Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Projecto de fin de carrera

Estimación del coeficiente de endogamia por máxima verosimilitud

Silvia Juanes Márquez

Director: Jan Graffelman

Departamento de Estadística y Investigación Operativa

Resumen

En este proyecto vamos a estimar por el método de máxima verosimilitud frecuencias genotípicas suponiendo distribución multinomial, también estimaremos el coeficiente de endogamia y frecuencias alélicas. Desarrollaremos dos ejemplos dando las estimaciones de los parámetros y medidas de dispersión mediante diferentes métodos. Realizaremos un estudio de simulación de Monte Carlo bajo el dominio de la mínima frecuencia alélica y variando tamaños muestrales, para analizar las propiedades del estimador del coeficiente de endogamia.

Abstract

In this project we will be estimated by the maximum likelihood method assuming multinomial distribution genotypic frequencies also will estimate the inbreeding coefficient and allelic frequencies. We will develop two examples giving parameter estimates and measures of dispersion by different methods. Be a study of Monte Carlo simulation under the range of minor allele frequency and varying sample sizes to analyze the properties of the estimator of the inbreeding coefficient.

Índice general

Capítulo 1
Introducción a la genética

A lo largo de la historia de la humanidad, se utilizó la genética tanto seleccionando los animales domésticos destinados a la cría en la domesticación de animales, como en la agricultura, llevando a cabo la polinización cruzada de cultivos. Solo mucho después, en el siglo XX, se llega a confirmar la teoría de los principios de la genética.

Los seres humanos son difíciles de estudiar ya que, al contrario que con la genética de plantas y animales, los cruces experimentales no son posibles, además hay que tener en cuenta que los factores ambientales son difíciles de controlar, ya que es complicado separarlos de los factores genéticos para una correcta asociación con la enfermedad.

El genoma humano esta compuesto por 23 pares de cromosomas homólogos, uno de cada par proviene de la madre y los otros del padre. Un cromosoma esta constituido por una cadena de ADN de doble hebra compuesta por ácidos y proteínas que forman bases nitrogenadas, unidas entre ellas dos a dos. Los genes son segmentos de esta cadena, y un alelo es cada una de las diversas formas que ocupa la misma posición en cada par de cromosomas homólogos encontrados en la población.

A esta variación entre sujetos en un lugar determinado de la cadena de ADN, se la conoce como polimorfismo, ya sea una sola base, una secuencia o en el número de repeticiones de una secuencia determinada. Estos son los responsables de las diferencias en una población de las expresiones individuales o rasgos, ya sea riesgo de enfermedades, características físicas, diferentes rendimientos, etc...

La composición genética particular de cada individuo lo referimos como genotipo y su expresión queda codificada en el fenotipo.

Utilizamos los términos rasgos y fenotipos para definir las características individuales (color de ojos, de piel, ...)

En el caso de tener un gen con dos alelos A_1 y A_2, los tres posibles genotipos que podríamos encontrar en una población serían A_1A_1 y A_2A_2, con los dos alelos iguales, los llamaríamos homocigotos y el genotipo A_1A_2 que serían heterocigotos, por lo que en este caso el alelo adquirido del padre es diferente que el heredado por parte de la madre. Si tuvieramos 3 alelos (A_1, A_2 y A_3), dispondriamos de $\frac{3\times4}{2} = 6$

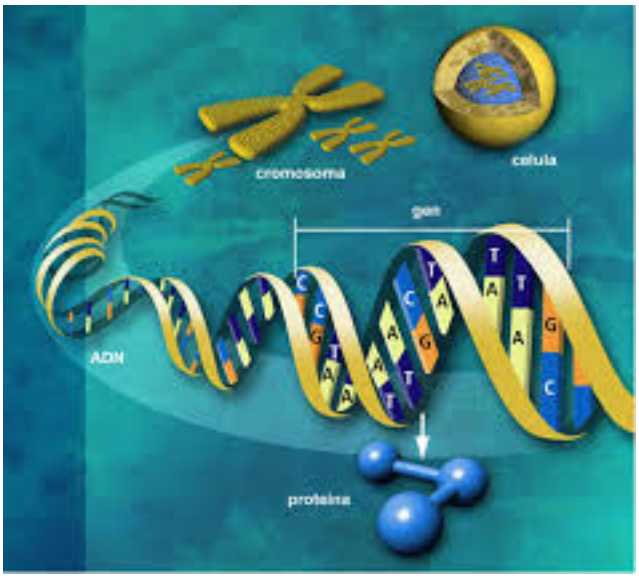

FIGURA 1. Genoma humano

posibles genotipos (A_1A_1, A_1A_2, A_2A_2, A_2A_3, A_3A_3 y A_1A_3). Generalizando, para un gen con k alelos, existen $\frac{k(k+1)}{2}$ genotipos posibles.

En una población, podemos encontrar polimorfismos o variaciones entre individuos, la variable que recoge los diferentes genotipos posibles en un locus concreto se llama marcador. Estos son interesantes, ya que si estudiamos su asociación con un determinado rasgo podremos averiguar cuales son los factores genéticos que afectan a dicho rasgo.

Estadística genética es una rama de la estadística que trata y analiza datos genéticos (material génetico heredado) para obtener información sobre rasgos o fenotipos heredados y enfermedades relacionadas con éstos.

El objetivo de este proyecto es el estudio de una medida de desviación del equilibrio de Hardy-Weinberg que se llama coeficiente de endogamia. Para ello estimaremos las frecuencias alélicas y dicho coeficiente mediante el método de máxima verosimilitud. A partir de un estudio de simulación, veremos propiedades interesantes del coeficiente de endogamia. También propondremos un par de ejemplos que ilustren el uso de este coeficiente con datos de estudios de genotipado.

En el capítulo 2 nos centraremos en los marcadores y citaremos algunos de los tipos más utilizados, después procederemos en el capítulo 3 a definir el equilibrio de Hardy-Weinberg y el llamado coeficiente de endogamia. En el cuarto capítulo estimaremos por máxima verosimilitud las frecuencias genotípicas, frecuencias alélicas y el coeficiente de endogamia a partir de la distribución multinomial. Aclararemos la relación directa entre coeficiente de endogamia y heterocigosidad en el capítulo 5. En el capítulo 6 hallaremos las estimaciones maximizando la función de verosimilitud en entorno R. El capítulo 7 contiene dos ejemplos numéricos y por último, el

número 8 esta dedicado al estudio de las propiedades del estimador del coeficiente de endogamia. El capítulo 9 contiene la discusión y las principales conclusiones del trabajo. Y por último, se presenta el código R más relevante en un apéndice y se completa esta memoria con la bibliografía.

Capítulo 2
Marcadores genéticos

Los últimos estudios del proyecto genoma humano estiman que existen alrededor de 28.000 genes en el genoma humano, con más de 3200 millones de pares de bases distribuidas en los genes, algunos genes son pequeños, de unos pocos miles de pares de bases, y otros de millones.

Por ello, entre tanto material genético, los marcadores son esenciales para saber qué polimorfismos infieren en según que rasgo y poder entender en el futuro la codificación de cada secuencia del ADN.

Existen varios tipos de marcadores, entre ellos:

SNP (o Single nucleotide polymorphism): Es el marcador más utilizado. Consiste en un poliformismo de un único par de nucleótidos situados en un locus determinado de la cadena de ADN.

Existen 4 nucleótidos Adenina (A), Tinina (T), Guanina (G) y Citosina (C), que combinados pueden dar lugar a $4 \times 4 = 16$ posibles genotipos que formarán una variable categórica, o SNP. Aunque en la práctica la mayoría de los SNPs son bi-alélicos, por lo que únicamente observaríamos tres genotipos diferentes. Por ejemplo, supongamos que tenemos un polimorfismo A / G, los posibles genotipos serían AA, AG y GG.

Una de las claves para que los SNPs sean tan populares es el bajo coste de la genotipificación de SNPs múltiples.

En la siguiente figura podemos distinguir un SNP en un segmento de la cadena de ADN:

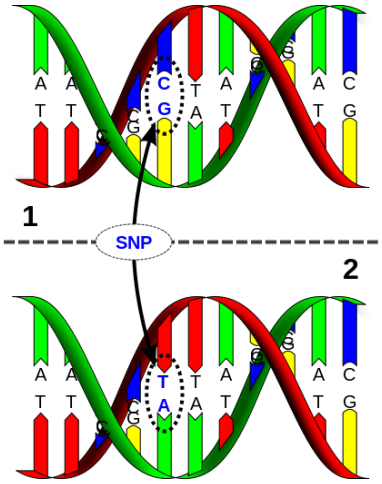

FIGURA 1. Ejemplo de un SNP, un polimorfismo A/G.

RFLP (Restriction fragment length polymorphism): Existe un gran número de enzimas de restricción que cortan el ADN de diferentes formas según el individuo, provocando polimorfismos en la población.

Para ser analizados, son separados en un gel en el laboratorio, se recoge la información en forma de datos binarios y así se puede inferir la presencia o ausencia de sitios de restricción.

Un ejemplo de este tipo de enzima es la BamHI, que corta el ADN en la secuencia de reconocimiento GGATCC / CCTAGG.

Este marcador se utiliza mucho menos que un SNP, ya que su estudio es mucho más preciso, y se centra más concretamente en un solo par de nucleótidos.

Microsatélites, SSR o STR: Consiste en una secuencia corta, que se repite un determinado número de veces. Este marcador junto con el SNP, son los más recurridos para estudios genéticos.

Por ejemplo, la secuencia corta podría ser ATT, la cual, en ATTATTATTATT se repite 4 veces.

Indel: Un polimorfismo de inserción / deleción, comúnmente abreviado *indel* es un tipo de variación genética en la que una secuencia nucleotídica específica está presente (inserción) o ausente (deleción). Si bien no es tan común como SNPs, los indeles se encuentran ampliamente distribuidos en todo el genoma. En las regiones codificantes del genoma, a menos que la longitud de un INDEL sea múltiplo de 3, se producirá una mutación puntual, y dará como resultado una proteína con aminoácidos adicionales (inserción) o pérdida de aminoácidos (supresión).

Capítulo 3
Equilibrio de Hardy-Weinberg

El equilibrio de Hardy-Weinberg es uno de los fundamentos de la genética moderna, muy relevante para el estudio de los marcadores geneticos, este capítulo detallaremos este principio.

Supongamos que tenemos dos tipos de alelos A y B, los cuales según los alelos heredados por parte de padre y madre, pueden dar lugar a los genotipos AA, AB y BB.

La ley de Hardy-Weinberg afirma que para un marcador bialélico, con alelos A y B (con frecuencias p y q respectivamente), las frecuencias genotípicas relativas de AA, AB y BB se distribuyen según la siguiente tabla:

	A	B
A	p^2	pq
B	pq	q^2

Tabla 1. Frecuencias genotípicas relativas

Y como no distinguiremos entre el orden de los alelos (entre n_{AB} y n_{BA}), la frecuencias genotípicas relativas serían:

$P(AA) = p^2$
$P(AB) = 2pq$
$P(BB) = q^2$

Definimos también las frecuencias genotípicas absolutas como:

n_{AA} = Número de individuos con el genotipo AA
n_{AB} = Número de individuos con el genotipo AB
n_{BB} = Número de individuos con el genotipo BB

Cuyas esperanzas son las frecuencias genotípicas relativas multiplicadas por el tamaño de la muestra.

Para demostrar la existencia de HWE han de cumplirse las condiciones anteriores, ésto implica que a lo largo del tiempo, éstas probabilidades de heredar cualquier genotipo, son constantes de generación en generación.

