\frac{1}{p_j}$	$\frac{1}{2p_j}$	$\frac{1}{(p_i+p_j)}$	$\frac{1}{(p_i+p_j)}$	$\boldsymbol{\frac{1}{2(p_i+p_j)}}$	$\frac{1}{2p_j}$	$\frac{1}{p_j}$	$\frac{1}{2p_j}$	$\boldsymbol{\frac{1}{2p_j}}$

Tabla 3.2: Índices de paternidad para el caso H heterocigótico.

3.2.2. Probabilidad de paternidad

A continuación se calcula un índice que es muy utilizado en la genética forense. Para ello serán necesarias las conocidas como probabilidades "a priori" y probabilidades "a posteriori". La probabilidad a priori de que un individuo sea el padre de un niño, esto es, la probabilidad de que el genotipo del supuesto padre, SP, coincida con el del padre biológico del niño, es de $1/2$ cuando no se tiene ninguna otra información.

Una vez conocidos los genotipos de los individuos involucrados (la madre, el niño y el supuesto padre), la probabilidad de que el genotipo del supuesto padre coincida con el genotipo del padre biológico, esto es, la probabilidad de que el genotipo del supuesto padre pueda dar lugar al genotipo del niño (dado el genotipo de la madre), es lo que se conoce como probabilidad a posteriori de que un individuo sea el padre biológico.

Se denotará por $\mathrm{P}(H_0)$ a la probabilidad a priori de que SP sea el padre del niño, es decir, la probabilidad de que el genotipo del supuesto padre coincida con el genotipo del padre biológico, y se denotará por $\mathrm{P}(H_a)$ a la probabilidad a priori de que SP no sea el padre biológico del niño. Estas se tomarán como equiprobables en la mayoría de situaciones puesto que no se tiene ninguna información adicional que otorgue preferencia a alguna de las dos hipótesis sobre la otra, es decir:

$$\mathrm{P}(H_0) = \mathrm{P}(H_a) = \frac{1}{2}.$$

De esta misma forma, se denotará por $\mathrm{P}(H_0 \mid G_H, G_M, G_{SP})$ a la probabilidad de que el

supuesto padre sea el padre biológico, una vez conocidos los genotipos de los individuos involucrados, es decir, la probabilidad de que el genotipo del supuesto padre pueda dar lugar al genotipo del niño, conocido el genotipo de la madre.

En concreto, esta probabilidad a posteriori, $\mathrm{P}(H_0 \mid G_H, G_M, G_{SP})$, es la conocida como *probabilidad de paternidad*. En genética forense recibe el nombre de *índice de Essen-Möller*, W, y se puede calcular empleando la regla de Bayes:

$$\mathrm{P}(H_0 \mid G_H, G_M, G_{SP}) = \frac{\mathrm{P}(H_0) \cdot \mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP})}.$$

Donde la probabilidad asociada a los datos, $\mathrm{P}(G_H, G_M, G_{SP})$, es la probabilidad de obtener los genotipos del niño, la madre y el supuesto padre en la población, y se calcula empleando la ley de las probabilidades totales:

$$\mathrm{P}(G_H, G_M, G_{SP}) = \mathrm{P}(G_H, G_M, G_{SP} \mid H_0) \cdot \mathrm{P}(H_0) + \mathrm{P}(G_H, G_M, G_{SP} \mid H_a) \cdot \mathrm{P}(H_a),$$

donde $\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)$ es la probabilidad de obtener en la población el genotipo del niño, G_H, el genotipo de la madre, G_M, que debe compartir al menos uno de sus alelos con el del niño, y el genotipo del supuesto padre, G_{SP}, que bajo la hipótesis nula es el padre biológico del niño y por lo tanto debe compartir al menos uno de sus alelos con él. Y de la misma forma, $\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)$ es la probabilidad de obtener en la población el genotipo de la madre, G_M, el genotipo del niño, G_H, que debe compartir al menos uno de sus alelos con el genotipo de la madre, y el genotipo del supuesto padre, G_{SP}, que bajo la hipótesis alternativa no es el padre biológico del niño.

Teniendo esto en cuenta, el índice de Essen-Möller, W, se reescribe como sigue:

$$\begin{aligned} W &= \mathrm{P}(H_0 \mid G_H, G_M, G_{SP}) = \\ &= \frac{\mathrm{P}(H_0) \cdot \mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0) \cdot \mathrm{P}(H_0) + \mathrm{P}(G_H, G_M, G_{SP} \mid H_a) \cdot \mathrm{P}(H_a)} = \\ &= \frac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0) + \mathrm{P}(G_H, G_M, G_{SP} \mid H_a)}. \end{aligned} \tag{3.1}$$

Sabiendo que el índice de paternidad es:

$$IP = \frac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)},$$

se tiene que, al ser la probabilidad $\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)$ un valor positivo, se puede reescribir W en función de los índices de paternidad a partir de la ecuación previa:

$$\begin{aligned} W &= \frac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0) + \mathrm{P}(G_H, G_M, G_{SP} \mid H_a)} = \\ &= \frac{\dfrac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)}}{\dfrac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)} + \dfrac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)}} = \frac{IP}{IP + 1}. \end{aligned} \tag{3.2}$$

3.2.3. Ejemplo standard trio

Se plantea un caso de determinación de paternidad para el problema *standard trio*. Se han obtenido los perfiles de ADN de la madre, M, el hijo, H, y el supuesto padre, SP, para tres marcadores genéticos de tipo STR (*Short Tandem Repeat*) distintos e independientes: **D3S1358**, **vWA** y **FGA**; y se han extraído los datos sobre sus probabilidades alélicas de una base de datos de una población noruega tomados del libro de Egeland, Kling y Mostad [3], y con los que se trabajará en el paquete *Familias* del sofware **R**.

El contraste de hipótesis en este caso es el planteado para el problema *standard trio*:

$$\begin{cases} H_0: & \text{El supuesto padre, } SP, \text{ es el padre biológico del niño, } H, \\ H_a: & \text{El padre biológico es un hombre cualquiera de la población no relacionado (no } SP). \end{cases}$$

Para este caso, se supone que los genotipos para cada uno de los individuos involucrados en el problema (el niño, la madre y el supuesto padre), son los que se presentan en la siguiente tabla:

Marcador	**D3S1358**	**vWA**	**FGA**
G_H	15:17	18:18	20:21
G_M	15:16	18:19	20:21
G_{SP}	17:18	18:18	19:20

Tabla 3.3: Genotipos supuestos para los individuos involucrados en el caso.

Para el primer marcador genético, D3S1358, $G_{SP} = Y_iY_j = 17:18$, es decir, "17" representa el alelo 17 del marcador ($Y_i = 17$) y "18" representa el alelo 18 ($Y_j = 18$). Para el marcador vWA, considerando de nuevo el genotipo del supuesto padre, "18" es el alelo 18 del marcador. Y por último, para FGA, "19" su alelo 19 y de forma análoga "20" representa el alelo 20.

Las probabilidades en la población para cada uno de los alelos involucrados se muestran en la siguiente tabla:

Marcador	**D3S1358**				**vWA**		**FGA**		
Alelo	15	16	17	18	18	19	19	20	21
Probabilidad	0,2635	0,2367	0,2040	0,1394	0,2107	0,0903	0,0602	0,1551	0,1725

Tabla 3.4: Probabilidades alélicas en una población noruega [1].

En la figura 3.1 se representa la situación genealógica y genotípica, tanto bajo la hipótesis nula como bajo la hipótesis alternativa. En ella, *added 1* representa al padre biológico (desconocido)

y del que no se conoce por lo tanto su genotipo para cada uno de los marcadores (lo que se representa por $-:-$). Se trata del padre biológico del niño en el caso de la hipótesis alternativa (cuando el supuesto padre, SP, no es el padre biológico). Además, debajo del nombre de cada uno de los individuos de este problema se han indicado sus genotipos para cada uno de los marcadores genéticos analizados (D3S1358, vWA y FGA).

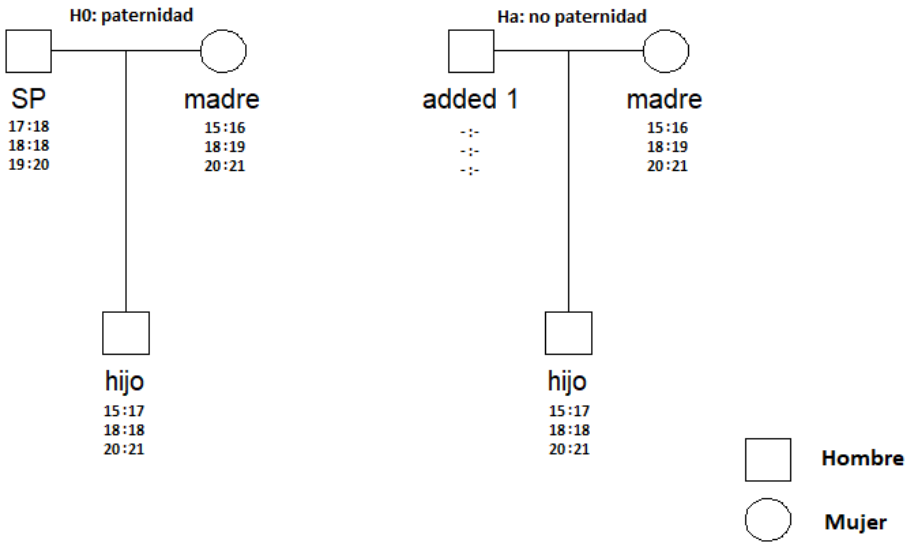

Figura 3.1: Ejemplo standard trio.

El primer objetivo será calcular el índice de paternidad:

$$IP = \frac{\mathrm{P}(G_H \mid G_M, G_{SP}, H_0)}{\mathrm{P}(G_H \mid G_M, H_a)}.$$

Para ello se comenzará calculando este índice para cada uno de los marcadores:

Marcador D3S1358:

Los genotipos de la madre, el hijo y el supuesto padre para este marcador vienen dados por:

$$G_M = 15{:}16, \quad G_H = 15{:}17, \quad G_{SP} = 17{:}18.$$

Primero se calcula la probabilidad $\mathrm{P}(G_H = 15:17 \mid G_M = 15:16,\ G_{SP} = 17:18,\ H_0)$, es decir, el numerador de la ecuación del índice de paternidad. Bajo la suposición de la hipótesis nula, que el supuesto padre, SP, es el padre biológico del niño, este último tiene una probabilidad

de 1/2 de heredar el alelo 15 de su madre, y una probabilidad de 1/2 de heredar el alelo 17 de su padre. Con lo que el numerador quedaría igual a $(1/2) \cdot (1/2) = 1/4$.

Para calcular el denominador de este índice, es decir, $\mathrm{P}(G_H = 15:17 \mid G_M = 15:16,\ H_a)$ basta tener en cuenta que la probabilidad de recibir el alelo 15 de su madre sigue siendo de 1/2, pero, bajo la hipótesis alternativa, cuando SP no es el padre biológico, la probabilidad de que el niño tenga un alelo 17 es $p_{17} = 0{,}2040$ como se muestra en la tabla 3.4. Con esto, el denominador es $(1/2) \cdot p_{17} = \frac{0{,}2040}{2} = 0{,}102$.

Haciendo los cálculos y sabiendo que el genotipo del hijo es de la forma Y_iY_j, el de la madre es Y_iY_k y el del supuesto padre es Y_jY_l; el índice de paternidad es:

$$IP_1 = \frac{\mathrm{P}(G_H = 15:17 \mid G_M = 15:16,\ G_{SP} = 17:18,\ H_0)}{\mathrm{P}(G_H = 15:17 \mid G_M = 15:16,\ H_a)} = \frac{1/4}{p_{17}/2} = \frac{0{,}25}{0{,}102} = 2{,}45098$$

que coincide con el valor obtenido utilizando la información de la tabla:
$IP_1 = 1/(2p_{17}) = 1/0{,}2040 = 2{,}45098$.

Por último, para este marcador, se calculan las verosimilitudes de los datos ($G_H = 15:17$, $G_M = 15:16,\ G_{SP} = 17:18$) bajo cada hipótesis, esto es:

- $\mathrm{P}(G_H = 15:17,\ G_M = 15:16,\ G_{SP} = 17:18 \mid H_0)$, que es la verosimilitud de los datos bajo la hipótesis nula, y que usando la regla del producto es:

$$\begin{aligned}
&\mathrm{P}(G_H = 15:17,\ G_M = 15:16,\ G_{SP} = 17:18 \mid H_0) = \\
&\quad = \mathrm{P}(G_H = 15:17 \mid G_M = 15:16,\ G_{SP} = 17:18,\ H_0) \cdot \\
&\qquad \cdot \mathrm{P}(G_M = 15:16 \mid G_{SP} = 17:18,\ H_0) \cdot \mathrm{P}(G_{SP} = 17:18 \mid H_0) = \\
&\quad = \frac{1}{4} \cdot \mathrm{P}(G_M = 15:16) \cdot \mathrm{P}(G_{SP} = 17:18) = \frac{1}{4} \cdot 2p_{15}p_{16} \cdot 2p_{17}p_{18} = \\
&\quad = 0{,}2635 \cdot 0{,}2367 \cdot 0{,}2040 \cdot 0{,}1394 = 0{,}001774.
\end{aligned} \tag{3.3}$$

  Nótese que es necesario recordar la independencia de los genotipos de la madre y del supuesto padre con las hipótesis, así como la independencia entre ellos y la suposición de población en *equilibrio de Hardy-Weinberg*.
- $\mathrm{P}(G_H = 15:17,\ G_M = 15:16,\ G_{SP} = 17:18 \mid H_a)$, que es la verosimilitud de los datos bajo la hipótesis alternativa y se calcula de la siguiente forma:

$$\begin{aligned}
&\mathrm{P}(G_H = 15:17,\ G_M = 15:16,\ G_{SP} = 17:18 \mid H_a) = \\
&\quad = \mathrm{P}(G_H = 15:17 \mid G_M = 15:16,\ G_{SP} = 17:18,\ H_a) \cdot \\
&\qquad \cdot \mathrm{P}(G_M = 15:16 \mid G_{SP} = 17:18,\ H_a) \cdot \mathrm{P}(G_{SP} = 17:18 \mid H_a) = \\
&\quad = \frac{p_{17}}{2} \cdot \mathrm{P}(G_M = 15:16) \cdot \mathrm{P}(G_{SP} = 17:18) = \frac{p_{17}}{2} \cdot 2p_{15}p_{16} \cdot 2p_{17}p_{18} = \\
&\quad = 2(p_{17})^2 p_{15}p_{16}p_{18} = 2 \cdot (0{,}2040)^2 \cdot 0{,}2635 \cdot 0{,}2367 \cdot 0{,}1394 = 0{,}000724.
\end{aligned} \tag{3.4}$$

Marcador vWA:

Los genotipos de la madre, hijo y supuesto padre para este marcador eran:

$$G_M = 18{:}19, \quad G_H = 18{:}18, \quad G_{SP} = 18{:}18.$$

Para obtener el numerador del índice de paternidad, es decir,

$$\mathrm{P}(G_H = 18{:}18 \mid G_M = 18{:}19,\ G_{SP} = 18{:}18,\ H_0),$$

hay que calcular la probabilidad de que el hijo obtenga el genotipo $18{:}18$ teniendo en cuenta que el de la madre es $18{:}19$ y el del supuesto padre es $18{:}18$, que bajo la hipótesis nula se considera como el padre biológico del niño. A la vista de los datos, el niño, H, heredará un alelo 18 de su padre con probabilidad 1, y heredará su otro alelo 18 de su madre con probabilidad $1/2$, es decir, el numerador vendrá dado por el valor $1/2$.

Para el denominador, $\mathrm{P}(G_H = 18:18 \mid G_M = 18:19,\ H_a)$, se calcula la probabilidad del genotipo del niño, $G_H = 18{:}18$ teniendo en cuenta que bajo la hipótesis alternativa heredará el alelo 18 de su madre con probabilidad $1/2$ y el supuesto padre SP no es su padre biológico; con lo que la probabilidad de que tenga el otro alelo 18 es p_{18}. Con lo que el denominador tendrá un valor de $(1/2) \cdot p_{18}$.

Con esto, como el hijo tiene un genotipo de la forma Y_iY_i, la madre tiene un genotipo de la forma Y_iY_j y el padre tiene un genotipo Y_iY_i, haciendo los cálculos, la razón de verosimilitudes o índice de paternidad es:

$$IP_2 = \frac{\mathrm{P}(G_H = 18{:}18 \mid G_M = 18{:}19,\ G_{SP} = 18{:}18, H_0)}{\mathrm{P}(G_H = 18{:}18 \mid G_M = 18{:}19,\ H_a)} = \frac{1/2}{p_{18}/2} = \frac{1}{p_{18}} = \frac{1}{0{,}2107} = 4{,}74608$$

que coincide con el resultado obtenido empleando la información de las tablas.

Para este segundo marcador genético se calculan también las verosimilitudes de los datos $(G_H = 18{:}18,\ G_M = 18{:}19,\ G_{SP} = 18{:}18)$ bajo las hipótesis del contraste:

- La verosimilitud bajo la hipótesis nula, $\mathrm{P}(G_H = 18{:}18,\ G_M = 18{:}19,\ G_{SP} = 18{:}18 \mid H_0)$, empleando la regla del producto de la misma forma que para el marcador anterior, se obtiene de la siguiente manera:

$$\begin{aligned} &\mathrm{P}(G_H = 18{:}18,\ G_M = 18{:}19,\ G_{SP} = 18{:}18 \mid H_0) = \\ &= \frac{1}{2} \cdot 2p_{18}p_{19} \cdot (p_{18})^2 = (p_{18})^3 \cdot p_{19} = (0{,}2107)^3 \cdot 0{,}0903 = 0{,}000845. \end{aligned} \tag{3.5}$$

  Donde se han tenido de nuevo en cuenta las condiciones de independencia, *equilibrio de Hardy-Weinberg* y la tabla 3.4 (probabilidades alélicas en una población noruega).

- La verosimilitud de los datos bajo la hipótesis alternativa se calcula de forma similar empleando la regla del producto:

$$\begin{aligned} &\mathrm{P}(G_H = 18{:}18,\ G_M = 18{:}19,\ G_{SP} = 18{:}18 \mid H_a) = \\ &= \frac{p_{18}}{2} \cdot 2p_{18}p_{19} \cdot (p_{18})^2 = (p_{18})^4 \cdot p_{19} = (0{,}2107)^4 \cdot 0{,}0903 = 0{,}000178. \end{aligned} \tag{3.6}$$

Marcador FGA:

Para este tercer STR , los genotipos de la madre, el hijo y el supuesto padre se suponen:

$$G_M = 20{:}21, \quad G_H = 20{:}21, \quad G_{SP} = 19{:}20.$$

Al igual que en los casos anteriores, se comienza calculando el numerador de la fórmula del índice de paternidad, en este caso, la probabilidad de que el niño herede el alelo 20 de su supuesto padre, SP, que bajo la hipótesis nula se supone que es el padre biológico, es de 1/2, y de la misma forma, la probabilidad de que el niño herede el alelo 21 de su madre es de 1/2; con lo que el numerador será: $\mathrm{P}(G_H = 20{:}21 \mid G_M = 20{:}21,\ G_{SP} = 19{:}20,\ H_0) = 1/4$.

El siguiente paso es calcular la probabilidad del denominador. Bajo la hipótesis alternativa, el niño, H, tiene una probabilidad de 1/2 de heredar el alelo 20 de su madre, así como una probabilidad de 1/2 de heredar el alelo 21; por lo tanto, teniendo en cuenta que bajo la hipótesis alternativa el supuesto padre no es su padre biológico, la probabilidad de que tenga el alelo 21 si heredó el 20 de su madre es p_{21}, y la probabilidad de que tenga el 20 si heredó el 21 es p_{20}. Con esto, la probabilidad del denominador viene dada por: $\mathrm{P}(G_H = 20:21 \mid G_M = 20:21,\ H_a) = (1/2) \cdot p_{21} + (1/2) \cdot p_{20}$.

Así, como el genotipo del hijo es de la forma Y_iY_j, el de la madre es de la forma Y_iY_j y el del supuesto padre es del tipo Y_iY_k, haciendo los cálculos se tiene que el índice de paternidad es:

$$\begin{aligned} IP_3 &= \frac{\mathrm{P}(G_H = 20{:}21 \mid G_M = 20{:}21,\ G_{SP} = 19{:}20,\ H_0)}{\mathrm{P}(G_H = 20{:}21 \mid G_M = 20{:}21,\ H_a)} = \\ &= \frac{1/4}{(p_{20} + p_{21})/2} = \frac{1/4}{(1/2) \cdot (0{,}1551 + 0{,}1725)} = 1{,}52625. \end{aligned}$$

Que coincide con el resultado que se obtendría empleando la información de las tablas.

Finalmente se calculan las verosimilitudes de los datos ($G_H = 20{:}21,\ G_M = 20{:}21,\ G_{SP} = 19{:}20$) bajo cada una de las hipótesis para este tercer marcador genético:

- $\mathrm{P}(G_H = 20:21,\ G_M = 20:21,\ G_{SP} = 19:20 \mid H_0)$ es la probabilidad de los datos bajo la hipótesis nula, que se calcula de la misma forma que en los dos casos anteriores,

obteniéndose el siguiente resultado:

$$\begin{aligned} &\mathrm{P}(G_H = 20{:}21,\ G_M = 20{:}21,\ G_{SP} = 19{:}20 \mid H_0) = \\ &= \frac{1}{4} \cdot 2p_{20}p_{21} \cdot 2p_{19}p_{20} = p_{19}(p_{20})^2 p_{21} = 0{,}0602 \cdot (0{,}1551)^2 \cdot 0{,}1725 = 0{,}000250. \end{aligned} \tag{3.7}$$

Donde las probabilidades de los alelos se han extraído de la tabla 3.4 (probabilidades alélicas en una población noruega).

- $\mathrm{P}(G_H = 20{:}21,\ G_M = 20{:}21,\ G_{SP} = 19{:}20 \mid H_a)$ es la probabilidad de los datos bajo la hipótesis alternativa, que se calcula de la misma manera que en los casos anteriores y que tiene el siguiente valor para este marcador genético:

$$\begin{aligned} &\mathrm{P}(G_H = 20{:}21,\ G_M = 20{:}21,\ G_{SP} = 19{:}20 \mid H_a) = \\ &= \frac{(p_{20} + p_{21})}{2} \cdot 2p_{20}p_{21} \cdot 2p_{19}p_{20} = 2(p_{20} + p_{21})p_{19}(p_{20})^2 p_{21} = \\ &= 2 \cdot (0{,}1551 + 0{,}1725) \cdot 0{,}0602 \cdot (0{,}1551)^2 \cdot 0{,}1725 = 0{,}000164. \end{aligned} \tag{3.8}$$

Donde, al igual que en los otros dos marcadores genéticos, se han tenido en cuenta todas las hipótesis de independencia, así como el *equilibrio de Hardy-Weinberg*, y se ha hecho uso de la regla del producto.