Podemos obtener la frecuencia alélica para la próxima generación (p') mediante el siguiente cálculo:

$p' = (2p^2 + 2pq)/2 = p$

la frecuencia alélica es también p. Por lo que efectivamente, las frecuencias alélicas bajo HWE no varían entre generaciones. Y como consecuencia las frecuencias genotípicas AA, AB y BB también se mantienen en las mismas proporciones a lo largo de las generaciones.

Esto significa que podemos determinar la frecuencia de un alelo en concreto de la siguiente forma:

$n_A = 2n_{AA} + n_{AB}$ número de alelos A en la muestra
$n_B = 2n_{BB} + n_{AB}$ número de alelos B en la muestra

Hay muchos supuestos requeridos para el cumplimiento del equilibrio de Hardy-Weinberg: el apareamiento ha de ser al azar, sin presencia de endogamia, suponemos población infinita, con generaciones discretas, las frecuencias de los alelos han de ser iguales en hombres y mujeres, y no puede haber ningún tipo de factor de mutación, migración, o de selección. A pesar de que ninguno de estos supuestos es probable que se mantenga exactamente en cualquier población, el principio de Hardy-Einberg suele proporcionar una buena aproximación para las frecuencias genotípicas en una población.

1. Coeficiente de endogámia

El coeficiente de endogamia (f) es una manera de parametrizar la desviación del equilibrio de HW. Cuando $f = 0$, las frecuencias genotípicas son las frecuencias de HW.

En una población se puede dar un fenómeno conocido como la endogámia. Éste ocurre o bien por un aislamiento geográfico o cuando hay una preferencia para el apareamiento. En cualquier caso, existe la posibilidad de que una descendencia herede dos copias del mismo alelo ancestral. A esta probabilidad de que un individuo al azar en la población herede dos copias del mismo alelo de un ancestro común lo llamaremos coeficiente de endogamia. Aunque en genética de poblaciones tenga este significado, cualquier otro factor que provoque un desvio del HWE que conduzca a un f distinto de cero también se identificará como tal. De modo que podemos definir el coeficiente de endogamia como una medida general de desequilibrio.

En las grandes poblaciones de apareamiento al azar las posibilidades de que cualquiera de los dos padres tengan un alelo ancestro común es baja, por lo tanto, f será insignificante, incluso cero.

Las probabilidades de heredar los diferentes genotipos bajo el supuesto de endogamia, se reparten según la siguiente tabla:

	A	B	
A	$p^2+\epsilon$	$pq-\epsilon$	p
B	$qp-\epsilon$	$q^2+\epsilon$	q
	p	q	1

TABLA 2. Frequencias genotípicas con desviación ϵ de HWE.

Por lo que las estimaciones de las frecuencias genotípicas bajo HWE se modificarán obteniéndose las siguientes probabilidades de obtener los distintos genotipos:

Genotipo	AA	AB	BB
P(Genotipo)	p^2+fpq	$2pq(1-f)$	q^2+fpq

TABLA 3. Frecuencias genotípicas bajo desequilibrio parametrizado con coeficiente de endogamia f.

Si el coeficiente de endogamia resultante es $f > 0$ significa que hay un exceso de homocigotos, y sin embargo, si resulta $f < 0$ existe un exceso de heterocigotos.

Para definir el rango de f, antes definiremos *Minor Allele Frequency* (MAF) como la más pequeña frecuencia alélica relativa, que sería el mínimo entre p y q, es decir, $p_m = min(p, q)$. El dominio de f es $\frac{-p_m}{1-p_m} \leq f \leq 1$. Por lo tanto, el menor valor posible depende de los valores de las frecuencias alélicas relativas.

En el siguiente capítulo hallaremos las estimaciones de los parámetros necesarios para el cálculo de estas probabilidades bajo dicho desequilibrio.

Capítulo 4
Estimación de parámetros

1. Caso multinomial

1.1. Función de verosimilitud. Para hallar las estimaciones máximas verosimiles de las frecuencias genotípicas que llamaremos g_i en el caso de la distribución multinomial, siendo i=$\{$AA, AB, BB$\}$ y $\Sigma_i n_i = N$, empezaremos por mostrar su función máxima verosimil:

$$f(n_i|g_i) = N! \prod_i \frac{g_i^{n_i}}{n_i!} \tag{1}$$

1.2. Estimación de parámetros. Seguidamente le aplicamos el logaritmo:

$$lnL = ln(f(n_i|g_i)) = ln(N!) - \Sigma_i ln(n_i!) + \Sigma_i n_i ln(g_i)$$

Si derivamos respecto g_i, no podremos hallar el estimador, por lo que añadiremos un multiplicador λ de la siguiente forma:

$$lnL = ln(f(n_i|g_i)) = ln(N!) - \Sigma_i ln(n_i!) + \Sigma_i n_i ln(g_i) + \lambda(1 - \Sigma_i g_i)$$

Y ahora sí, podremos derivar, y mediante el siguiente sistema de ecuaciones hallaremos las estimaciones:

$$\left.\begin{array}{r} \frac{\partial L}{\partial g_i} = \frac{n_i}{g_i} - \lambda = 0 \\ \Sigma_i g_i = 1 \end{array}\right\} g_i = \frac{n_i}{\lambda}$$

$$g_i = \frac{n_i}{\lambda}$$
$$\Sigma_i g_i = \frac{\Sigma_i n_i}{\lambda}$$
$$1 = \frac{\Sigma_i n_i}{\lambda}$$
$$\lambda = \Sigma_i n_i = N$$

Los estimadores de las frecuencias genotípicas g_i son:

$$\hat{g}_i = \frac{n_i}{N} \tag{2}$$

Por lo tanto, los estimadores máximos verosímiles de las frecuencias genotípicas son las frecuencias genotípicas muestrales.

2. Distribución con parametros p y f

2.1. Función de verosimilitud. En este apartado expresaremos la función verosimil de la distribución multinomial (1) en función del coeficiente de endogamia (f) y la proporción alélica de A (p).

$$L(p,q,f|x) = \frac{N!}{n_{AA}!n_{AB}!n_{BB}!}(p^2+pqf)^{n_{AA}}(2pq(1-f))^{n_{AB}}(q^2+pqf)^{n_{BB}} \tag{3}$$

2.2. Estimación de p y f. Hacemos el logaritmo de la anterior función de verosimilitud (3):

$$lnL(p,q,f|x) = K + n_{AA}ln(p^2+pqf) + n_{AB}ln(2pq(1-f)) + n_{BB}ln(q^2+pqf)$$

donde $K = lnN! - (ln(n_{AA}!) + ln(n_{AB}!) + ln(n_{BB}!)$

Reescribimos la fórmula únicamente en función de p y f, teniendo en cuenta que $q = 1-p$

$$lnL(p,q,f|x) = K + n_{AA}ln(p^2+p(1-p)f) + n_{AB}ln(2pq(1-f)) + n_{BB}ln(q^2+q(1-q)f) =$$
$$= K + n_{AA}ln(p^2+pf-p^2f) + n_{AB}ln(2pq(1-f)) + n_{BB}ln(q^2+qf-q^2f)$$

(4)
$$lnL(p,f|x) = K+n_{AA}ln(p^2(1-f)+pf)+n_{AB}ln(2p(1-p)(1-f))+n_{BB}ln((1-p)^2(1-f)+(1-p)f)$$

Procedemos ahora a derivar (4). La derivada de K respecto p es 0, la del primer sumando sería:

$$\frac{dn_{AA}ln(p^2(1-f)+pf)}{dp} = n_{AA}\frac{2p(1-f)+f}{p^2(1-f)+pf}$$

dividimos por $1-f$ de manera que podremos expresar $\alpha = \frac{f}{1-f}$

$$\frac{dn_{AA}ln(p^2(1-f)+pf)}{dp} = n_{AA}\frac{2p+\alpha}{p(p+\alpha)}$$

la derivada de la segunda parte:

$$\frac{dn_{AB}ln(2p(1-p)(1-f))}{dp} = \frac{n_{AB}2(1-f)-4p(1-f)}{2p(1-p)(1-f)} = n_{AB}\frac{2(1-2p)}{2p(1-p)} = n_{AB}\frac{(1-2p)}{p(1-p)}$$

y la derivada de la tercera parte:

$$\frac{dn_{BB}ln((1-p)^2(1-f)+(1-p)f)}{dp} = n_{BB}\frac{-2(1-p)(1-f)-f}{(1-p)^2(1-f)+(1-p)f} =$$

$$= n_{BB}\frac{-2q(1-f)-f}{q^2(1-f)+qf} = n_{BB}\frac{-2q-\frac{f}{1-f}}{q(q+\frac{f}{1-f})} = n_{BB}\frac{-2q-\alpha}{q(q+\alpha)}$$

Por lo que la derivada de la función de verosimilitud respecto p es:

$$\frac{dlnL}{dp} = n_{AA}\frac{2p+\alpha}{p(p+\alpha)} + n_{AB}\frac{1-2p}{pq} - n_{BB}\frac{2q+\alpha}{q(q+\alpha)} = 0 \tag{5}$$

A continuación hallamos la derivada de la función de verosimilitud (1) respecto f:

$$\frac{dlnL}{df} = n_{AA}\frac{p(1-p)}{p^2(1-f)+pf} + n_{AB}\frac{-2p(1-p)}{2p(1-p)(1-f)} + n_{BB}\frac{qp}{q^2(1-f)+qf} =$$

$$\frac{dlnL}{df} = n_{AA}\frac{q}{p(1-f)+f} - \frac{n_{AB}}{(1-f)} + n_{BB}\frac{p}{q(1-f)+f} = 0 \tag{6}$$

Si multiplicamos por $(1-f)$ y substituimos α nos quedaría:

$$\frac{dlnL}{df} = n_{AA}\frac{q}{p+\alpha} - n_{AB} + n_{BB}\frac{p}{q+\alpha} = 0 \tag{7}$$

Ahora multiplicamos por $\frac{1-2p}{pq}$, el factor de n_{AB}:

$$n_{AA}\frac{q(1-2p)}{(p+\alpha)pq} - n_{AB}\frac{1-2p}{pq} + n_{BB}\frac{p(1-2p)}{(q+\alpha)pq} = 0$$

$$n_{AB}\frac{(1-2p)}{pq} = n_{AA}\frac{(1-2p)}{(p+\alpha)p} + n_{BB}\frac{(1-2p)}{(q+\alpha)q}$$

De manera que la podemos substituir en la equacion (5):

$$n_{AA}\frac{2p+\alpha}{p(p+\alpha)} + n_{AA}\frac{1-2p}{p(p+\alpha)} + n_{BB}\frac{1-2p}{q(q+\alpha)} - n_{BB}\frac{2q+\alpha}{q(q+\alpha)} =$$
$$= n_{AA}\frac{2p+\alpha+1-2p}{p(p+\alpha)} + n_{BB}\frac{1-2p-2q-\alpha}{q(q+\alpha)} = n_{AA}\frac{\alpha+1}{p(p+\alpha)} + n_{BB}\frac{-(1+\alpha)}{q(q+\alpha)} = 0$$

$$n_{AA}\frac{1+\alpha}{p(p+\alpha)} = n_{BB}\frac{1+\alpha}{q(q+\alpha)}$$

$$\frac{qn_{AA}}{p+\alpha} = \frac{pn_{BB}}{q+\alpha} \tag{8}$$

Por lo que teniendo en cuenta la expresión de la ecuación (7), podemos concluir:

$$\frac{qn_{AA}}{p+\alpha} = \frac{n_{AB}}{2} = \frac{pn_{BB}}{q+\alpha} \tag{9}$$

Con la primera igualdad de (9):

$$2(1-p)n_{AA} = (p+\alpha)n_{AB}$$
$$2n_{AA} - 2pn_{AA} = pn_{AB} + \alpha n_{AB}$$
$$\alpha n_{AB} = 2n_{AA} - 2pn_{AA} - pn_{AB}$$

$$\alpha n_{AB} = 2n_{AA} - pn_A \tag{10}$$

Y utilizando la segunda igualdad de (9):

$$2pn_{BB} = (1-p+\alpha)n_{AB}$$
$$2pn_{BB} = n_{AB} - pn_{AB} + \alpha n_{AB}$$
$$\alpha n_{AB} = 2pn_{BB} + pn_{AB} - n_{AB}$$

$$\alpha n_{AB} = pn_B - n_{AB} \tag{11}$$

Igualamos las ecuaciones (10) y (11):

$$2n_{AA} - pn_A = pn_B - n_{AB}$$
$$pn_A + pn_B = 2n_{AA} + n_{AB}$$
$$p(n_A + n_B) = 2n_{AA} + n_{AB}$$
$$p2N = 2n_{AA} + n_{AB}$$

Y teniendo en cuenta que $n_A + n_B = 2N$, la estimación de p es:

$$\hat{p} = \frac{2n_{AA} + n_{AB}}{2N} \tag{12}$$

Por último, si substituimos la estimación de p en la ecuacion (11):

$$\alpha n_{AB} = \frac{2n_{AA} + n_{AB}}{2N} n_B - n_{AB}$$
$$\alpha n_{AB} = \frac{n_A n_B}{2N} - n_{AB}$$
$$\alpha = \frac{n_A n_B}{2N n_{AB}} - 1$$

Y sabiendo que $f = \frac{\alpha}{\alpha+1}$:

$$f = \frac{\frac{n_A n_B}{2N n_{AB}} - 1}{\frac{n_A n_B}{2N n_{AB}}}$$

$$f = \frac{n_A n_B - 2N n_{AB}}{n_A n_B} \tag{13}$$

$$f = \frac{(2n_{AA} + n_{AB})(2n_{BB} + n_{AB}) - 2N n_{AB}}{n_A n_B}$$

$$f = \frac{4n_{AA}n_{BB} + 2n_{AA}n_{AB} + 2n_{AB}n_{BB} + n_{AB}^2 - 2n_{AB}n_{AA} - 2n_{AB}^2 - 2n_{AB}n_{BB}}{n_A n_B}$$

La estimación de f es:

$$\hat{f} = \frac{4n_{AA}n_{BB} - n_{AB}^2}{n_A n_B} \tag{14}$$

Las varianzas para ambos estimadores se han obtenido a partir del libro Genetic Data Analisis II de Weir [**3**]:

$$V(\hat{p}) = \frac{p(1-p)(1+f)}{2n} \tag{15}$$

$$V(\hat{f}) = \frac{(1-f)^2(1-2f)}{N} + \frac{f(1-f)(2-f)}{2Np(1-p)} \tag{16}$$

El gradiente definido como:

$$\nabla lnL = (\frac{\partial lnL}{\partial p}, \frac{\partial lnL}{\partial f})$$

donde:

$$\frac{\partial lnL}{\partial p} = \frac{n_{AA}(2p(1-f)+f)}{p^2(1-f)+pf} + \frac{n_{AB}(1-2p)}{p(1-p)} - \frac{n_{BB}(2(1-p)(1-f)+f)}{(1-p)^2(1-f)+(1-p)f}$$

$$\frac{\partial lnL}{\partial f} = \frac{n_{AA}(1-p)}{p(1-f)+f} - \frac{n_{AB}}{1-f} + \frac{n_{BB}p}{(1-p)(1-f)+f}$$

A continuación hallaremos la hessiana, que tiene la forma:

$$H(lnL) = \begin{pmatrix} \frac{\partial^2 lnL}{\partial p^2} & \frac{\partial^2 lnL}{\partial p \partial f} \\ \frac{\partial^2 lnL}{\partial p \partial f} & \frac{\partial^2 lnL}{\partial f^2} \end{pmatrix}$$

Empezamos por la doble derivada respecto de p a partir de la ecuación (5):

$$\frac{\partial^2 lnL}{\partial p^2} = \frac{2n_{AA}p(p+\alpha) - (2p+\alpha)n_{AB}(2p+\alpha)}{p^2(p+\alpha)^2} + \frac{-2n_{AB}p(1-p) - (1-2p)n_{AB}(1-2p)}{p^2(1-p)^2} - \frac{(-2n_{BB})(1-p)(1-p+\alpha) - (2(-1+p)-\alpha)n_{BB}(2(1-p)+\alpha)}{(1-p)^2(1-p+\alpha)^2}$$

cambiamos (-1+p) por -(1-p) y simplificamos:

$$\frac{n_{AA}(2p(p+\alpha) - (2p+\alpha)^2)}{p^2(p+\alpha)^2} - \frac{n_{AB}(1-2p(1-p))}{p^2(1-p)^2} - \frac{n_{BB}((2(1-p)+\alpha)^2 - 2(1-p)((1-p)+\alpha))}{(1-p)^2(1-p+\alpha)^2}$$

y substituyendo $\alpha = \frac{f}{1-f}$ nos queda:

$$\frac{\partial^2 lnL}{\partial p^2} = \frac{n_{AA}(2p(p+\frac{f}{1-f}) - (2p+\frac{f}{1-f})^2)}{p^2(p+\frac{f}{1-f})^2} - \frac{n_{AB}(1-2p(1-p))}{p^2(1-p)^2} - \frac{n_{BB}((2(1-p)+\frac{f}{1-f})^2 - 2(1-p)((1-p)+\frac{f}{1-f}))}{(1-p)^2(1-p+\frac{f}{1-f})^2}$$

si multiplicamos arriba y abajo por $(1-f)^2$ en el primer y tercer término nos queda:

$$\frac{\partial^2 lnL}{\partial p^2} = \frac{n_{AA}(2(1-f)(p^2(1-f)+pf) - (2p(1-f)+f)^2}{p^2(p(1-f)+f)^2} - \frac{n_{AB}(1-2p(1-p))}{p^2(1-p)^2} - \frac{n_{BB}(2(1-p)(1-f)+f)^2 - 2(1-f)((1-p)^2(1-f)+(1-p)f)}{(1-p)^2((1-p)(1-f)+f)^2}$$

Para calcular $\frac{\partial^2 lnL}{\partial f^2}$ recurrimos a la primera derivada de la función de verosimilitud respecto f, pero antes de hacer el cambio de variable con α (6) y volvemos a derivar:

$$\frac{\partial^2 lnL}{\partial f^2} = \frac{-n_{AA}(1-p)(-p+1)}{(p(1-f)+f)^2} - \frac{-n_{AB}(-1)}{(1-f)^2} + \frac{-n_{BB}p(-(1-p)+1)}{((1-p)(1-f)+f)^2}$$

$$\frac{\partial^2 lnL}{\partial f^2} = -\frac{n_{AA}((1-p)^2)}{(p(1-f)+f)^2} - \frac{n_{AB}}{(1-f)^2} - \frac{n_{BB}p^2}{((1-p)(1-f)+f)^2}$$

Y los otros dos términos de la matriz hessiana derivando respecto p a partir de la ecuación (6):

$$\frac{\partial^2 lnL}{\partial p \partial f} = -\frac{n_{AA}(p(1-f)+f+(1-f)q)}{(p(1-f)+f)^2} + \frac{n_{BB}(q(1-f)+f+(1-f)p)}{(q(1-f)+f)^2} = \\ = -\frac{n_{AA}}{(p(1-f)+f)^2} + \frac{n_{BB}}{(q(1-f)+f)^2}$$

Para verificarla derivaremos también la ecuacion (5) respecto f:

$$\frac{\partial^2 lnL}{\partial p \partial f} = \left(\frac{n_{AA}p(p+\alpha) - pn_{AA}(2p+\alpha)}{p^2(p+\alpha)^2} - \frac{n_{BB}q(q+\alpha) - qn_{BB}(2q+\alpha)}{q^2(q+\alpha)^2})d\alpha \right)$$

$$= \left(\frac{n_{AA}p(p+\alpha-2p-\alpha)}{p^2(p+\alpha)^2} + \frac{n_{BB}q(q+\alpha-2q-\alpha)}{q^2(q+\alpha)^2} \right) d\alpha =$$

$$= \left(\frac{n_{BB}}{(q+\alpha)^2} - \frac{n_{AA}}{(p+\alpha)^2} \right) d\alpha$$

Hacemos el cambio de variable sabiendo que $d\alpha = \frac{df}{(1-f)^2}$

$$\frac{\partial^2 lnL}{\partial p \partial f} = \left(\frac{n_{BB}}{(q+\frac{f}{1-f})^2} - \frac{n_{AA}}{(p+\frac{f}{1-f})^2} \frac{1}{(1-f)^2} \right) df \\ = \frac{n_{BB}}{(q(1-f)+f)^2} - \frac{n_{AA}}{(p(1-f)+f)^2}$$

Esta matriz nos será útil para calcular la varianza mínima de los estimadores y su covarianza.

Capítulo 5
Relación entre el coeficiente de endogamia y heterocigosidad

El coeficiente de endogamia f también podríamos definirlo como la probabilidad de que un individuo de una muestra aleatoria herede la misma copia de ambos padres si el valor de f es positivo.

En este apartado reescribiremos la estimación de f de manera que podamos observar claramente la relación entre heterocigosidad y el coeficiente de endogamia.

Recurriendo a la ecuación (13):

$$f = \frac{n_A n_B - 2N n_{AB}}{n_A n_B} = 1 - \frac{2N n_{AB}}{n_A n_B}$$

Como ya habiamos utilizado, sustituiremos $n_B = 2N - nA$:

$$f = 1 - \frac{2N n_{AB}}{n_A(2N - n_A)}$$

Y de la estimación de p (12) que expresaremos de la siguiente forma:

$$p = \frac{n_A}{2N}$$

Sustituimos n_A por $2Np$:

$$f = 1 - \frac{2N n_{AB}}{2Np(2N - 2Np)}$$

$$f = 1 - \frac{n_{AB}}{2np(1-p)}$$

Dejándonos con la ecuación:

$$f = 1 - \frac{n_{AB}}{2npq} \tag{17}$$

Por lo que el coeficiente de endogamia es uno menos el cociente entre la heterocigosidad observada y la esperada. Sabiendo que en una distribución binomial $P(A \cap B) = P(A)P(B) = pq$, la probabilidad de que tengan un alelo A y uno B

$$P(A \cap B) + P(B \cap A) = 2(P(A)P(B) = 2pq.$$

Si nos fijamos en la tabla (3):

$$NP(AB) = 2Npq(1 - f)$$

Vemos como coincide con la esperanza de n_{AB} o heterocigosidad esperada.

Si f es constante de generación en generación, se puede aplicar la ley del equilibrio de Sewall Wright, el cual nos expone que f es positiva si la heterocigosidad observada es menos que la esperada y f es negativa si ocurre justo lo contrario. El coeficiente f será 0 si ambas son iguales.