Nótese que una vez calculadas las verosimilitudes de los datos bajo las dos hipótesis, para cada uno de los marcadores, empleando la fórmula:

$$IP = \frac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_a)},$$

es sencillo comprobar que se obtienen los índices de paternidad de cada caso.

Una vez realizadas estas operaciones individuales se puede obtener mediante la fórmula:

$$IP_T = \prod_{i=1}^{n} IP_i.$$

el índice de paternidad para el perfil de ADN completo que se está analizando con tan solo multiplicar los resultados obtenidos de los tres marcadores ($IP_1 = 2{,}45098,\ IP_2 = 4{,}74608,\ IP_3 = 1{,}52625$):

$$IP_T = IP_1 \cdot IP_2 \cdot IP_3 = 2{,}45098 \cdot 4{,}74608 \cdot 1{,}52625 = 17{,}75418.$$

Es decir, con tan solo el análisis de tres marcadores genéticos distintos, la probabilidad de obtener los perfiles bajo la hipótesis de paternidad es entre 17 y 18 veces mayor que la probabilidad de obtenerlos bajo la hipótesis alternativa, es decir, la no paternidad. Es evidente que esto no son pruebas suficientes para tomar una decisión acerca del contraste, pero en los laboratorios se

analizarán muchos más loci, siguiendo este mismo procedimiento, con el objetivo de obtener un cociente de verosimilitudes muy grande, de forma que los resultados sean concluyentes y se puedan tomar decisiones coherentes y rigurosas acerca del contraste.

Asimismo, se puede calcular la verosimilitud global de los datos para cada hipótesis, multiplicando las obtenidas para cada marcador:

- La verosimilitud global bajo H_0 se obtiene multiplicando los resultados de las ecuaciones (3.3), (3.5) y (3.7) (las verosimilitudes de los datos bajo H_0 para cada uno de los marcadores):

$$\begin{aligned}\mathrm{P}(G_H, G_M, G_{SP} \mid H_0) &= \mathrm{P}(D3S1358 \mid H_0) \cdot \mathrm{P}(vWA \mid H_0) \cdot \mathrm{P}(FGA \mid H_0) = \\ &= 0{,}001774 \cdot 0{,}000845 \cdot 0{,}000250 = 3{,}7475 \times 10^{-10}.\end{aligned}$$

  Donde $\mathrm{P}(D3S1358 \mid H_0)$ representa la probabilidad de los tres genotipos, G_H, G_M, G_{SP}, bajo H_0, para el marcador D3S1358; y para los otros marcadores se está empleando una notación análoga.

- La verosimilitud global bajo H_a se obtiene multiplicando los resultados obtenidos en (3.4), (3.6) y (3.8) (las verosimilitudes de los datos bajo H_a para cada uno de los marcadores):

$$\begin{aligned}\mathrm{P}(G_H, G_M, G_{SP} \mid H_a) &= \mathrm{P}(D3S1358 \mid H_a) \cdot \mathrm{P}(vWA \mid H_a) \cdot \mathrm{P}(FGA \mid H_a) = \\ &= 0{,}000724 \cdot 0{,}000178 \cdot 0{,}000164 = 2{,}1135 \times 10^{-11}.\end{aligned}$$

El último concepto estudiado para el *standard trio* que se aplica a este ejemplo es el del índice de Essen-Möller, W, que convierte la razón de verosimilitudes en una probabilidad a posteriori para la hipótesis de paternidad. Este índice se puede calcular mediante la fórmula (3.1) (empleando las verosimilitudes globales calculadas anteriormente):

$$\begin{aligned}W = \mathrm{P}(H_0 \mid G_H, G_M, G_{SP}) &= \frac{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_M, G_{SP} \mid H_0) + \mathrm{P}(G_H, G_M, G_{SP} \mid H_a)} = \\ &= \frac{3{,}7475 \times 10^{-10}}{3{,}7475 \times 10^{-10} + 2{,}1135 \times 10^{-11}} = 0{,}946679.\end{aligned}$$

Y es sencillo comprobar que se obtiene el mismo resultado empleando la caracterización del índice de Essen-Möller que emplea el índice de paternidad general (3.2):

$$W = \frac{IP_T}{IP_T + 1} = \frac{17{,}75418}{17{,}75418 + 1} = 0{,}946679.$$

Esto quiere decir, que suponiendo que ambas hipótesis son equiprobables a priori, es decir, la probabilidad de paternidad sin haber realizado el contraste con el conocimiento de los datos es del 50 %, la probabilidad de paternidad obtenida tras el análisis de los perfiles de ADN ha aumentado

hasta el 94,67 %. Con esto se puede deducir que el perfil es coincidente y se podría confirmar la hipótesis de paternidad, teniendo en cuenta que este análisis se ha hecho para únicamente tres marcadores genéticos y que en la realidad se obtendrán las conclusiones a partir del análisis de numerosos loci.

3.3. Standard duo

En algunos problemas de paternidad no se podrá disponer del perfil de ADN materno, en ese caso se estará tratando con el problema denominado como *standard duo*. Aun sin este perfil de ADN, todavía será de interés el contraste de hipótesis sobre la verdadera relación del supuesto padre, SP, con el niño, H.

$$\begin{cases} H_0: & \text{el supuesto padre, } SP, \text{ es el verdadero padre del niño, } H \\ H_a: & \text{el supuesto padre, } SP, \text{ no es el padre biológico del niño, } H \end{cases}$$

De nuevo, la hipótesis alternativa, H_a, puede reescribirse como "el verdadero padre es un hombre de la población, que no es SP". Este contraste también se podría ver como H_0 : "el genotipo del padre biológico coincide con el del supuesto padre" frente a H_a : "el genotipo del padre biológico no coincide con el del supuesto padre".

Este caso es similar al *standard trio*, pero se dispone de menos información para realizar deducciones acerca de las hipótesis. De nuevo se supone que la madre (desconocida) y el supuesto padre no están relacionados y que todos los individuos pertenecen a una población en *equilibrio de Hardy-Weinberg*; y se desea conocer si el supuesto padre, SP, es realmente el padre biológico del niño o no.

3.3.1. Índice de paternidad

Sean, al igual que en el *standard trio*, G_{SP} y G_H los genotipos del supuesto padre y del hijo respectivamente. En este caso el índice de paternidad (o razón de verosimilitudes) viene dado por la siguiente fórmula:

$$IP = \frac{\mathrm{P}(G_H, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_{SP} \mid H_a)}.$$

Pues los datos disponibles en este caso son los genotipos del supuesto padre y del hijo. Aquí, $\mathrm{P}(G_H, G_{SP} \mid H_0)$ representa la probabilidad de obtener en la población el genotipo del niño, G_H, y el genotipo del supuesto padre, G_{SP}, que bajo H_0 es el padre biológico del niño y por lo tanto deberá compartir al menos uno de sus alelos con él. Y $\mathrm{P}(G_H, G_{SP} \mid H_a)$ representa la probabilidad de obtener en la población el genotipo del niño, G_H, y el del supuesto padre, G_{SP},

sabiendo que bajo la hipótesis alternativa SP no es el padre biológico del niño y por lo tanto no está relacionado con él.

De nuevo, al igual que en el *standard trio*, el genotipo del supuesto padre es independiente del hecho de verificarse la hipótesis nula o la hipótesis alternativa. Esto es, su genotipo no será distinto sea el padre biológico del niño o no, y por lo tanto se verifica que:

$$\mathrm{P}(G_{SP} \mid H_0) = \mathrm{P}(G_{SP} \mid H_a) = \mathrm{P}(G_{SP}).$$

Donde $\mathrm{P}(G_{SP} \mid H_0)$ representa la probabilidad en la población del genotipo del supuesto padre, sabiendo que es el padre biológico del niño, $\mathrm{P}(G_{SP} \mid H_a)$ es la probabilidad de obtener en la población el genotipo del supuesto padre sabiendo que no es el padre biológico del niño, y $\mathrm{P}(G_{SP})$ es la probabilidad en la población del genotipo del supuesto padre.

Y en el caso de la hipótesis alternativa (SP no es el padre biológico de H), el genotipo del niño será independiente del genotipo del supuesto padre, con lo que se tiene:

$$\mathrm{P}(G_H \mid G_{SP}, H_a) = \mathrm{P}(G_H \mid H_a) = \mathrm{P}(G_H).$$

Esto es, la probabilidad de obtener en la población el genotipo del niño, G_H, sabiendo que el supuesto padre no es el padre biológico, es igual a la probabilidad de obtener en la población el genotipo del niño.

Dado lo anterior y empleando nuevamente la regla del producto, se tiene que el índice de paternidad es:

$$IP = \frac{\mathrm{P}(G_H, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_{SP} \mid H_a)} = \frac{\mathrm{P}(G_H \mid G_{SP}, H_0)}{\mathrm{P}(G_H \mid G_{SP}, H_a)} \cdot \frac{\mathrm{P}(G_{SP} \mid H_0)}{\mathrm{P}(G_{SP} \mid H_a)} = \frac{\mathrm{P}(G_H \mid G_{SP}, H_0)}{\mathrm{P}(G_H \mid H_a)}.$$

Siendo $\mathrm{P}(G_H \mid G_{SP}, H_0)$ la probabilidad de obtener en la población el genotipo del niño, G_H, sabiendo que el genotipo del supuesto padre es G_{SP}, que bajo la hipótesis nula es el padre biológico con lo que tendrá que compartir por lo menos uno de sus alelos con el niño.

A continuación se presenta un caso concreto en el que se supone que los genotipos del hijo y del supuesto padre, para un determinado marcador genético, son $G_H = Y_iY_j$ y $G_{SP} = Y_iY_k$ respectivamente, siendo Y_i, Y_j y Y_k alelos distintos del marcador analizado. En este caso, bajo la hipótesis nula H_0, el supuesto padre SP tiene una probabilidad de $1/2$ de transmitir su alelo Y_i al niño (dado que bajo H_0 se está suponiendo que es el verdadero padre), y la probabilidad de que el niño herede un alelo Y_j de su madre biológica (que es desconocida) viene dada, según la ley de Hardy-Weinberg, por p_j. Con esto, el numerador del índice de paternidad será:

$$\mathrm{P}(G_H = Y_iY_j \mid G_{SP} = Y_iY_k,\ H_0) = \frac{1}{2} \cdot p_j.$$

Para el denominador, como se tiene que el genotipo del niño es independiente del genotipo del supuesto padre bajo la hipótesis alternativa, y se desconoce el genotipo de la madre biológica,

se verifica por la ley de Hardy-Weinberg que su probabilidad en la población vendrá dada por:

$$\mathrm{P}(G_H = Y_iY_j \mid H_a) = \mathrm{P}(G_H = Y_iY_j) = 2p_ip_j.$$

Con lo que, en el caso en el que tanto el padre como el hijo sean heterocigóticos, el índice de paternidad se calculará como:

$$IP = \frac{\mathrm{P}(G_H = Y_iY_j \mid G_{SP} = Y_iY_k,\ H_0)}{\mathrm{P}(G_H = Y_iY_j \mid H_a)} = \frac{p_j/2}{2p_ip_j} = \frac{1}{4p_i}.$$

En la tabla siguiente se presentan los índices de paternidad para otras combinaciones alélicas posibles para el hijo y el supuesto padre, que se pueden obtener de forma similar al caso anterior.