Capítulo 6
Estimación de parámetros por métodos numéricos en R

En el capítulo 5 hemos obtenido fórmulas explícitas para los estimadores de máxima verosimilitud de p y f. En este capítulo trabajaremos la estimación de estos parámetros por métodos numéricos.

Hallaremos las estimaciones de los parámetros p y f usando la función *nlminb* de R. Para ello primero creamos la función que da como resultado el valor negativo de la función de máxima verosimilitud según f y p.

```
logver <- function(par,x){
  p <- par[1]
  f <- par[2]

  pi<-rep(NA,3)

  pi[1] <- p^2+p*(1-p)*f
  pi[2] <- 2*p*(1-p)*(1-f)
  pi[3] <- (1-p)^2+p*(1-p)*f

  vo <- -sum(x*log(pi))
  return(-vo)
}
```

El valor obtenido tiene que ser negativo porque la función que introduciremos como argumento objective de *nlminb* será minimizada.

```
nlminb(c(1/2,0), logver, x, lower = -1, upper = 1,
control=list(trace=TRUE))
```

Al ejecutar la orden se estiman los parametros p y f, minimizando el valor negativo que se obtiene con la función anterior *logver*.

También adjuntamos la función mediante la que obtendríamos el vector gradiente:

```
hwe.gradiente <- function(p,x) {
  pa <- p[1]
  f <- p[2]

  dldpa <- x[1]*(2*pa*(1-f)+f)/(pa^2*(1-f)+pa*f) +
    x[2]*(1-2*pa)/(pa*(1-pa)) -
    x[3]*(2*(1-pa)*(1-f)+f)/(((1-pa)^2)*(1-f)+(1-pa)*f)

  dldf <- x[1]*(1-pa)/(pa+(1-pa)*f) - x[2]/(1-f) + x[3]*pa/((1-pa)*(1-f) + f)
  return(c(dldpa,dldf))
}
```

Y la función con la que calcular la matriz hessiana:

```
hwe.hessiana <- function(ps,x) {
   pa <- ps[1]
   f <- ps[2]

  dldpdp <- x[1]*(2*(1-f)*(pa^2+pa*(1-pa)*f) - ((2*pa*(1-f)+f)^2))/
     ((pa^2 + pa*(1-pa)*f)^2)
   - x[2]*(1-2*pa*(1-pa))/((pa^2)*((1-pa)^2))
   - x[3]*((2*(1-pa)+f/(1-f))^2-2*(((1-pa)^2)+(1-pa)*(f/(1-f))))/
     (((1-pa)^2)*((1-pa+f/(1-f))^2))

  dldpdf <- -x[1]/((pa + (1-pa)*f)^2) + x[3]/(((1-pa)*(1-f)+f)^2)


  dldfdf <- -x[1]*((1-pa)^2)/((pa +(1-pa)*f)^2) - x[2]/((1-f)^2) -
   x[3]*(pa^2)/(((1-pa)*(1-f)+f)^2)


  HH <- matrix(c(dldpdp,dldpdf,dldpdf,dldfdf),ncol=2)
  rownames(HH) <- c("pa","f")
  colnames(HH) <- rownames(HH)
  return(HH)
}
```

Capítulo 7
Ejemplos numéricos

En este capítulo propondremos dos ejemplos de datos para compararlos y así poder apreciar las diferencias entre las estimaciones y los métodos utilizados, y poder sacar conclusiones al respecto.

1. Ejemplo 1

Este primer ejemplo, ha sido seleccionado de una base de datos de donantes de sangre de 1000 individuos en Inglaterra, esta base de datos se ha obtenido de un libro de genética de Hedrick [**4**]. Hemos escogido $x = \{298, 489, 213\}$ frecuencias genotípicas absolutas correspondientes a los genotipos MM, MN y NN. Antes de realizar ningún cálculo observamos que la frecuencia de heterocigotos es muy alta, por lo que podemos intuir que el valor de f no deberia ser alto, y que no hay grandes diferencias entre las frecuencias alélicas.

Primero estimaremos los parámetros p y f mediante Máxima verosimilitud, utilizando las fórmulas en el capítulo 4 :

$$\hat{p} = \frac{2n_{AA} + n_{AB}}{2N} = \frac{2 \times 298 + 489}{2 \times 1000} = 0{,}5425$$

$$\hat{f} = \frac{4n_{AA}n_{BB} - n_{AB}^2}{n_A n_B} = \frac{4 \times 298 \times 213 - 489^2}{(2 \times 298 + 489)(2 \times 213 + 489)} = 0{,}0149$$

$$V(\hat{p}) = \frac{p(1-p)(1+f)}{2n} = 0{,}000126$$

$$V(\hat{f}) = \frac{(1-f)^2(1-2f)}{N} + \frac{f(1-f)(2-f)}{2Np(1-p)} = 0{,}001$$

Basándonos en la normalidad asintótica del estimador de máxima verosimilitud [**5**] podemos obtener el IC para p:

$$I.C._{\alpha=0,05}(p) = [\hat{p} - z_{(1-\alpha/2)}\sqrt{V(\hat{p})}, \hat{p} + z_{(1-\alpha/2)}\sqrt{V(\hat{p})} = [0{,}5205\ 0{,}5645]$$

Y análogamente el intervalo de confianza para f:

$$I.C._{\alpha=0,05}(f) = [-0{,}0471\ 0{,}0769]$$

Observamos que el cero se encuentra dentro del intervalo de confianza, lo cual indica que no podemos rechazar el equilibrio de Hardy-Weinberg.

A partir de la función *hwe.gradiente* hemos obtenido el gradiente:

$$\nabla lnL = (\frac{\partial lnL}{\partial p}, \frac{\partial lnL}{\partial f}) = [-1{,}14e-13\ 0]$$

Observamos que el vector gradiente es un vector de ceros, lo cual indica que el punto ($\hat{p} = 0{,}5425$, $\hat{f} = 0{,}0149$) es un punto estacionario de la función de verosimilitud.

Y utilizando *hwe.hessiana*, extraemos la matriz hessiana:

$$H = \begin{pmatrix} \frac{\partial^2 lnL}{\partial p^2} & \frac{\partial^2 lnL}{\partial p \partial f} \\ \frac{\partial^2 lnL}{\partial p \partial f} & \frac{\partial^2 lnL}{\partial f^2} \end{pmatrix} = \begin{pmatrix} -7940{,}077 & -4{,}9464 \\ -4{,}9464 & -999{,}8010 \end{pmatrix}$$

La matriz hessiana es definida negativa, porque todos los autovalores de la matriz son menores que cero(`eigen(H)`), esto significa que los estimadores son máximos.

La matriz de información de Fisher es igual a menos la esperanza de la matriz hessiana, y sabemos mediante la cota de Cramér-Rao es:

$$V(\hat{\theta}) \geq I(\theta)^{-1}$$

La matriz de Información en este caso es:

$$I(\hat{p}, \hat{f}) = \begin{pmatrix} 7940{,}077 & 4{,}9464 \\ 4{,}9464 & 999{,}8010 \end{pmatrix}$$

La inversa de la matriz de información de fisher nos ofrece la varianza mínima de los estimadores, y al tratarse de estimadores de máxima verosimilitud, se cumple la igualdad de esta inecuación.

Las varianzas resultantes de la inversa de la matriz de información son:

$S^2_{\hat{p}} = 0{,}000126$
$S^2_{\hat{f}} = 0{,}00100$

Y realizando la raiz cuadrada obtenemos las desviaciones estándar:

$S_{\hat{p}} = 0{,}0112$
$S_{\hat{f}} = 0{,}0316$

Podemos observar que estos resultados son muy parecidos a los obtenidos por las fórmulas anteriores extraídas de el libro de Weir [**3**].

La correlación entre $\hat{p}$ y $\hat{f}$ es:

$$r(\hat{p}, \hat{f}) = \frac{Cov(\hat{p}, \hat{f})}{S_{\hat{p}} S_{\hat{f}}} = \frac{-6{,}23 \times 10^{-}7}{0{,}01122 \times 0{,}031626} = -0{,}00175$$

La covarianza tiene un valor muy bajo en este caso, por ello la correlación también es despreciable.

El método *nlminb* explicado en el capítulo 6 nos proporciona la siguiente salida:

```
0:     1047.3454: 0.500000  0.00000
1:     1043.7677: 0.572806 0.00471096
2:     1041.5601: 0.541568 0.0706435
3:     1040.3224: 0.549177 -0.00191677
4:     1040.1079: 0.542218 0.000275378
5:     1040.0311: 0.543338 0.00748475
6:     1040.0191: 0.541718 0.00927874
7:     1040.0064: 0.542433 0.0115875
8:     1040.0009: 0.542501 0.0148798
9:     1040.0009: 0.542500 0.0148830
```

Nos ofrece el valor objetivo, y las estimaciones de los parametros para cada iteración.

El output resultante contiene los parámetros estimados para los que se alcanza el óptimo en la función objetivo, el valor de dicha función, la convergencia del método, el número de iteraciones que se han necesitado para llegar al óptimo, la función gradiente y el tipo de convergencia que se ha conseguido.

Vemos como las estimaciones de los parámetros coinciden y también las suposiciones iniciales. Para saber si estas estimaciones son significativas, utilizaremos varios tests de la libreria $HardyWeinberg$ [**6**] en R a continuación:

Test de Chi-Cuadrado para HWE:

$$\chi^2 = \sum_{genotipos} \frac{(observados - esperados)^2}{esperados} = 0{,}2214896$$

Y mediante la orden `HWChisq` en R obtenemos la siguiente salida:

```
Chi-square test for Hardy-Weinberg equilibrium
Chi2 =  0.2214896 p-value =  0.6379073 D =  -3.69375 f =  0.01488253
```

Por lo que podemos decir que no rechazamos la hipotesis nula de que exista HWE, eso significa que el parametro f no será significativo.

Realizamos el mismo test pero corrigiendo por el parámetro de continuidad:

```
Chi-square test with continuity correction for Hardy-Weinberg equilibrium
Chi2 =  0.1789563 p-value =  0.6722717 D =  -3.69375 f =  0.01488253
```

Vemos que esta corrección no afecta al resultado.

También realizamos el Test Exacto en R con la orden `HWExact`:

```
Haldane's Exact test for Hardy-Weinberg equilibrium
using SELOME p-value
sample counts: nMM =  298 nMN =  489 nNN =  213
H0: HWE (D==0), H1: D <> 0
D =  -3.69375 p =  0.6556635
```

De nuevo, no podemos rechazar la hipotesis nula de la existencia de equilibrio de Hardy-Weinberg, por lo que f no es significativamente diferente de 0. Destacamos que los tres tests (IC para $\hat{f}$, chi-cuadrado y exact test) conducen a la misma conclusión.

Ahora presentaremos las curvas de nivel resultantes de la orden *contourplot* de la librería *lattice* [**7**] en la figura 1:

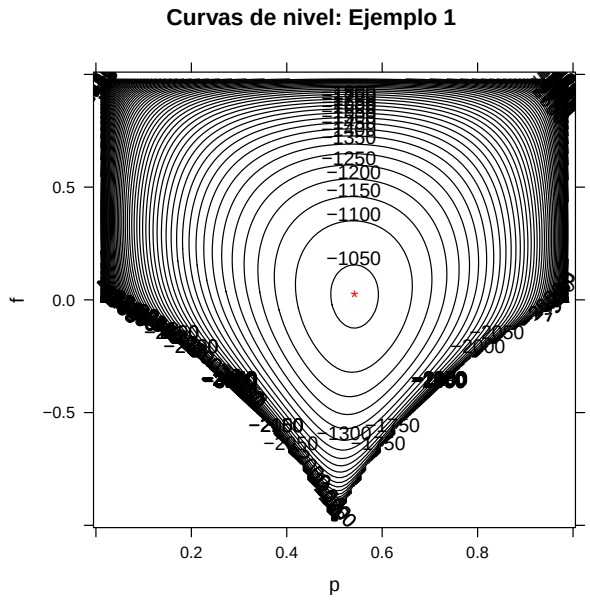

FIGURA 1. Curvas de nivel: Ejemplo 1

Podemos observar que para este caso, el nivel más alto nos indica que la estimación de la frecuencia alélica de M se situaría alrededor de 1/2 y la del coeficiente de endogamia se localicaría muy próxima a 0. La función de verosimilitud aparece como concava con un único punto estacionario interior.