G_H	Y_iY_i		Y_iY_j		
G_{SP}	Y_iY_i	Y_iY_j	Y_iY_i	Y_iY_j	Y_iY_k
IP	$1/p_i$	$1/(2p_i)$	$1/(2p_i)$	$\boldsymbol{(p_i + p_j)/(4p_ip_j)}$	$\boldsymbol{1/(4p_i)}$

Tabla 3.5: Índices de paternidad en el standard duo.

A continuación se presenta de manera detallada otro de los cálculos de esta tabla, en el que tanto el supuesto padre como el hijo son heterocigóticos, con genotipos: $G_H = Y_iY_j,\ G_{SP} = Y_iY_j$.

Para el numerador del índice de paternidad, se puede observar que la probabilidad de que el niño herede el alelo Y_i del supuesto padre (que bajo la hipótesis nula H_0 se supone que es el padre biológico) es de $1/2$; si hereda este alelo de su padre, existirá, por la ley de Hardy-Weinberg, una probabilidad p_j de que herede el alelo Y_j de su madre biológica (que es desconocida). Además, existe una probabilidad de $1/2$ de que el niño herede el alelo Y_j de su padre, SP, con lo que tiene una probabilidad p_i de heredar el alelo Y_i de su madre biológica en este caso. Con esto, el numerador resulta: $(1/2) \cdot p_j + (1/2) \cdot p_i = (p_i + p_j)/2$.

Para calcular el denominador, se trata únicamente de obtener la probabilidad del genotipo $G_H = Y_iY_j$ en la población, calculándolo mediante la ley de Hardy-Weinberg; pues bajo la hipótesis alternativa, H_a, se supone que SP no es el padre biológico, y la madre es desconocida en este problema. Con esto, el denominador del índice de paternidad es: $2p_ip_j$.

En conjunto, el índice de paternidad para el caso en el que tanto el supuesto padre como el hijo son heterocigóticos con genotipo Y_iY_j es:

$$IP = \frac{\mathrm{P}(G_H = Y_iY_j \mid G_{SP} = Y_iY_j,\ H_0)}{\mathrm{P}(G_H = Y_iY_j \mid H_a)} = \frac{(p_i + p_j)/2}{2p_ip_j} = \frac{p_i + p_j}{4p_ip_j}.$$

En el standard duo, o caso en el que se desconoce la componente materna, la probabilidad de paternidad se calcula de manera análoga al caso del *standard trio*; esto es, el índice de Essen-Möller, W, se calculará de nuevo empleando las razones de verosimilitud (los llamados índices

de paternidad):

$$W = \mathrm{P}(H_0 \mid G_H, G_{SP}) = \frac{\mathrm{P}(G_H, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_{SP} \mid H_0) + \mathrm{P}(G_H, G_{SP} \mid H_a)} = \frac{IP}{IP+1}.$$

Pues la probabilidad $\mathrm{P}(G_H, G_{SP} \mid H_a)$ es positiva.

En esta fórmula, $\mathrm{P}(H_0 \mid G_H, G_{SP})$ representa la probabilidad a posteriori de que un individuo sea el padre biológico del niño, esto es, conocido el genotipo del niño, la probabilidad de que el genotipo del supuesto padre coincida con el del padre biológico, sabiendo que este debe compartir al menos uno de sus alelos con el del niño. Además, $\mathrm{P}(G_H, G_{SP} \mid H_0)$ representa la probabilidad de obtener en la población el genotipo del niño, G_H, y el del supuesto padre, G_{SP}, sabiendo que bajo la hipótesis nula este es el padre biológico y, por tanto, debe compartir al menos uno de sus alelos con el niño; y $\mathrm{P}(G_H, G_{AP} \mid H_a)$ representa la probabilidad de obtener en la población el genotipo del niño y el del supuesto padre, sabiendo que bajo la hipótesis alternativa no es el padre biológico del niño, por lo que no está relacionado con él.

3.3.2. Ejemplo standard duo

Se presenta un caso de determinación de la paternidad en un problema de tipo *standard duo*. Se han obtenido los perfiles de ADN del hijo, H, y del supuesto padre, SP, para tres marcadores genéticos de tipo STR (*Short Tandem Repeat*) distintos e independientes: **D13S317**, **TPOX** y **CSF1PO** y el genotipo de la madre es desconocido. De nuevo, los datos sobre las probabilidades alélicas en la población de estos STR se han extraído de una base de datos de una población noruega, gracias al libro de Egeland, Kling y Mostad [3], y serán los empleados en el paquete *Familias* del software **R**.

El contraste de hipótesis a resolver es el general del problema *standard duo*:

$$\begin{cases} H_0: & \text{el supuesto padre, } SP, \text{ es el verdadero padre del niño, } H. \\ H_a: & \text{el supuesto padre, } SP, \text{ no es el padre biológico del niño, } H. \end{cases}$$

Se supone que, para el STR D13S317, los genotipos del supuesto padre y del hijo son, respectivamente, $G_{SP} = 10:13,\ G_H = 9:13$. Nótese que al igual que en el ejemplo del caso *standard trio*, "10", "13" y "9" representan tres alelos diferentes de este marcador genético, es decir, se corresponderían con Y_i, Y_j, Y_k. Para el marcador TPOX, se supone que $G_{SP} = 8:12,\ G_H = 8:8$. Y para el STR CSF1PO se suponen: $G_{SP} = 11:11,\ G_H = 10:11$.

La situación genealógica y genotípica de este caso se representa en la siguiente figura, en la que se puede ver, que el genotipo de la madre no se conoce para ninguno de los marcadores, tanto bajo la hipótesis nula como bajo la alternativa. Esto se debe a que en el caso *standard duo* la madre es desconocida, y al igual que en el ejemplo del *standard trio* esta situación se representa

mediante $-:-$. Asimismo, *added 1* representa al padre biológico del niño, que bajo la hipótesis nula, H_a, es desconocido también.

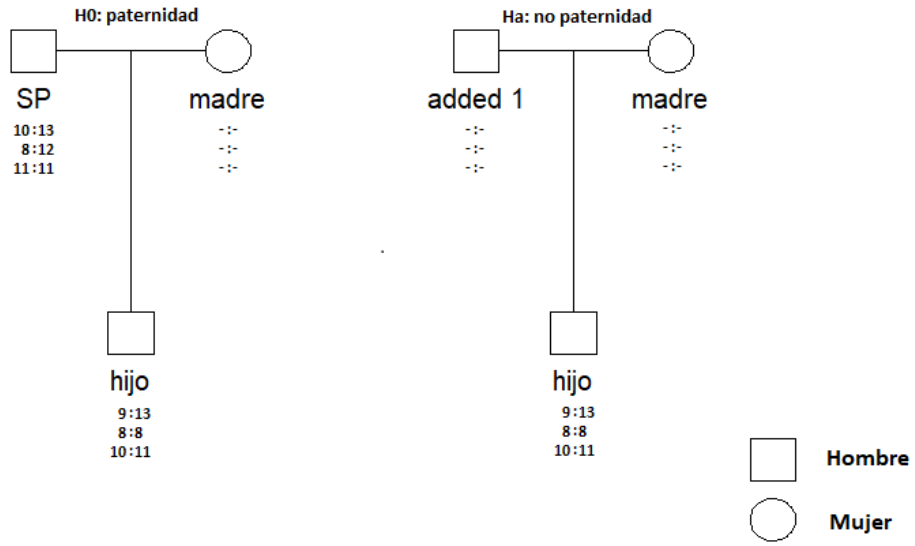

Figura 3.2: Ejemplo standard duo.

Las probabilidades en la población para cada uno de los alelos involucrados en este ejemplo se muestran en la siguiente tabla:

Marcador	**D13S317**			**TPOX**		**CSF1PO**	
Alelo	9	10	13	8	12	10	11
Probabilidad	0,0754	0,0889	0,1116	0,5539	0,0373	0,2472	0,2989

Tabla 3.6: Probabilidades alélicas en una población noruega [1].

Se comienza calculando el índice de paternidad para cada uno de los STR estudiados, cuya fórmula es:

$$IP = \frac{\mathrm{P}(G_H \mid G_{SP}, H_0)}{\mathrm{P}(G_H \mid H_a)}.$$

Marcador D13S317:

El genotipo del niño para este STR es $G_H = 9:13$, y el del supuesto padre $G_{SP} = 10:13$. Nótese que, relacionándolo con la tabla expuesta previamente en esta sección, se trata del caso en el que el genotipo del hijo es de la forma Y_iY_j y el del supuesto padre Y_iY_k.

Para calcular el numerador del índice de paternidad hay que tener en cuenta que la probabilidad de que el niño herede el alelo 13 de SP, que bajo la hipótesis nula es su padre biológico, es de 1/2; y al ser desconocido el genotipo de la madre, tiene una probabilidad p_9 de obtener su alelo 9. Con lo que el numerador resulta $\mathrm{P}(G_H = 9:13 \mid G_{SP} = 10:13,\ H_0) = (1/2) \cdot p_9$, que calculado con la información de la tabla de probabilidades alélicas expuesta previamente es: $\mathrm{P}(G_H = 9{:}13 \mid G_{SP} = 10{:}13,\ H_0) = (0{,}0754)/2 = 0{,}0377$.

Para calcular el denominador, como el genotipo de la madre es desconocido y, bajo la hipótesis alternativa, SP no es el padre biológico del niño, se tendrá que calcular la probabilidad del genotipo del niño dentro de la población mediante la ley de Hardy-Weinberg:
$\mathrm{P}(G_H = 9{:}13 \mid H_a) = 2p_9p_{13}$, que empleando de nuevo las probabilidades alélicas resulta:
$\mathrm{P}(G_H = 9{:}13 \mid H_a) = 2 \cdot 0{,}0754 \cdot 0{,}1116 = 0{,}01683$.

Con lo que el índice de paternidad viene dado por:

$$IP_1 = \frac{\mathrm{P}(G_H = 9{:}13 \mid G_{SP} = 10{:}13,\ H_0)}{\mathrm{P}(G_H = 9{:}13 \mid H_a)} = \frac{0{,}0377}{0{,}01683} = 2{,}24014.$$

Y se puede comprobar que este resultado coincide con el calculado mediante la fórmula de la tabla de índices de paternidad del caso *standard duo*: $IP_1 = 1/(4p_{13}) = 1/(4 \cdot 0{,}1116) = 2{,}24014$.

Marcador TPOX:

Para calcular el índice de paternidad en este caso se recuerda que los genotipos del niño y el supuesto padre son $G_H = 8:8$ y $G_{SP} = 8:12$. Para relacionarlo con la tabla de índices de paternidad calculada en esta sección, el genotipo del hijo es de la forma $G_H = Y_iY_i$ y el del supuesto padre es $G_{SP} = Y_iY_j$.

Para calcular la probabilidad del numerador hay que tener en cuenta que hay una probabilidad de 1/2 de que el niño herede el alelo 8 de SP, que bajo la hipótesis nula es su padre biológico, con lo que, al ser el genotipo de la madre desconocido, hay una probabilidad p_8 de que obtenga su otro alelo 8. Así, empleando la tabla de probabilidades alélicas, el numerador queda:
$\mathrm{P}(G_H = 8{:}8 \mid G_{SP} = 8{:}12,\ H_0) = (1/2) \cdot p_8 = 0{,}5539/2 = 0{,}27695$.

Para el denominador se tiene de nuevo la probabilidad del genotipo del hijo en la población, siguiendo la ley de Hardy-Weinberg: $\mathrm{P}(G_H = 8{:}8 \mid H_a) = (p_8)^2 = (0{,}5539)^2 = 0{,}3068$.

Con lo que se puede ver que el resultado será el mismo que se obtiene empleando la fórmula de la tabla 3.5:

$$IP_2 = \frac{\mathrm{P}(G_H = 8{:}8 \mid G_{SP} = 8{:}12,\ H_0)}{\mathrm{P}(G_H = 8{:}8 \mid H_a)} = \frac{0{,}27695}{0{,}3068} = 0{,}90269 = \frac{1}{2p_8}.$$

Marcador CSF1PO:

Los genotipos del niño y el supuesto padre para este STR se suponen $G_H = 10:11$ y $G_{SP} = 11:11$. Es decir, se trata de la situación de la tabla de índices de paternidad del caso *standard duo* donde $G_H = Y_iY_j$ y $G_{SP} = Y_iY_i$. Por lo tanto, el índice de paternidad se calculará con la fórmula:

$$IP_3 = \frac{\mathrm{P}(G_H = 10{:}11 \mid G_{SP} = 11{:}11,\ H_0)}{\mathrm{P}(G_H = 10{:}11 \mid H_a)} = \frac{1}{2p_{11}} = \frac{1}{2 \cdot 0{,}2989} = 1{,}6728.$$

Se puede comprobar de manera sencilla la procedencia de este resultado. Para el numerador del índice de paternidad, bajo la hipótesis nula H_0, se tiene que el niño heredará el alelo 11 de su padre con probabilidad 1, por ser este último (SP) homocigótico. Al ser desconocido el genotipo de la madre, la probabilidad de que el niño obtenga un alelo 10 viene dada por la ley de Hardy-Weinberg: p_{10}. Con esto, el numerador del índice de paternidad resulta
$\mathrm{P}(G_H = 10{:}11 \mid G_{SP} = 11{:}11,\ H_0) = p_{10} = 0{,}2472$, según la tabla de probabilidades alélicas.

Para calcular el denominador, de nuevo y como en los casos anteriores, bajo la hipótesis alternativa se tiene que SP no es el padre biológico del niño y que el genotipo de la madre también es desconocido, con lo que $\mathrm{P}(G_H = 10{:}11 \mid H_a) = 2p_{10}p_{11} = 2 \cdot 0{,}2472 \cdot 0{,}2989 = 0{,}14778$.

Y con esto, el índice de paternidad resulta, como se adelantaba con la fórmula de la tabla previa, $IP_3 = 0{,}2472/0{,}14778 = 1{,}6728$.

Una vez calculados los índices de paternidad para cada uno de los tres STR analizados, se puede obtener el índice de paternidad para el perfil completo teniendo en cuenta las hipótesis de independencia y de individuos en una población verificando el *equilibrio de Hardy-Weinberg*:

$$IP_T = \prod_{i=1}^{3} IP_i = IP_1 \cdot IP_2 \cdot IP_3 = 2{,}24014 \cdot 0{,}90269 \cdot 1{,}6728 = 3{,}38266.$$

Por lo tanto, se puede interpretar que, con tan solo el uso de tres marcadores genéticos, la probabilidad de obtener los perfiles de ADN del hijo y del supuesto padre son entre 3 y 4 veces más probables bajo la hipótesis de paternidad que bajo la hipótesis de no paternidad. Esto no proporciona pruebas suficientes para tomar una decisión acerca del contraste, pero permite ilustrar el procedimiento llevado a cabo en los laboratorios con más de 10 marcadores genéticos.

Para completar este problema se puede calcular el índice de Essen-Möller, W, tanto mediante las verosimilitudes de los perfiles de ADN del niño y del supuesto padre (G_H, G_{SP}) bajo cada una de las hipótesis del contraste, como empleando la fórmula de la razón de verosimilitudes.

De forma similar al ejemplo del caso *standard trio*, para calcular la verosimilitud de los genotipos dados, se deben multiplicar las verosimilitudes obtenidas para estos perfiles, para cada uno de los SRT independientemente.

Se calcula primero la probabilidad $\mathrm{P}(G_H, G_{SP} \mid H_0)$ para el marcador D13S317:

$$\begin{aligned}\mathrm{P}(G_H = 9{:}13,\ G_{SP} = 10{:}13 \mid H_0) = \\ = \mathrm{P}(G_H = 9{:}13 \mid G_{SP} = 10{:}13,\ H_0) \cdot \mathrm{P}(G_{SP} = 10{:}13 \mid H_0) = \\ = \frac{p_9}{2} \cdot 2p_{10}p_{13} = p_9p_{10}p_{13} = 0{,}0754 \cdot 0{,}0889 \cdot 0{,}1116 = 0{,}000748.\end{aligned}$$

Para el marcador TPOX, esta probabilidad es:

$$\mathrm{P}(G_H = 8{:}8,\ G_{SP} = 8{:}12 \mid H_0) = \frac{p_8}{2} \cdot 2p_8p_{12} = (p_8)^2 \cdot p_{12} = (0{,}5539)^2 \cdot 0{,}0373 = 0{,}011444.$$

Y para el STR CSF1PO se tiene que:

$$\mathrm{P}(G_H = 10{:}11,\ G_{SP} = 11{:}11 \mid H_0) = p_{10} \cdot (p_{11})^2 = 0{,}2472 \cdot (0{,}2989)^2 = 0{,}022085.$$

Para continuar, se calcula para cada marcador la verosimilitud de los perfiles bajo la hipótesis alternativa, H_a. Comenzando con el STR D13S317:

$$\begin{aligned}\mathrm{P}(G_H = 9{:}13,\ G_{SP} = 10{:}13 \mid H_a) = \\ = \mathrm{P}(G_H = 9{:}13 \mid G_{SP} = 10{:}13,\ H_a) \cdot \mathrm{P}(G_{SP} = 10{:}13 \mid H_a) = \\ = \mathrm{P}(G_H = 9{:}13 \mid H_a) \cdot \mathrm{P}(G_{SP} = 10{:}13 \mid H_a) = 2p_9p_{13} \cdot 2p_{10}p_{13} = \\ = 4p_9p_{10}(p_{13})^2 = 4 \cdot 0{,}0754 \cdot 0{,}0889 \cdot (0{,}1116)^2 = 0{,}000334.\end{aligned}$$

Para el marcador genético TPOX, esta probabilidad viene dada por:

$$\mathrm{P}(G_H = 8{:}8,\ G_{SP} = 8{:}12 \mid H_a) = (p_8)^2 \cdot 2p_8p_{12} = 2(p_8)^3p_{12} = 2 \cdot (0{,}5539)^3 \cdot 0{,}0373 = 0{,}012677.$$

Y por último, para CSF1PO, se calcula:

$$\mathrm{P}(G_H = 10{:}11,\ G_{SP} = 11{:}11 \mid H_a) = 2p_{10}p_{11} \cdot (p_{11})^2 = 2 \cdot 0{,}2472 \cdot (0{,}2989)^3 = 0{,}013203.$$

Con esto, las probabilidades totales de los perfiles bajo cada una de las hipótesis son:

$$\begin{aligned}\mathrm{P}(G_H, G_{SP} \mid H_0) = 0{,}000748 \cdot 0{,}011444 \cdot 0{,}022085 = 1{,}8905 \times 10^{-7} \\ \mathrm{P}(G_H, G_{SP} \mid H_a) = 0{,}000334 \cdot 0{,}012677 \cdot 0{,}013203 = 5{,}5903 \times 10^{-8}\end{aligned}$$

Ahora, el índice de Essen-Möller se calcula como sigue:

$$W = \frac{\mathrm{P}(G_H, G_{SP} \mid H_0)}{\mathrm{P}(G_H, G_{SP} \mid H_0) + \mathrm{P}(G_H, G_{SP} \mid H_a)} = \frac{1{,}8905 \times 10^{-7}}{1{,}8905 \times 10^{-7} + 5{,}5903 \times 10^{-8}} = 0{,}771780.$$

Que es el mismo resultado obtenido al calcular este índice mediante las razones de verosimilitud:

$$W = \frac{IP_T}{IP_T + 1} = \frac{3{,}38266}{3{,}38266 + 1} = 0{,}7718.$$

Esto quiere decir que, supuesto que ambas hipótesis son equiprobables a priori, esto es, la probabilidad de paternidad antes de haber realizado el contraste es del 50 %; la probabilidad de paternidad tras el análisis de los perfiles de ADN ha aumentado hasta el 77,18 %, con lo que se podría decir que el perfil es coincidente y se confirmaría la hipótesis de paternidad. En la realidad se trabajará con numerosos marcadores genéticos y esta probabilidad aumentará hasta el 99 % en la mayoría de los casos en los que se confirme la hipótesis de paternidad.

3.4. Probabilidades de exclusión

Además de conocer el índice de paternidad y el índice de Essen-Möller, será también interesante conocer la probabilidad de no paternidad. De nuevo cabe recordar que se está suponiendo que todo individuo involucrado en el caso pertenece a una población en *equilibrio de Hardy-Weinberg*.

3.4.1. Exclusión de un hombre cualquiera del caso de paternidad

En algunas situaciones, como los casos de paternidad, será de utilidad conocer la probabilidad de excluir a un hombre cualquiera de la población como posible padre del niño.

Dado que el padre biológico del niño debe compartir al menos un alelo de cada locus con su hijo, se puede calcular, haciendo uso de las frecuencias alélicas en la población, la proporción de individuos que no pueden ser excluidos de la paternidad (NE: probabilidad de no exclusión). De esta forma, se puede obtener la *probabilidad de exclusión*, PE, pues la probabilidad de que un hombre cualquiera no sea descartado como padre es: $NE = 1 - PE$.

Se observa que esta probabilidad, PE, puede ser calculada antes de que haya sido obtenido el genotipo de cualquier supuesto padre en la población; con tan solo el genotipo del niño en el caso del *standard duo*. Esta se calcula de forma diferente dependiendo de si el niño es homocigótico para el locus analizado o si, por el contrario, es heterocigótico.

Si el genotipo del niño para el locus estudiado, l, es $G_H = Y_iY_i$, es decir, es homocigótico para el locus; entonces la probabilidad de descartar a un hombre cualquiera de la población como posible padre vendrá dada por la probabilidad genotípica de aquellos hombres que no posean el alelo Y_i para este locus l, que será $(1-p_i)^2$. Es decir, para este locus donde el hijo es homocigótico, la probabilidad de exclusión es: $PE = (1 - p_i)^2$.

Si se tiene ahora que el genotipo del niño para un locus l' es heterocigótico, es decir, $G_H = Y_iY_j$, entonces la probabilidad de excluir a un hombre de la población como posible padre será aquella probabilidad genotípica correspondiente a los genotipos que no contengan ni el alelo Y_i ni el Y_j. Esto es, para el locus l', en el caso en el que el hijo sea heterocigótico, la probabilidad de exclusión se calcula como:$PE = (1 - p_i - p_j)^2$.