2. Ejemplo 2

Este segundo ejemplo se basa en un SNP, un polimorfismo A/T de un estudio de genotipado con frecuencias genotípicas $x = \{0, 12, 88\}$. Estos datos han sido escogidos con el obgetivo de obtener frecuencia alélica de A pequeña, vemos que la frecuencia genotípica de TT es muy alta, lo que sugiere que, a pesar de que la frecuencia de AA sea 0, quizás estemos en presencia de un f no despreciable.

Procedemos al cálculo de las estimaciones máximas verosimiles obtenidas:

$$\hat{p} = \frac{2n_{AA} + n_{AB}}{2N} = \frac{2 \times 0 + 12}{2 \times 100} = 0{,}06$$

$$\hat{f} = \frac{4n_{AA}n_{BB} - n_{AB}^2}{n_A n_B} = \frac{4 \times 0 \times 88 - 12^2}{(2 \times 0 + 12)(2 \times 88 + 12)} = -0{,}0638$$

Las varianzas de Weir [**3**] y sus desviaciones son:

$$V(\hat{p}) = \frac{p(1-p)(1+f)}{2N} = 0{,}00026$$
$$s.d. = \sqrt{V(\hat{p})} = 0{,}01625$$

$$V(\hat{f}) = \frac{(1-f)^2(1-2f)}{N} + \frac{f(1-f)(2-f)}{2Np(1-p)} = 0{,}00034$$
$$s.d. = \sqrt{V(\hat{f})} = 0{,}01839$$

El intervalo de confianza de p es:

$$I.C._{\alpha=0{,}05}(p) = [\hat{p} - z_{(1-\alpha/2)}\sqrt{V(\hat{p})}, \hat{p} + z_{(1-\alpha/2)}\sqrt{V(\hat{p})} = [0{,}0282\ 0{,}0918]$$

Y para f es:

$$I.C._{\alpha=0{,}05}(f) = [-0{,}0999\ -0{,}0278]$$

Es interesante remarcar en este punto que el intervalo de confianza para f no contiene el 0.

Mediante la función *hwe.gradiente* obtenemos el gradiente:

$$\nabla lnL = (\frac{\partial lnL}{\partial p}, \frac{\partial lnL}{\partial f}) = [-6{,}382979 - 5{,}640000]$$

Observamos que el vector gradiente está lejos de ser un vector de ceros, lo cual indica que, contrariamente al ejemplo anterior, el punto ($\hat{p} = 0{,}06$, $\hat{f} = -0{,}0638$) no es un punto estacionario de la función de verosimilitud.

La matriz hessiana:

$$H(lnL) = \begin{pmatrix} \frac{\partial^2 lnL}{\partial p^2} & \frac{\partial^2 lnL}{\partial p \partial f} \\ \frac{\partial^2 lnL}{\partial p \partial f} & \frac{\partial^2 lnL}{\partial f^2} \end{pmatrix} = \begin{pmatrix} -3560{,}1431 & 100{,}40909 \\ 100{,}40909 & -10{,}96467 \end{pmatrix}$$

La matriz hessiana es definida negativa, porque todos los autovalores de la matriz son menores que cero, esto significa que los estimadores son máximos.

La matriz de Información en este caso es:

$$I(\hat{p}, \hat{f}) = \begin{pmatrix} 3560{,}1431 & -100{,}40909 \\ -100{,}40909 & 10{,}96467 \end{pmatrix}$$

Y la inversa de la matriz de Información de Fisher resultante es:

$$I^{-1}(\hat{p}, \hat{f}) = \begin{pmatrix} 0{,}0003787 & 0{,}0034679 \\ 0{,}0034679 & 0{,}1229593 \end{pmatrix}$$

Las varianzas resultantes de la inversa de la matriz de Información son:

$S^2_{\hat{p}} = 0{,}000379$
$S^2_{\hat{f}} = 0{,}122959$

Y realizando la raiz cuadrada obtenemos las desviaciones estándar:

$S_{\hat{p}} = 0{,}0195$
$S_{\hat{f}} = 0{,}3506$

Notemos, que las medidas de desviación obtenidas a partir de la matriz hessiana son mayores que las obtenidas mediante las fórmulas de Weir, sobre todo para $\hat{f}$, este resultado contradice el supuesto de que la inversa de la matriz Información de Fisher proporciona las varianzas mínimas para los estimadores máximos verosímiles.

La correlación entre $\hat{p}$ y $\hat{f}$ es:

$$r(\hat{p}, \hat{f}) = \frac{Cov(\hat{p}, \hat{f})}{S_{\hat{p}} S_{\hat{f}}} = \frac{0{,}00347}{0{,}01946 \times 0{,}35066} = 0{,}50821$$

La correlación es muy alta, y positiva, lo que significa que existe una relación directa entre $\hat{p}$ y $\hat{f}$.

La salida ofrecida por *nlminb* es:

```
 0:     130.31167: 0.500000  0.00000
 1:     98.367591: 0.402252 0.0211046
 2:     53.267854: 0.204921 0.0536673
 3:     45.648203: 0.156014 0.0564466
 4:     37.491442: 0.0580723 0.0588302
 5:     36.841582: 0.0610680 -0.0390950
 6:     36.821283: 0.0583293 -0.0411786
 7:     36.773026: 0.0626162 -0.0542590
 8:     36.741596: 0.0600441 -0.0552405
 9:     36.719453: 0.0564720 -0.0594306
10:     36.710470: 0.0573226 -0.0601300
11:     36.697828: 0.0585373 -0.0619671
12:     36.697497: 0.0585290 -0.0620229
13:     36.696835: 0.0585128 -0.0621346
14:     36.696808: 0.0585122 -0.0621391
15:     36.696803: 0.0585120 -0.0621400
16:     36.696761: 0.0585111 -0.0621471
17:     36.696760: 0.0585111 -0.0621471
18:     36.696760: 0.0585111 -0.0621472
19:     36.696760: 0.0585111 -0.0621472
20:     36.696760: 0.0585111 -0.0621473
21:     36.696760: 0.0585111 -0.0621473
22:     36.696759: 0.0585110 -0.0621473
23:     36.696759: 0.0585110 -0.0621473
24:     36.696759: 0.0585110 -0.0621473
25:     36.696759: 0.0585110 -0.0621473
26:     36.696759: 0.0585110 -0.0621473
```

Las estimaciones obtenidas mediante las fórmulas del capítulo 4 y las obtenidas a partir de *nlminb* son muy parecidas. El test X^2 para HWE:

```
Chi-square test for Hardy-Weinberg equilibrium
Chi2 =  0.4074242 p-value =  0.5232798 D =  0.36 f =  -0.06382979
```

Este test no es adecuado para este tipo de muestra, ya que la frecuencia genotípica de AA es menor que 5. Pero de todas formas, vemos que no se rechaza la hipotesis nula de que exista equilibrio de Hardy-Weinberg. El mismo test con parámetro de continuidad igual a 0.5:

```
Chi-square test with continuity correction for Hardy-Weinberg equilibrium
Chi2 =  0.05895704 p-value =  0.808152 D =  0.36 f =  -0.06382979
```

Vemos como la introducción del parámetro de continuidad no afecta al resultado. Pero como hemos dicho, no es un test adecuado.

El siguiente test es más adiente:

```
Haldane's Exact test for Hardy-Weinberg equilibrium
using SELOME p-value
sample counts: nAA =  0 nAT =  12 nTT =  88
```

```
H0: HWE (D==0), H1: D <> 0
D =  0.36 p =  1
```

Podemos observar que el p-valor es 1, por lo que no rechazamos la hipotesis nula de que halla HWE y por lo tanto, el coeficiente de endogamia no es significativamente diferente de 0, cosa que contradice el resultado obtenido mediante el intervalo de confianza que, recordemos, no contenia el 0. Lo cual supone una clara contradicción. Ésto significa que la solución asimptótica no es adecuada para esta muestra, ya sea por un tamaño muestral pequeño o bien por el hecho de que estamos trabajando con una frecuencia alélica muy baja. Este aspecto lo estudiaremos en el próximo capítulo.

La figura 2 representa las curvas de nivel:

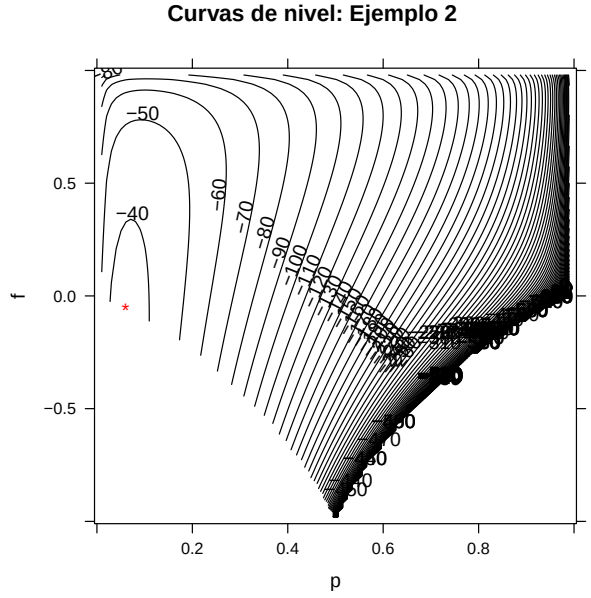

FIGURA 2. Curvas de nivel: Ejemplo 2

Los valores bajos de p son los que se situan dentro de la primera curva, el valor del estimador del coeficiente f se encuentra muy próximo a 0, es exactamente $\frac{-0{,}06}{(1-0{,}06)}$, por lo que coincide con el límite inferior de su rango, también hemos de citar que su valor es negativo, lo que nos sugiere que pueda existir un exceso de heterocigotos.

Capítulo 8
Simulaciones

En este apartado estudiaremos las propiedades del estimador de máxima verosimilitud de f (sesgo, varianza, probabilidad de cobertura de los intervalos de confianza) y su comportamiento frente a diferentes valores que pueda tomar la frecuencia alélica p.

Utilizado la orden `HWData` del paquete $HardyWeinberg$ [**6**], simulamos marcadores genéticos bajo el supuesto de equilibrio de HW, tomando muestras de una distribución multinomial con probabilidades p^2, $2pq$ y q^2.

Antes de generar las simulaciones hemos reproducido $nm = 10000$ réplicas de las muestras escogidas, creamos un vector de longitud nm con los tamaños de las muestras ($n < -rep(1000, nm)$), también un vector con el valor de f (hemos escogido $f = 0$) repetido nm veces, y otro vector con nm veces el valor de p (en este caso $p = 0{,}5$).

```
sim<-HWData(nm, n, f, p, pfixed = FALSE, exactequilibrium = FALSE)
```

Ahora tenemos *sim* que es una lista con Xt que son las frecuencias genotípicas absolutas simuladas, y Xc son las relativas.

Calculamos $\hat{p}$, $\hat{f}$, la varianza de $\hat{f}$ y los intervalos de confianza de $\hat{f}$ para cada una de las muestras simuladas a partir del siguiente bucle:

```
for (i in 1:nm){
  p[i] <- (2*Xt[i,1]+Xt[i,2])/(2*N)

  f[i] <- (4*Xt[i,1]*Xt[i,3]-Xt[i,2]^2)/
  ((2*Xt[i,1]+Xt[i,2])*(2*Xt[i,3]+Xt[i,2]))

  vf[i] <- ((1-f[i])^2)*(1-2*f[i])/N +
  f[i]*(1-f[i])*(2-f[i])/(2*N*p[i]*(1-p[i]))

  liminf[i] <- f[i]-qnorm(1-alpha/2)*sqrt(vf[i])
  limsup[i] <- f[i]+qnorm(1-alpha/2)*sqrt(vf[i])
}
```

Presentamos a continuación la distribución de las frecuencias relativas alélicas estimadas a partir de las muestras simuladas en un histograma:

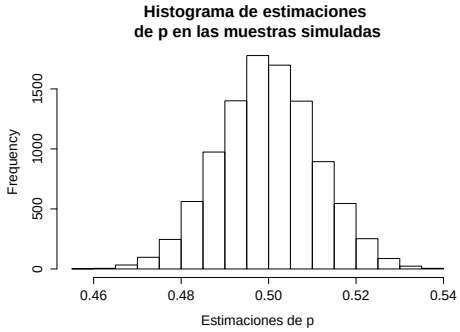

FIGURA 1. Histograma de estimaciones de p

Vemos como la distribución se asemeja a una normal de media 0.5, lo cual es lógico ya que hemos simulado un número razonable de muestras como para asegurar que asimptóticamente sigue una distribución normal.

En la figura 2 presentamos el histograma distribución de las f generadas:

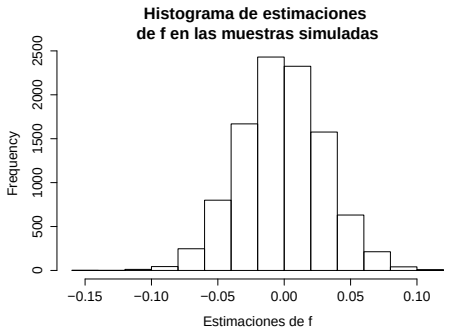

FIGURA 2. Histograma de estimaciones de f

En este gráfico ocurre lo mismo que con la distribución de estimaciones de p, asimptóticamente sigue una distribución normal solo que con media 0.

En la figura 3 presentamos el histograma de las varianzas de Weir de las estimaciones de p. Podremos ver que las varianzas de Weir de las $\hat{p}$ representan una normal con media 0.0001248, el valor teórico de la varianza de $\hat{p}$ que es 0.000125 por lo que son muy parecidos.

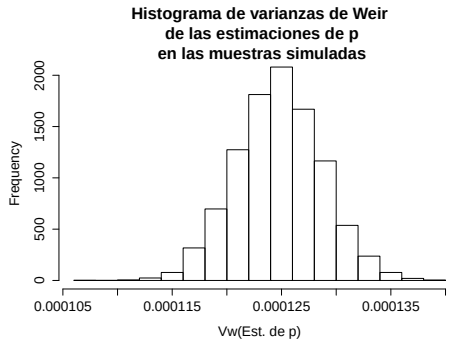

FIGURA 3. Histograma de varianzas de Weir de las estimaciones de p

En la siguiente figura presentamos el histograma de las varianzas de Weir para las estimaciones de f:

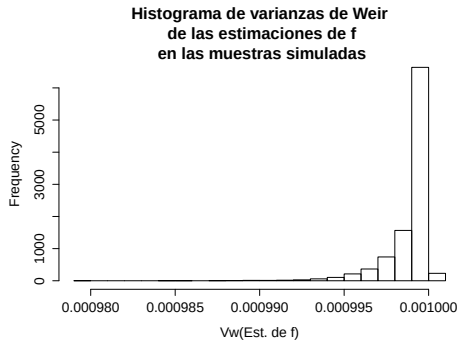

FIGURA 4. Histograma de varianzas de Weir de las estimaciones de f

y vemos como el valor medio de las varianzas de Weir obtenidas (0.000999) es ligeramente menor al valor de la varianza poblacional de $\hat{f}$ (0.001).

Hallamos $n0f$ que es el número de muestras en las que $f = 0$ queda comprendido en el intervalo de confianza.

El ratio de este número entre el total de muestras simuladas, $\frac{n0f}{nm}$ da lugar a la llamada probabilidad de cobertura, que para los valores introducidos ha resultado igual a 0.9497, lo que significa que el $94,97\,\%$ de las muestras tienen un coeficiente de endogamia no significativo. Son menos del $95\,\%$ esperado aunque muy cercano, esto puede ser debido a que la media de varianzas calculadas para las estimaciones de f de las muestras simuladas es algo menor que la varianza poblacional.

Si escogieramos un valor para la frecuencia alélica de 0.1, y y 0 para el factor de endogamia, los histogramas resultantes serían los siguientes:

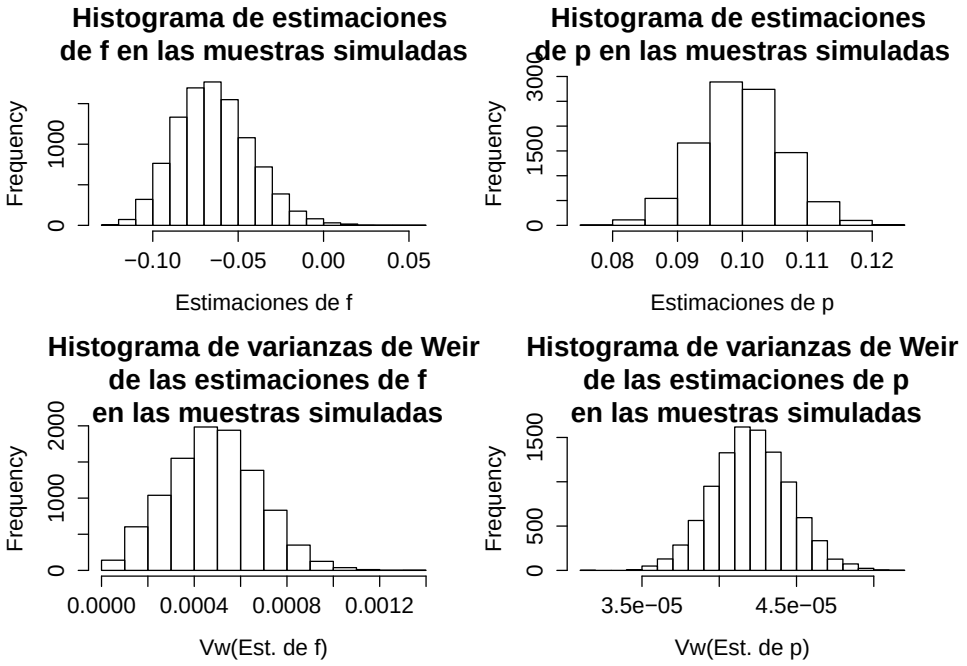

FIGURA 5. Histogramas ($p = 0{,}1$, $f = 0$, $n = 1000$)

En el primer histograma de las estimaciones de f, podemos ver como contrariamente al caso anterior, su media se ha desplazado del 0 hacia cantidades negativas a pesar de que escogimos el valor de 0 para las simulaciones, recordemos que esto implicaba que existía un exceso de heterocigotos, tiene sentido, ya que el número de homozigotos del alelo con frecuencia mínima es muy pequeño. Las estimaciones de p se distribuyen normalmente con media 0.1 (tal y como escogimos en la simulación).

El valor de la varianza poblacional de p en este caso es $4{,}5e-05$ que es mayor que la media de las varianzas de $\hat{p}$ que tiene el valor de 4.2e-05 y comparando la media de las varianzas de $\hat{f}$ (0.00049) con la varianza poblacional de f (0.001), la poblacional es nada menos que el doble que la media de las varianzas de $\hat{f}$.

Calculando la probabilidad de cobertura obtenemos un $23,32\,\%$ este valor es muy bajo, y es debido a la diferencia enorme que existe entre la varianza poblacional y la media de las varianzas de $\hat{f}$.

Si reducimos el tamaño muestral a 100 individuos:

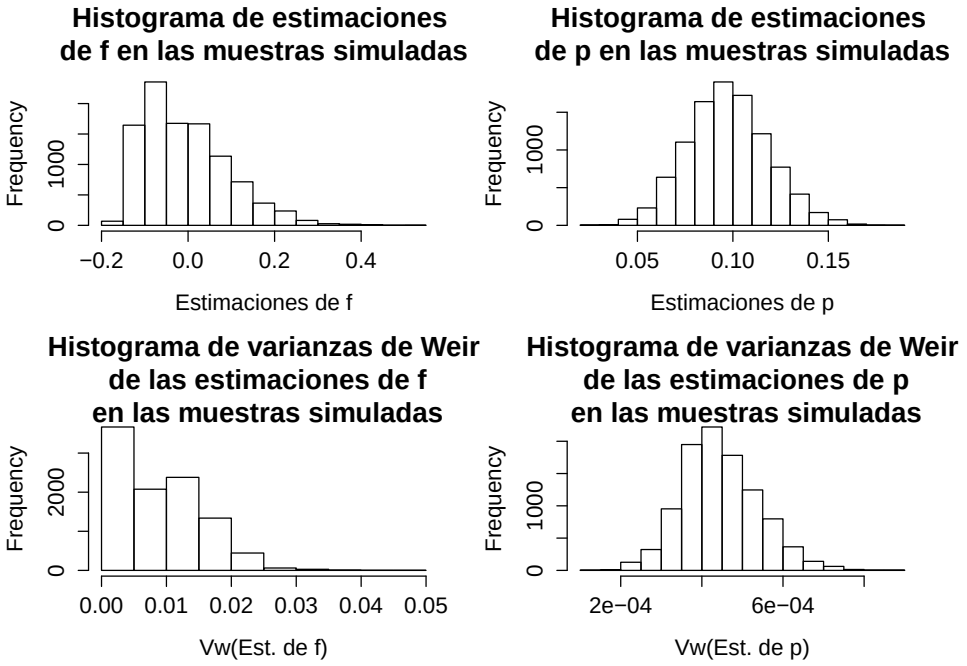

FIGURA 6. Histogramas ($p = 0{,}1$, $f = 0$, $n = 100$)

Vemos que las estimaciones de f no se concentran tanto en los valores negativos como en el histograma anterior(con $n = 1000$), aunque su media sea negativa, se situa muy cerca de 0. La varianza poblacional de f en este caso es de 0.01 que es mayor que la media de las varianzas de $\hat{f}$ que nos ha dado 0.00845. En cuanto a la frecuencia alélica, las diferencias entre estos histogramas y los obtenidos con $n = 1000$ se basan en que las varianzas son mayores. La varianza poblacional de p es 0.00045 aunque un poco más alta, es muy parecida a la media de las varianzas de $\hat{p}$ obtenida con valor 0.000445.

Respecto a la probabilidad de cobertura, esta vez un 62,6 % de los intervalos de confianza contienen el cero, esto es debido a que la media de las varianzas de $\hat{f}$ aunque menor que la varianza poblacional, es mayor que la obtenida en el anterior ejemplo con n=1000.