Cuando se considera un conjunto de K loci, como puede ser en un perfil de ADN donde se analizan numerosos marcadores genéticos de forma simultánea, y llamando PE_i a las probabilidades de exclusión de cada locus $i = 1, \ldots, n$; se calcula la probabilidad de exclusión general como:

$$PE = 1 - \prod_{i=1}^{K}(1 - PE_i).$$

Teniendo en cuenta que en general $NE = 1 - PE$ es la probabilidad de que un hombre cualquiera no sea excluido de la paternidad. Se tiene entonces que la probabilidad general de que un hombre no sea descartado, para los K loci analizados en el perfil de ADN del niño, es:

$$NE = \prod_{i=1}^{K}(1 - PE_i).$$

Ejemplo 3.1. Considerando el ejemplo del caso *standard duo*, se puede calcular la probabilidad de exclusión para cada uno de los SRT implicados (D13S317, TPOX y CSF1PO), así como la probabilidad de exclusión total.
Para el marcador D13S317, como el genotipo del niño es $G_H = 9{:}13$, la probabilidad de exclusión para este primer STR se calcula como: $PE_1 = (1 - p_9 - p_{13})^2$, que, empleando las probabilidades alélicas de la tabla 3.6, resulta:

$$PE_1 = (1 - 0{,}0754 - 0{,}1116)^2 = 0{,}66097.$$

Para el segundo marcador genético, TPOX, se ha supuesto que el genotipo del niño es $G_H = 8{:}8$, con lo que la probabilidad de exclusión para este segundo STR es:

$$PE_2 = (1 - p_8)^2 = (1 - 0{,}5539)^2 = 0{,}19901.$$

Por último, para CSF1PO, el genotipo del niño se supone $G_H = 10 : 11$, que se trata de nuevo de un caso heterocigótico; por lo que la probabilidad de exclusión calculada mediante las probabilidades alélicas para este tercer marcador es:

$$PE_3 = (1 - p_{10} - p_{11})^2 = (1 - 0{,}2472 - 0{,}2989)^2 = 0{,}20603.$$

Una vez calculada la probabilidad de exclusión para estos tres marcadores genéticos, se obtienen sus respectivas probabilidades de no exclusión, con el objetivo de conseguir la probabilidad de exclusión general:

$$NE_1 = 1 - PE_1 = 1 - 0{,}66097 = 0{,}33903,$$
$$NE_2 = 1 - PE_2 = 1 - 0{,}19901 = 0{,}80099,$$
$$NE_3 = 1 - PE_3 = 1 - 0{,}20603 = 0{,}79397.$$

Con lo que la probabilidad de no exclusión total viene dada por:

$$NE = NE_1 \cdot NE_2 \cdot NE_3 = 0{,}33903 \cdot 0{,}80099 \cdot 0{,}79397 = 0{,}21561.$$

Y entonces la probabilidad de exclusión general para este ejemplo es:

$$PE = 1 - NE = 1 - 0{,}21561 = 0{,}78439.$$

A la vista de estos resultados, se puede interpretar que hay una probabilidad de 0,78439 de que un hombre cualquiera de la población sea excluido del caso de paternidad dado el genotipo del

niño para los tres STR analizados (D13S317, TPOX y CSF1PO). Dicho de otra forma, hay cerca de un 22 % de probabilidades de que un hombre cualquiera de la población que, en efecto, no es el padre biológico del niño, no sea descartado como posible padre empleando este test de ADN con el genotipo del niño para los tres marcadores genéticos.

Para el caso del *standard trio*, cuando el genotipo de la madre es conocido, esta probabilidad de exclusión se calcula de forma análoga, teniendo en cuenta que cada probabilidad de exclusión para un locus concreto y un caso de paternidad concreto, PE, será la probabilidad de descartar a un hombre cualquiera como posible padre, considerando tanto el genotipo del niño como el de la madre.

Por ejemplo, si se tiene un locus l para el cual el niño es homocigótico, es decir, $G_H = Y_iY_i$, y el genotipo de la madre se supone que es $G_M = Y_iY_j$, siendo Y_i e Y_j dos alelos distintos del locus l; entonces cualquier hombre que no posea el alelo Y_i podrá ser descartado como posible padre. En resumen, la probabilidad de exclusión vendrá dada por: $PE = (1 - p_i)^2$.

En la siguiente tabla se presentan las probabilidades de exclusión, para un locus concreto l, dados los genotipos del niño y de la madre, con Y_i, Y_j e Y_k tres alelos distintos de ese locus. Estas se calculan todas de forma similar. Se presenta desarrollado el caso en el que los genotipos del niño y de la madre son, respectivamente: $G_H = Y_iY_j$, $G_M = Y_iY_j$. En esta situación, el alelo que el niño herede de su madre podrá ser tanto el Y_i como el Y_j, por lo que la probabilidad de exclusión se corresponderá con la de aquellos genotipos que no posean ninguno de estos dos alelos, esto es:

$$PE = (1 - p_i - p_j)^2.$$

G_M	Y_iY_i		Y_iY_j				
G_H	Y_iY_i	Y_iY_j	Y_iY_i	Y_jY_j	Y_iY_j	Y_iY_k	Y_jY_k
PE	$(1-p_i)^2$	$(1-p_j)^2$	$\mathbf{(1-p_i)^2}$	$(1-p_j)^2$	$\mathbf{(1-p_i-p_j)^2}$	$(1-p_k)^2$	$(1-p_k)^2$

Tabla 3.7: Probabilidades de exclusión caso *standard trio* para un locus l.

Ejemplo 3.2. Considerando el ejemplo del caso *standard trio*, se pueden calcular las probabilidades de exclusión para cada uno de los marcadores genéticos analizados (D3S1358, vWA y FGA) además de la probabilidad de exclusión general.

Comenzando con el marcador D3S1358, para el cual se ha supuesto que el genotipo del niño es $G_H = 15{:}17$ y el genotipo de la madre $G_M = 15{:}16$; para calcular la probabilidad de exclusión hay que obtener la probabilidad de los genotipos que no posean el alelo 17 de este STR; pues el alelo que el niño ha heredado de su madre tiene que ser el 15. Es decir, empleando las frecuencias

alélicas de la tabla 3.4, la probabilidad de exclusión para el primer marcador es:

$$PE_1 = (1 - p_{17})^2 = (1 - 0{,}2040)^2 = 0{,}63362.$$

Para el segundo marcador, vWA, se ha supuesto que los genotipos de madre e hijo son $G_H = 18{:}18$, $G_M = 18{:}19$. Entonces se consideran excluidos todos los individuos que no posean el alelo 18 de este STR, es decir, la probabilidad de exclusión para el segundo marcador es:

$$PE_2 = (1 - p_{18})^2 = (1 - 0{,}2107)^2 = 0{,}62300.$$

Finalmente, para FGA, con los genotipos $G_H = 20{:}21$ y $G_M = 20{:}21$, se tiene que el alelo que el niño ha heredado de la madre podría ser tanto el 20 como el 21. Con esto, se excluirá de la paternidad a cualquier hombre que no posea ninguno de estos dos alelos, pues al existir ambas posibilidades para el alelo materno heredado, el alelo heredado del padre puede ser también cualquiera de los dos; es decir, la probabilidad de exclusión para el tercer marcador es:

$$PE_3 = (1 - p_{20} - p_{21})^2 = (1 - 0{,}1551 - 0{,}1725)^2 = 0{,}45212.$$

A la vista de estos resultados, la probabilidad de exclusión total para este caso es:

$$\begin{aligned} PE &= 1 - (1 - PE_1) \cdot (1 - PE_2) \cdot (1 - PE_3) = \\ &= 1 - (1 - 0{,}63362) \cdot (1 - 0{,}62300) \cdot (1 - 0{,}45212) = \\ &= 1 - 0{,}36638 \cdot 0{,}377 \cdot 0{,}54788 = 1 - 0{,}07568 = 0{,}92432. \end{aligned}$$

Esto se puede interpretar como que, dados los genotipos de la madre y del niño para los tres marcadores genéticos, entonces hay una probabilidad de 0,92432 de que un hombre cualquiera de la población sea descartado como posible padre. Es decir, hay cerca de un 8 % de probabilidades de que un hombre cualquiera de la población no relacionado (que realmente no sea el padre biológico del niño) no sea excluido del caso de paternidad empleando el test de ADN para estos tres marcadores genéticos y dados los genotipos del niño y de la madre.

Esto pone fin a la exposición de los métodos de resolución de problemas de paternidad que se presentan de forma detallada en este trabajo. Para finalizar, se introduce como ejemplo otro de los numerosos puntos de vista desde los cuales se pueden abordar los problemas de paternidad: el *poder de exclusión* de un marcador genético. Se presentará el método en el caso *standard duo*.

3.4.2. Poder de exclusión de un marcador genético

Se llama poder de exclusión, $PoEx$, de un marcador genético, a su efectividad como herramienta para resolver un caso de paternidad. Esta puede ser caracterizada mediante su habilidad

para excluir "padres falsos", es decir, la capacidad del marcador para descartar a una persona cualquiera de la población como posible padre. La probabilidad de exclusión, PE, definida anteriormente se refiere a la probabilidad de excluir a un hombre cualquiera de la población, dado un caso de paternidad concreto. En cambio, el poder de exclusión de un marcador genético, $PoEx$, se refiere a su capacidad para excluir a un hombre cualquiera de la población empleando todas las posibles combinaciones para la paternidad, sin restringirse a ningún caso en particular.

Se considera un locus con n alelos: $Y_1, \ldots, Y_n$, cuyas probabilidades alélicas son, respectivamente, $p_1, \ldots p_n$. Entonces, se denomina poder de exclusión individual, PEI, a la proporción de hombres cualesquiera de la población que pueden ser descartados de la paternidad a la vista del genotipo del niño.

Si el genotipo del niño es $G_H = Y_iY_i$, es decir, el niño es homocigótico, entonces cualquier hombre que no posea el alelo Y_i será excluido del conjunto de individuos que podrían ser el padre biológico, con lo que:

$$PEI = \mathrm{P}(\text{un hombre cualquiera se excluye de la paternidad} \mid G_H = Y_iY_i) = (1 - p_i)^2.$$

Además, la probabilidad, en una población en *equilibrio de Hardy-Weinberg*, del genotipo del niño es $\mathrm{P}(G_H = Y_iY_i) = p_i^2$.

Si, por el contrario, el niño es heterocigótico, $G_H = Y_iY_j$, donde Y_i e Y_j son dos alelos diferentes, entonces se tendrá:

$$PEI = \mathrm{P}(\text{un hombre cualquiera se excluye de la paternidad} \mid G_H = Y_iY_j) = (1 - p_i - p_j)^2.$$

Y la probabilidad en la población para el genotipo del niño es $\mathrm{P}(G_H = Y_iY_j) = 2p_ip_j$.

El poder de exclusión ($PoEx$) del marcador genético se obtiene entonces sumando los poderes de exclusión individuales para todas las posibles combinaciones genotípicas del hijo, ponderadas por la probabilidad de este genotipo en la población:

$$PoEx = \sum_{G_H} PEI \cdot \mathrm{P}(G_H) = \sum_{i=1}^{n} (1 - p_i)^2 \cdot p_i^2 + \sum_{i<j} (1 - p_i - p_j)^2 \cdot 2p_ip_j.$$

Donde el primer término se corresponde con el caso en el que el niño es homocigótico y el segundo con el caso heterocigótico. Se puede observar que en este segundo término se pide sumar en $i < j$ para que no aparezcan casos duplicados (el caso $G_H = Y_iY_j$ es el mismo que $G_H = Y_jY_i$).

Para presentar un ejemplo en el que se ilustre el cálculo del poder de exclusión, se toma un marcador genético tipo SNP, que tendrá solamente dos alelos, basado en los datos extraídos de la página web *PharmGKB* [9].

Ejemplo 3.3. Se quiere calcular el poder de exclusión del SNP **rs6311**. Este SNP presenta dos alelos diferentes: T (timina) y C (citosina). Sus probabilidades en una población africana son,

respectivamente, $p_T = 0{,}3910$ y $p_C = 0{,}6090$.
Al existir solamente dos alelos para este SNP, los posibles genotipos para el niño son:

$$G_H = \text{T T},\ G_H = \text{T C},\ G_H = \text{C C}.$$

Donde, en el primer caso, el niño es homocigótico para este SNP, al igual que en el tercer caso; y en el segundo es heterocigótico. Cabe recordar que el genotipo "T C" se considera el mismo que el "C T".