Ahora volvemos a cambiar el valor de la frecuencia alélica a 0.5 y mantenemos el tamaño muestral de 100. Obtenemos los siguientes gráficos:

Observamos que los histogramas referentes a las estimaciones son muy parecidos a los obtenidos con tamaños muestrales de 1000 individuos. La varianza poblacional de f es 0.01, es ligeramente mayor que la media de varianzas de las estimaciones de f (0.00989), lo mismo pasa para la frecuencia alélica, el valor de la varianza poblacional es 0.00125 y la media de las varianzas de las estimaciones de p es 0.00124. La probabilidad de cobertura es de $94,09\,\%$, muy parecida a la obtenida en el caso de n=1000.

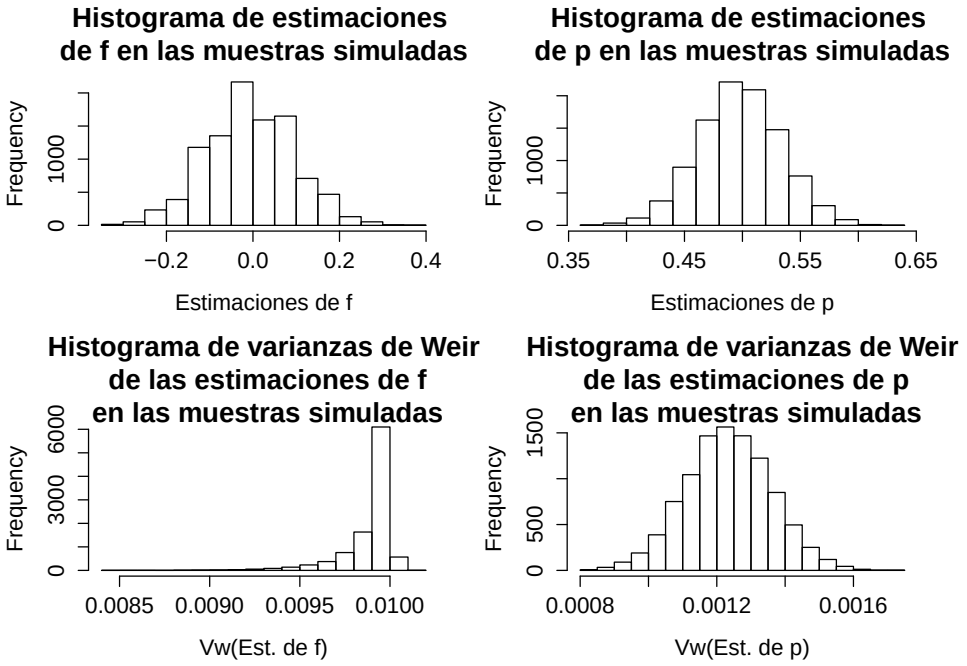

FIGURA 7. Histogramas ($p = 0{,}5$, $f = 0$, $n = 100$)

A continuación repetiremos el procedimiento para diferentes valores de frecuencias alélicas mínimas p_m. En el apéndice se encuentran las funciones utilizadas para este apartado. Tenemos la función *simular*, que se basa en un bucle que crea nm simulaciones para cada valor de p_m, y la función *sesgo.cober* que te crea un output con los sesgos de las estimaciones de p y f, las medias de las varianzas calculadas a partir de las fórmulas de Weir, las varianzas muestrales, y las probabilidades de cobertura.

Los resultados obtenidos los resumimos en la siguiente tabla:

p_m	Sesgo de $\hat{p}$	$\bar{V}_{Weir}(\hat{p})$	σ^2_p	Sesgo de $\hat{f}$	$\bar{V}_{Weir}(\hat{f})$	σ^2_f	P(cob.)
0.05	-0.000020	0.000024	0.000024	-0.000355	0.000967	0,001	0.8905
0.10	0.000035	0.000045	0.000045	-0.000234	0.000986	0,001	0.9272
0.15	0.000068	0.000064	0.000064	-0.000621	0.000991	0,001	0.9429
0.20	0.000109	0.000080	0.000080	-0.000548	0.000994	0,001	0.9473
0.25	-0.000042	0.000094	0.000094	-0.000688	0.000996	0,001	0.9500
0.30	0.000099	0.000105	0.000105	-0.000409	0.000998	0,001	0.9502
0.35	0.000014	0.000114	0.000114	-0.000100	0.000998	0,001	0.9480
0.40	0.000010	0.000120	0.000120	-0.000210	0.000999	0,001	0.9508
0.45	-0.000044	0.000124	0.000124	-0.000648	0.000999	0,001	0.9444
0.50	-0.000081	0.000125	0.000125	-0.000168	0.000999	0,001	0.9474

TABLA 1. Sesgos y Cobertura: n=1000

En esta tabla vemos que los sesgos son muy pequeños, todos son menores de 0.001, y son independientes de la magnitud del estimador. Si nos fijamos en las medias de las varianzas de las estimaciones, vemos que siempre son ligeremente menores que las poblacionales.

Podemos observar claramente que cuanto menor sea la frecuencia alélica, menor es también la probabilidad de cobertura, lo cual nos hace pensar que para frecuencias alélicas pequeñas, sería erróneo tomar soluciones asintóticas para los intervalos de f y otras pruebas de hipotesis, sino que la prueba exacta es la más apropiada para testar el valor de $\hat{f}$. Por lo que podemos afirmar que, para el segundo ejemplo, el resultado de no rechazo de la existencia de equilibrio de Hardy-Weinberg y por lo tanto la no significación de f es el resultado correcto, y no los intervalos de confianza.

En el siguiente gráfico ofrecemos un plot de la probabilidad de cobertura en función de la menor frecuencia alélica (p_m).

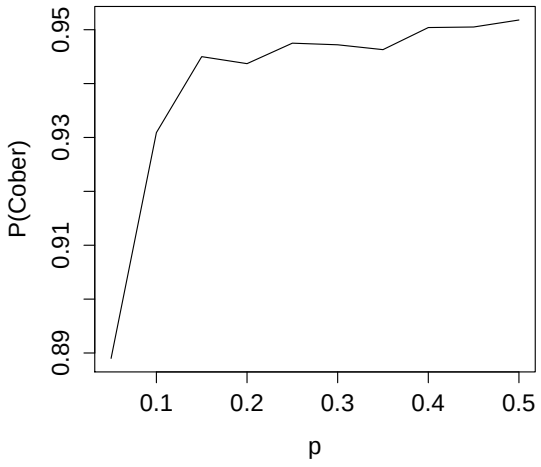

FIGURA 8. Plot p_m vs P(Cobertura)

Por curiosidad y para verlo más claramente, volvemos a realizar el mismo plot de probabilidades de cobertura en función de menor frecuencia alélica pero cambiando el rango de p_m, escogemos desde 0.01 hasta 0.5 en particiones de 0.01 y nos resulta el siguiente gráfico:

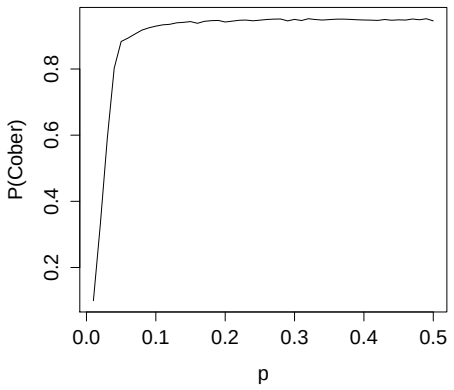

FIGURA 9. Plot p_m vs P(Cobertura)

Podemos observar que sólo a partir del valor de la menor frecuencia alélica de aproximadamente 0.15 se puede considerar que los intervalos de confianza son fiables, ya que la probabilidad de cobertura es menor cuanto menor el valor de p_m. Para más detalle, la tabla de sesgos y probabilidades de cobertura de esta última simulación se ha incorporado en el apéndice.

Ahora repetiremos el proceso para tamaños muestrales de 100, como en el ejemplo 2, presentamos la tabla resumen de sesgos, varianzas y probabilidades de cobertura:

p_m	Sesgo de $\hat{p}$	$\bar{V}_{Weir}(\hat{p})$	σ_p^2	Sesgo de $\hat{f}$	$\bar{V}_{Weir}(\hat{f})$	σ_f^2	P(cob.)
0.05	-0.000151	0.000234	0.000238	-0.004831	0.006550	0.01	0.2229
0.10	0.000169	0.000446	0.000450	-0.004969	0.008475	0.01	0.6314
0.15	-0.000185	0.000630	0.000638	-0.005149	0.009090	0.01	0.8672
0.20	0.000065	0.000792	0.000800	-0.005475	0.009424	0.01	0.9133
0.25	-0.000130	0.000929	0.000938	-0.004141	0.009634	0.01	0.9305
0.30	-0.000228	0.001038	0.001050	-0.005924	0.009726	0.01	0.9348
0.35	-0.000557	0.001126	0.001137	-0.004085	0.009816	0.01	0.9398
0.40	0.000176	0.001189	0.001200	-0.004476	0.009859	0.01	0.9372
0.45	-0.000392	0.001226	0.001238	-0.004508	0.009885	0.01	0.9441
0.50	0.000245	0.001238	0.001250	-0.004498	0.009895	0.01	0.9439

TABLA 2. Sesgos y Cobertura: n=100

Observamos que los sesgos son lógicamente mucho mayores para tamaños muestrales de $n = 100$ que con $n = 1000$, también podemos ver que las varianzas de $\hat{f}$ obtenidas a partir de la fórmula de Weir siguen siendo menores que las muestrales. Y respecto a la relación entre la frecuencia alélica mínima y la probabilidad de cobertura apunta en la misma dirección que en el caso anterior, pero más pronunciadamente.

En la siguiente figura presentamos las probabilidades de cobertura en función de los diferentes valores de frecuencia alélica mínima escogidos, para diferentes tamaños muestrales(= 100, $n = 500$, $n = 1000$):

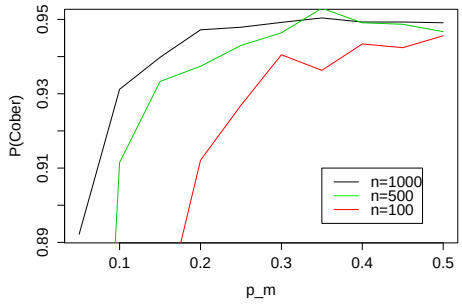

FIGURA 10. Plot p_m vs P(Cobertura) según tamaño muestral

Vemos que el decremento de la probabilidad de cobertura con valores pequeños de frecuencias alélicas mínimas es mucho más notable cuanto menor sean los tamaños muestrales.

Por último en la siguiente figura representamos las varianzas obtenidas a partir de las fórmulas de Weir respecto frecuencias mínimas, para los diferentes tamaños muestrales:

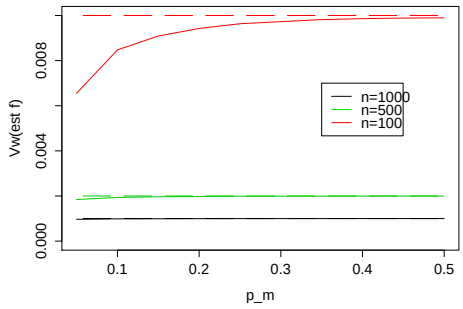

FIGURA 11. Plot p_m vs $V_w(estf)$ según tamaño muestral

Podemos observar que, como es de esperar, las varianzas relativas a tamaños muestrales pequeños son mucho mayores que las obtenidas a partir de muestras de tamaño considerable. Las líneas intermitentes representan las varianzas poblacionales de f, que por lo que podemos ver son mucho mayores que las medias de las varianzas de las estimaciones de f para frecuencias alélicas muy bajas, sin embargo para frecuencias alélicas mayores las varianzas se igualan.