Para el caso 1, cuando el genotipo del niño es homocigótico $G_H = \text{T T}$, los hombres de la población que se excluirán de la paternidad serán aquellos que no posean el alelo T de este SNP. Esto es, $PEI_1 = (1 - p_T)^2$.
Para el caso 3, cuando el genotipo del niño es homocigótico $G_H = \text{C C}$, se tendrá de forma análoga $PEI_3 = (1 - p_C)^2$.
Para el caso en el que el niño es heterocigótico, esto es, su genotipo es $G_H = \text{T C}$, se excluirían de la paternidad aquellos hombres que no posean ni el alelo T ni el alelo C, por lo que, como esto no es posible, en este caso no se excluye a ningún hombre de la posible paternidad ($PEI_2 = 0$).

Las probabilidades asociadas a cada uno de los genotipos en la población son:

$$\text{P}(G_H = \text{T T}) = p_T^2,\ \text{P}(G_H = \text{T C}) = 2p_T p_C,\ \text{P}(G_H = \text{C C}) = p_C^2.$$

Y con esto, empleando la fórmula del poder de exclusión y las probabilidades alélicas mencionadas anteriormente, se tiene que:

$$\begin{aligned} PoEx &= PEI_1 \cdot \text{P}(G_H = \text{T T}) + PEI_2 \cdot \text{P}(G_H = \text{T C}) + PEI_3 \cdot \text{P}(G_H = \text{C C}) = \\ &= (1 - p_T)^2 \cdot p_T^2 + (1 - p_C)^2 \cdot p_C^2 = \\ &= (1 - 0{,}3910)^2 \cdot (0{,}3910)^2 + (1 - 0{,}6090)^2 \cdot (0{,}6090)^2 = 2 \cdot 0{,}05670 = 0{,}1134. \end{aligned}$$

Con esto, el poder de exclusión de este SNP **rs6311** es de un 11,34 %, es decir, este marcador genético puede excluir correctamente a un supuesto padre con una probabilidad de 0,1134. Esta es una probabilidad muy baja, por ello, cuando se trabaja con SNPs en los laboratorios, serán necesarios muchos de ellos para poder obtener resultados concluyentes.

Para calcular el poder de exclusión de un marcador genético tipo STR (*Short Tandem Repeat*) será necesario el uso de un ordenador, pues al estar formados por numerosos alelos, son muchas las posibles combinaciones genotípicas para el niño. En el siguiente capítulo se presentará la resolución del cálculo del poder de exclusión de los marcadores genéticos D13S317, TPOX y CSF1PO, analizados en el ejemplo expuesto para el caso *standard duo*, mediante el paquete *paramlink* del software **R**.

Capítulo 4

Aplicación con el software R

El software **R** [10] es una herramienta muy relevante en programación y análisis estadístico, cuyo empleo está ampliamente extendido tanto a nivel académico como empresarial.

Se trata de un lenguaje informático que forma parte del proyecto GNU (`https://www.gnu.org/`). Es un conjunto de herramientas de software mediante las cuales se pueden manipular, visualizar y calcular conjuntos de datos. El entorno **R** incluye por defecto los *paquetes básicos* y recomendados, que contienen conjuntos de funciones e incluso bases de datos con distintas utilidades; pero se puede extender de forma sencilla ya que existe una gran disponibilidad de paquetes diferentes que se pueden descargar desde el repositorio de **R** *CRAN (Comprehensive R Archive Network).*

En este capítulo se verá el uso del paquete *Familias* [2], que se empleará para realizar los cálculos de los ejemplos vistos en el capítulo anterior. Además, se presentará una introducción al paquete *paramlink* [14], que será de utilidad para calcular probabilidades de exclusión.

El principal objetivo del paquete *Familias*, de Petter Mostad y Thore Egeland, es el cálculo de las razones de verosimilitud para los casos de paternidad expuestos.

4.1. Standard trio

Para introducir este paquete *Familias*, se comienza con el ejemplo del caso *standard trio* (subsección 3.2.3). Se aportará para cada función empleada una breve explicación y se implementará sobre este ejemplo resuelto previamente, comprobando finalmente que los resultados coinciden con lo obtenido.

Primeramente se definen los individuos involucrados en el caso de paternidad junto a su

genealogía, es decir, la forma en la que están relacionados. La función del paquete *Familias* que genera estas genealogías de los individuos involucrados es `FamiliasPedigree(id, dadid, momid, sex)`, donde `id` son los identificadores de los individuos involucrados en el caso, `dadid` indica el padre biológico de cada individuo, `momid` indica la madre biológica de cada uno y `sex` incluye la información acerca del sexo de cada uno de los individuos involucrados.

El siguiente paso para poder calcular la razón de verosimilitudes, de la que se recuerda su fórmula:

$$IP = \frac{\mathrm{P}(G_H \mid G_M, G_{SP}, H_0)}{\mathrm{P}(G_H \mid G_M, H_a)},$$

es generar una lista, en la que se asigne a cada genealogía un nombre que facilite la lectura de la salida de datos.

Una vez introducidos los datos para cada una de las hipótesis, se debe incluir la información de los marcadores genéticos que se van a analizar. Esto se puede conseguir de dos maneras diferentes empleando la función `FamiliasLocus`: de forma manual o desde una base de datos. Esta función tiene tres argumentos básicos: `frequencies`, que son las probabilidades alélicas, `allelenames`, que contiene los nombres de los alelos y `name`, el nombre del marcador genético. En este ejemplo del caso *standard trio* se utiliza la base de datos llamada `NorwegianFrequencies` del paquete *Familias* que contiene información sobre las probabilidades alélicas de un conjunto de marcadores genéticos de una población noruega [1].

Por último, deben incluirse los genotipos de cada individuo involucrado para cada uno de los marcadores analizados y crear una matriz con los datos. Y una vez creados estos elementos, se emplea la función `FamiliasPosterior (pedigrees, loci, datamatrix, prior, ref=1, kinship=0, simplifyMutations=FALSE)` para calcular la razón de verosimilitudes; donde: `pedigrees` es la lista de genealogías a estudiar, `loci` la lista de marcadores genéticos, `datamatrix` contiene los genotipos de cada individuo para cada marcador, `prior` contiene las probabilidades a priori de cada genealogía, `ref` indica la genealogía que debe tomarse como referencia para calcular la razón de verosimilitudes y `kinship = 0` indica una población en *equilibrio de Hardy-Weinberg*.

El código expuesto en el anexo I.1 da lugar a los siguientes resultados:

```
#Resultado de las probabilidades a posteriori y razones de verosimilitud:
resultados
##
## $posterior
##    noPadre      Padre
## 0.05335134 0.94664866
```

```
## $prior
## noPadre   Padre
##     0.5     0.5
##
## $LR
##  noPadre    Padre
##  1.00000 17.74367
## $LRperMarker
##         noPadre    Padre
## D3S1358       1 2.450403
## vWA           1 4.745047
## FGA           1 1.526038
##
## $likelihoods
##      noPadre        Padre
## 2.113538e-11 3.750192e-10
## $likelihoodsPerSystem
##              noPadre        Padre
## D3S1358 0.0007238629 0.0017737560
## vWA     0.0001782231 0.0008456768
## FGA     0.0001638286 0.0002500088
```

En esta lista de resultados se obtienen todos los valores buscados que se calcularon previamente de forma manual en el ejemplo de la subsección 3.2.3. En concreto:

- Los valores "posterior" indican las probabilidades a posteriori de los genotipos G_H, G_M y G_{SP}. El valor bajo el nombre *noPadre* es la verosimilitud de estos datos bajo la hipótesis alternativa, es decir, la no paternidad. El valor bajo el nombre *Padre* es la verosimilitud de estos tres genotipos bajo la hipótesis nula, es decir, el índice de Essen-Möller: $W = 0{,}946679$.
- Los valores "prior" son las probabilidades a priori bajo cada hipótesis, que como se adelantaba, se suponen equiprobables.
- En "LR" se encuentran las razones de verosimilitud, o índices de paternidad, de nuevo para cada una de las hipótesis. Bajo el nombre de *Padre* se encuentra el índice de paternidad total buscado ($IP_T = 17{,}75418$).
- "LRperMarker" se corresponde con los índices de paternidad calculados para cada marcador genético de forma individual. De nuevo, los índices de interés serán los calculados bajo la hipótesis nula, es decir, la hipótesis de paternidad (*Padre*). ($IP_1 = 2{,}45098$, $IP_2 = 4{,}74608$, $IP_3 = 1{,}52625$.)

- Los valores "likelihoods" son las razones de verosimilitud generales calculadas bajo cada una de las hipótesis que se emplearon en el ejemplo para calcular el índice de Essen-Möller, W, ($\mathrm{P}(G_H, G_M, G_{SP} \mid H_a) = 2{,}1135 \times 10^{-11}$, $\mathrm{P}(G_H, G_M, G_{SP} \mid H_0) = 3{,}7475 \times 10^{-10}$).
- Por último, los valores en "likelihoodsPerSystem" son las razones de verosimilitud calculadas para cada uno de los marcadores genéticos involucrados de forma independiente, bajo cada una de las hipótesis del contraste. Para H_a:
  $\mathrm{P}(G_{H1}, G_{M1}, G_{SP1} \mid H_a) = 0{,}000724$, $\mathrm{P}(G_{H2}, G_{M2}, G_{SP2} \mid H_a) = 0{,}000178$ y
  $\mathrm{P}(G_{H3}, G_{M3}, G_{SP3} \mid H_a) = 0{,}000164$.
  Y para H_0:
  $\mathrm{P}(G_{H1}, G_{M1}, G_{SP1} \mid H_0) = 0{,}001774$, $\mathrm{P}(G_{H2}, G_{M2}, G_{SP2} \mid H_0) = 0{,}000845$ y
  $\mathrm{P}(G_{H3}, G_{M3}, G_{SP3} \mid H_0) = 0{,}000250$.

Con lo que se observa que los resultados son análogos a los calculados manualmente en la subsección 3.2.3, salvo errores de redondeo.

4.2. Standard duo

Una vez comprendido el funcionamiento del paquete *Familias*, es sencillo adaptar el código de **R** empleado anteriormente para obtener la resolución del caso *standard duo* (subsección 3.3.2). Bastará tener en cuenta que el genotipo de la madre no está disponible, por lo que ahora los individuos involucrados en este caso de paternidad serán solamente el niño y el supuesto padre. La principal diferencia se encuentra, por lo tanto, en la introducción de los datos para las genealogías. El código que resuelve este ejemplo se puede ver en el Anexo I.2, y los resultados son los siguientes:

```
#Resultado de las probabilidades a posteriori y razones de verosimilitud:
resultados
##
## $posterior
##   noPadre     Padre
## 0.2282214 0.7717786
## $prior
## noPadre   Padre
##     0.5     0.5
##
## $LR
## noPadre   Padre
## 1.00000 3.38171
```

```
## $LRperMarker
##         noPadre     Padre
## D13S317       1 2.2392603
## TPOX          1 0.9026768
## CSF1PO        1 1.6730138
##
## $likelihoods
##      noPadre        Padre
## 5.588195e-08 1.889766e-07
## $likelihoodsPerSystem
##              noPadre        Padre
## D13S317 0.0003343607 0.0007487207
## TPOX    0.0126648490 0.0114322658
## CSF1PO  0.0131964248 0.0220778010
```

Y se puede comprobar que las soluciones son coincidentes con las obtenidas manualmente en la subsección 3.3.2:

Índices de paternidad: $IP_1 = 2{,}24014,\ IP_2 = 0{,}90269,\ IP_3 = 1{,}6728,\ IP_T = 3{,}38266.$
Verosimilitudes de los datos bajo H_0:
$\mathrm{P}(G_{H1}, G_{SP1} \mid H_0) = 0{,}000748,\ \mathrm{P}(G_{H2}, G_{SP2} \mid H_0) = 0{,}011444,$
$\mathrm{P}(G_{H3}, G_{SP3} \mid H_0) = 0{,}022085,\ \mathrm{P}(G_H, G_{SP} \mid H_0) = 1{,}8905 \times 10^{-7}.$
Verosimilitudes de los datos bajo H_a:
$\mathrm{P}(G_{H1}, G_{SP1} \mid H_a) = 0{,}000334,\ \mathrm{P}(G_{H2}, G_{SP2} \mid H_a) = 0{,}012677,$
$\mathrm{P}(G_{H3}, G_{SP3} \mid H_a) = 0{,}013203,\ \mathrm{P}(G_H, G_{SP} \mid H_a) = 5{,}5903 \times 10^{-8}.$
Índice de Essen-Möller: $W = 0{,}7718$.

4.3. Poder de exclusión

El paquete *paramlink* de **R** es un conjunto de diversas herramientas para el análisis de genealogías utilizando datos de marcadores genéticos. Dentro de las múltiples aplicaciones forenses que posee, será de interés el cálculo de probabilidades de exclusión, en concreto del poder de exclusión de un marcador genético. Se presenta como ejemplo el cálculo para los marcadores del caso *standard duo* mencionado en la subsección 3.4.2.

La función encargada del cálculo del poder de exclusión es `exclusionPower`. Esta calcula el poder de exclusión de un marcador genético. Tiene varios argumentos de entrada, algunos de los cuales se definirán empleando otras funciones de este mismo paquete. Los más relevantes son:

- `ped_claim`, que describe la genealogía supuesta para los individuos del problema a través

de la función `nuclearPed(noffs, sex)`, donde `noffs = 1` para indicar que el número de hijos es uno y `sex = 1` para indicar que es hombre.

- `ped_true`, que describe la verdadera relación entre los individuos involucrados. Para ello se emplea la función `singleton(id, sex, ...)` que genera una genealogía de un solo individuo, donde `id` es el identificador del indiviuo y `sex = 1` indica que es hombre.
- `ids` indica los individuos cuyos genotipos están disponibles.
- `alleles` indica el nombre de los alelos del marcador.
- `afreq` indica las probabilidades alélicas en la población, que deben sumar uno.

En el anexo I.3 se presenta el código de **R** empleado para calcular el poder de exclusión de los tres STR (D13S317, TPOX y CSF1PO) analizados en el ejemplo del caso *standard duo* (subsección 3.3.2). Para ello se utiliza de nuevo la base de datos `NorwegianFrequencies`. Los resultados obtenidos son los siguientes:

```
#Resultados poder de exclusión:
#Marcador D13S317:
PoEx1
## 0.4438022
#Marcador TPOX:
PoEx2
## 0.2111522
#Marcador CSF1PO:
PoEx3
## 0.3181626
```

Con lo que:

- D13S317 puede excluir correctamente a un supuesto padre, es decir, a un individuo que no es el padre biológico del niño, con una probabilidad de 0,4438.
- TPOX puede excluir correctamente a un supuesto padre en un 21,12 % de los casos.
- CSF1PO puede excluir correctamente a un supuesto padre con una probabilidad de 0,3182.

Estas probabilidades son bastante bajas. Es por esta razón, entre otras ya mencionadas, que en un caso real se analizarán numerosos marcadores genéticos de forma simultánea dentro de un perfil de ADN para poder aumentar la fiabilidad.

Con este último ejemplo concluye este capítulo de aplicaciones con el software **R**. Esto marca también el final de este trabajo, con el que se ha pretendido dar una visión general de la importancia de las matemáticas en la genética forense, y explicar su aplicación a algunos casos importantes de los problemas de parentesco, así como el uso de **R** para la ejecución de los métodos de resolución del caso *standard duo*, *standard trio*, y las *probabilidades de exclusión*.

Bibliografía

[1] Dupuy, B. M.; Kling, D. and Stenersen, M. (2013), *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, Vol. 4(1).

[2] Egeland, T.; Mostad, P. and Simonsson, I. (2025), *Familias: Probabilities for Pedigrees Given DNA Data,* R package version 2.6.3. URL: `https://CRAN.R-project.org/package=Familias`.

[3] Egeland, T.; Kling, D. and Mostad, P. (2016), *Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics*, Academic Press.

[4] Evett, I.W. and Weir, B.S. (1998), *Interpreting DNA Evidence. Statistical Genetics for Forensic Scientists*, Sinauer Associates.

[5] Fisz, M. (1963), *Probability Theory and Mathematical Statistics*, 3rd ed., John Wiley & Sons.

[6] Fung, W.K. and Hu, Y-Q. (2008), *Statistical DNA Forensics: Theory, Methods and Computation*,John Wiley & Sons.

[7] Klug, W.S.; Cummings, M.R.; Spencer, C.A. and Palladino, M.A. (2013), *Conceptos de genética*, 10th ed., Pearson Educación, S.A.

[8] Pierce, B. (2021), *Genetics Essentials. Concepts and Connections*, 5th ed., W. H. Freeman and Company.

[9] PharmGKB, (2023, 18 de marzo), *Variant PA166154673*, URL: `https://www.pharmgkb.org/variant/PA166154673`.

[10] R Core Team (2021), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

[11] Rohatgi, V.K., and Ehsanes Saleh, A.K. (2015), *An Introduction to Probability and Statistics*, 3rd ed., John Wiley & Sons.

[12] Vélez Ibarrola, R. (2004), *Cálculo de Probabilidades 2*, Ediciones Académicas.

[13] Vélez Ibarrola, R. and García Pérez, A. (1993), *Principios de Inferencia Estadística*, Madrid: Universidad Nacional de Educación a Distancia.

[14] Vigeland, M.D. (2022), *paramlink: Parametric Linkage and Other Pedigree Analysis in R*, R package version 1.1-5. URL: `https://CRAN.R-project.org/package=paramlink`.

Anexo I

Código y salidas de R

I.1. Standard trio

```
#Instalación del paquete:
install.packages("Familias", dependencies = TRUE)
#Se carga la librería:
library(Familias)

#Introducción de los datos:
id<-c("madre","hijo","SP")
sex<-c("female","male","male")
momid<-c(NA,"madre",NA)

#Hipótesis nula, H0 (paternidad):
dadid0<-c(NA,"SP",NA)
ped0<-FamiliasPedigree(id,dadid0,momid,sex)
#Hipótesis alternativa, Ha (no paternidad):
dadid_a<-c(NA,NA,NA)
ped_a<-FamiliasPedigree(id,dadid_a,momid,sex)
#Lista de genealogías:
genealogias<-list(NoPadre=ped0, Padre=ped_a)

#Se carga la base de datos:
data("NorwegianFrequencies")
```

```
#Marcadores contenidos en la base de datos:
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"

#STR analizados: D3S1358, vWA y FGA:
D3S1358<-FamiliasLocus(NorwegianFrequencies$D3S1358, name="D3S1358")
vWA<-FamiliasLocus(NorwegianFrequencies$vWA, name="vWA")
FGA<-FamiliasLocus(NorwegianFrequencies$FGA, name="FGA")
marcadores<-list(D3S1358,vWA,FGA)

#Se incluyen los genotipos en pares ordenados conforme a la lista anterior:
madre<-c(15,16,18,19,20,21) #G_M1=15:16, G_M2=18:19, G_M3=20:21.
hijo<-c(15,17,18,18,20,21) #G_H1=15:17, G_H2=18:18, G_H3=20:21.
SP<-c(17,18,18,18,19,20) #G_SP1=17:18, G_SP2=18:18, G_SP3=19:20.

matrizdatos<-rbind(madre, hijo, SP)
matrizdatos
##
##       [,1] [,2] [,3] [,4] [,5] [,6]
## madre   15   16   18   19   20   21
## hijo    15   17   18   18   20   21
## SP      17   18   18   18   19   20

#Cálculo de las probabilidades a posteriori y razones de verosimilitud:
resultados<-FamiliasPosterior(genealogias, marcadores, matrizdatos)
```

I.2. Standard duo

```
#Semejante al "standard trio" pero sin el genotipo materno:
#Introducción de los datos:
id<-c("hijo","SP")
sex<-c("male","male")
momid<-c(NA,NA)

#Hipótesis nula, H0 (paternidad):
dadid0<-c("SP",NA)
ped0<-FamiliasPedigree(id,dadid0,momid,sex)
#Hipótesis alternativa, Ha (no paternidad):
dadid_a<-c(NA,NA)
ped_a<-FamiliasPedigree(id,dadid_a,momid,sex)
#Lista de genealogías:
genealogias<-list(NoPadre=ped0, Padre=ped_a)

#Se carga la base de datos:
data("NorwegianFrequencies")
#STR analizados: D13S317, TPOX y CSF1PO:
D13S317<-FamiliasLocus(NorwegianFrequencies$D13S317, name="D13S317")
TPOX<-FamiliasLocus(NorwegianFrequencies$TPOX, name="TPOX")
CSF1PO<-FamiliasLocus(NorwegianFrequencies$CSF1PO, name="CSF1PO")
marcadores<-list(D13S317,TPOX,CSF1PO)

#Se incluyen los genotipos en pares ordenados conforme a la lista anterior:
hijo<-c(9,13,8,8,10,11) #G_H1=9:13, G_H2=8:8, G_H3=10:11.
SP<-c(10,13,8,12,11,11) #G_SP1=10:13, G_SP2=8:12, G_SP3=11:11.
matrizdatos<-rbind(hijo, SP)
matrizdatos
##
##      [,1] [,2] [,3] [,4] [,5] [,6]
## hijo    9   13    8    8   10   11
## SP     10   13    8   12   11   11

#Cálculo de las probabilidades a posteriori y razones de verosimilitud:
resultados<-FamiliasPosterior(genealogias, marcadores, matrizdatos)
```

I.3. Poder de exclusión

```
#Instalación del paquete:
install.packages("paramlink", dependencies = TRUE)
#Se carga la librería:
library(paramlink)
#Se carga la base de datos:
data("NorwegianFrequencies")

#Genealogía supuesta: SP es el padre biológico del niño
ped_claim<-nuclearPed(noffs=1, sex=1)
#Genealogía real: SP no es el padre biológico del niño
ped_true<-list(singleton(id=1, sex=1),singleton(id=3, sex=1))
#Individuos cuyos genotipos están disponibles:
ids<-c(1,3)

#Se definen las frecuencias alélicas para los tres STR:
afreq1<-NorwegianFrequencies$D13S317
afreq2<-NorwegianFrequencies$TPOX
afreq3<-NorwegianFrequencies$CSF1PO

#Alelos de cada marcador genético:
alelos1<-names(afreq1)
alelos2<-names(afreq2)
alelos3<-names(afreq3)

#Cálculo del poder de exclusión:
#Marcador D13S317:
PoEx1<-exclusionPower(ped_claim, ped_true, ids, alleles=alelos1, afreq=afreq1)
#Marcador TPOX:
PoEx2<-exclusionPower(ped_claim, ped_true, ids, alleles=alelos2, afreq=afreq2)
#Marcador CSF1PO:
PoEx3<-exclusionPower(ped_claim, ped_true, ids, alleles=alelos3, afreq=afreq3)
```