Capítulo 9
Conclusión

Hemos obtenido expresiones explícitas para f y p, que hemos comparado con estimaciones obtenidas mediante métodos numéricos y coinciden. Y se han aplicado los estimadores a datos empíricos.

Hemos visto que las medidas adecuadas para la dispersión de f dependen notablemente del valor de la frecuencia alélica.

Hemos demostrado, mediante simulación, que en el caso de que la frecuencia mínima alélica sea pequeña, los valores de f tienden a ser menores de 0, y las medias de las varianzas de las estimaciones de f en las muestras simuladas, resultan menores que las poblacionales, esto provoca un decremento en la probabilidad de cobertura.

También hemos podido observar que el tamaño de las muestras interfiere totalmente en las conclusiones anteriores. Hemos observado que cuanto menor sea el tamaño muestral, más aumenta la diferencia entre las varianzas de las estimaciones y la varianza poblacional. Consecuentemente, con tamaños muestrales pequeños aun menor es la probabilidad de cobertura en el caso de tener frecuencias mínimas alélicas bajas. Por lo que se refuerzan aún más las conclusiones anteriores.

Podemos decir entonces, que los intervalos de confianza basados en la distribución asintótica del parámetro no son representativos de la realidad, en presencia de una menor frecuencia alélica baja, y sobre todo con tamaños muestrales pequeños, ya que son demasiado estrechos y fácilmente se obtiene un coeficiente de endogamia significativo. En esos casos, es aconsejable realizar la *Prueba Exacta* para determinar la significación del coeficiente de endogamia.

Capítulo 10
Apéndice

```
#install.packages("HardyWeinberg")
#install.packages("lattice")
library(HardyWeinberg)
library(lattice)

x <- c(298,489,213)
names(x) <- c("MM","MN","NN")

x <- c(0,12,88)
names(x) <- c("AA","AT","TT")


N<-sum(x)
p <- (2*x[1]+x[2])/(2*N) #E(p)
f <- (4*x[1]*x[3]-x[2]^2)/((2*x[1]+x[2])*(2*x[3]+x[2])) #E(f)
vp<-p*(1-p)*(1+f)/(2*N) #var(p)
vf<-((1-f)^2)*(1-2*f)/N + f*(1-f)*(2-f)/(2*N*p*(1-p)) #var(f)
round(f-qt(alpha/2,N-1)*sqrt(vf),4) #liminf
round(f+qt(alpha/2,N-1)*sqrt(vf),4)  #limsup

logver <- function(par){
  p <- par[1]
  f <- par[2]

  pi<-rep(NA,3)

  pi[1] <- p^2+p*(1-p)*f
  pi[2] <- 2*p*(1-p)*(1-f)
  pi[3] <- (1-p)^2+p*(1-p)*f

  vo <- sum(x*log(pi))
  return(-vo)
}
```

```
nlminb(rep(1/2,2), logver, lower = 0, upper = 1)

results <- HWChisq(x,cc=0,verbose=TRUE)#chi square
results <- HWExact(x,pvaluetype="selome",verbose=TRUE)
#exact test


#z matriz con valores de la funcion de verosimilitud
#segun los de $p$ y $f$
#Preparamos los datos para graficos de curvas de nivel

x1<-p   #estimaciones de p y f
x2<-f

p <- seq(0, 1, length.out = 100)
f <- seq(-1, 1, length.out = 100)
grid <- expand.grid(p=p, f=f)

#attach(grid)
for(i in 1:nrow(grid)){
  paa <- grid$p[i]^2+grid$p[i]*(1-grid$p[i])*grid$f[i]
  pab <- 2*grid$p[i]*(1-grid$p[i])*(1-grid$f[i])
  pbb <- (1-grid$p[i])^2+grid$p[i]*(1-grid$p[i])*grid$f[i]

  grid$z[i] <- x[1]*log(paa)+x[2]*log(pab)+x[3]*log(pbb)
}


contourplot(z ~ p * f, data = grid,
            add=T,
            cuts = 100, region = F,
            xlab = "p",
            ylab = "f",
            main = "Curvas de nivel: Ejemplo 1")

trellis.focus("panel",1, 1,highlight=F)
ltext(x1,x2,"*",col=2,font=33)
trellis.unfocus()
simular <- function(part,pf,nm,ns){
  ##part numero de particiones en el rango del parametro p
  # pf parametro f escogido
  # nm  numero de simulaciones
  # ns tamaÃ±o de las muestras
  ##############################

  pp <- seq(0.05, 0.5, length.out =part)
  n <- rep(ns, nm)
  f <- rep(pf, nm)
```

```
  sim<- vector('list', part)
  names(sim) <- 1:part  #paste0('item:', seq_along(mybiglist)

  for (i in 1:part){
  p <- rep(pp[i],nm)
  temp<-HWData(nm, n, f, p, pfixed = FALSE, exactequilibrium = FALSE)
  sim[[i]] <- temp
}
return(sim)
}
n0f<-NA
sesgo.cober<-function(alpha,sim){

  restot <- data.frame()

  for (obj in 1:length(sim)){

    pp <- seq(0.05, 0.5, length.out=length(sim))
    p<-f<-vf<-vp<-liminf<-limsup<-matrix(NA,nm,)
    Xt<-sim[[obj]]$Xt

    for (i in 1:nm){
      p[i] <- (2*Xt[i,1]+Xt[i,2])/(2*N) #E(p)
      f[i] <- (4*Xt[i,1]*Xt[i,3]-Xt[i,2]^2)/((2*Xt[i,1]+Xt[i,2])*(2*Xt[i,3]+Xt[i,2]))
      vf[i] <- ((1-f[i])^2)*(1-2*f[i])/N + f[i]*(1-f[i])*(2-f[i])/(2*N*p[i]*(1-p[i]))
      vp[i] <- p[i]*(1-p[i])*(1+f[i])/(2*N)
      liminf[i] <- f[i]-qnorm(1-alpha/2)*sqrt(vf[i])
      limsup[i] <- f[i]+qnorm(1-alpha/2)*sqrt(vf[i])
      vpm<-var(p)
      vfm<-var(f)
    }
    vpp<-pp[obj]*(1-pp[obj])*(1+pf)/(2*N)
    vfp<-((1-pf)^2)*(1-2*pf)/N + pf*(1-pf)*(2-pf)/(2*N*pp[obj]*(1-pp[obj]))
    n0f<-sum(liminf<0 & limsup>0)
    res<-c(pp[obj],mean(p)-pp[obj], mean(vp),vpp, mean(f), mean(vf),vfp, n0f/nm)
    restot<-rbind(restot,res)
  }
  colnames(restot)<-c("Parametro p","sesgo de p", "mean(Var_weir (p))", "var_pob(p)",
  "sesgo de f", "mean(Var_weir (f))", "var_pob(f)", "prob. de cobertura")
  return(restot)
}
nm <- 10000
ns <- 1000
pf <- 0
part<-50

sim<-simular(part,pf,nm,ns)
```

```
(res<-sesgo.cober(alpha,sim))
plot(t(res[1]),t(res[4]),type="l",xlab="p",ylab="P(Cober)")

colnames(x) <- c("X1","X2")
rownames(x) <- c("MM","MN","NN")
HWTernaryPlot(t(x),1000,region=1,hwcurve=TRUE,mafbounds=TRUE,
vertex.cex=1, addmarkers = TRUE)

hwe.gradiente <- function(p,x) {
  pa <- p[1]
  f <- p[2]
  dldpa <- x[1]*(2*pa*(1-f)+f)/(pa^2*(1-f)+pa*f) +
    x[2]*(1-2*pa)/(pa*(1-pa)) -
    x[3]*(2*(1-pa)*(1-f)+f)/(((1-pa)^2)*(1-f)+(1-pa)*f)

  dldf <- x[1]*(1-pa)/(pa*(1-f)+f) - x[2]/(1-f) +
    x[3]*pa/((1-pa)*(1-f) + f)
  return(c(dldpa,dldf))
}

hwe.hessiana <- function(ps,x) {
   pa <- ps[1]
   f <- ps[2]

  dldpdp <-
    x[1]*(2*pa*(pa+(f/(1-f)))-((2*pa+f/(1-f))^2))/
      ((pa^2)*((pa+f/(1-f))^2)) -
    x[2]*(1-2*pa*(1-pa))/((pa^2)*((1-pa)^2)) -
    x[3]*((2*(1-pa)+f/(1-f))^2-2*(((1-pa)^2)+f/(1-f)))/
    (((1-pa)^2)*((1-pa+f/(1-f))^2))
    # alpha= f/(1-f)

  dldpdf <- -x[1]/((pa*(1-f)+f)^2) + x[3]/(((1-pa)*(1-f)+f)^2)


  dldfdf <- -x[1]*((1-pa)^2)/((pa*(1-f)+f)^2) - x[2]/((1-f)^2) -
  x[3]*(pa^2)/(((1-pa)*(1-f)+f)^2)

  HH <- matrix(c(dldpdp,dldpdf,dldpdf,dldfdf),ncol=2)
  rownames(HH) <- c("pa","f")
  colnames(HH) <- rownames(HH)
  return(HH)
}
```

p_m	Sesgo de $\hat{p}$	$\bar{V}_{Weir}(\hat{p})$	σ_p^2	Sesgo de $\hat{f}$	$\bar{V}_{Weir}(\hat{f})$	σ_f^2	P(cob.)
0.01	-0.000032	0.000005	0.000005	-0.000555	0.000802	0,001	0.0938
0.02	-0.000026	0.000010	0.000010	-0.000267	0.000913	0,001	0.3342
0.03	0.000038	0.000015	0.000015	-0.000518	0.000939	0,001	0.5942
0.04	-0.000026	0.000019	0.000019	-0.000187	0.000963	0,001	0.7988
0.05	0.000006	0.000024	0.000024	-0.000431	0.000967	0,001	0.8931
0.06	0.000034	0.000028	0.000028	-0.000518	0.000971	0,001	0.8864
0.07	-0.000034	0.000033	0.000033	-0.000115	0.000980	0,001	0.9105
0.08	-0.000062	0.000037	0.000037	-0.000472	0.000980	0,001	0.9141
0.09	0.000024	0.000041	0.000041	-0.000694	0.000981	0,001	0.9251
0.10	-0.000009	0.000045	0.000045	-0.000568	0.000984	0,001	0.9322
0.11	-0.000072	0.000049	0.000049	-0.000853	0.000984	0,001	0.9327
0.12	0.000056	0.000053	0.000053	-0.000712	0.000987	0,001	0.9351
0.13	0.000067	0.000057	0.000057	-0.000176	0.000991	0,001	0.9419
0.14	-0.000047	0.000060	0.000060	-0.000172	0.000992	0,001	0.9413
0.15	0.000076	0.000064	0.000064	-0.000537	0.000991	0,001	0.9445
0.16	-0.000061	0.000067	0.000067	-0.000709	0.000992	0,001	0.9424
0.17	-0.000144	0.000070	0.000071	-0.000530	0.000993	0,001	0.9435
0.18	0.000122	0.000074	0.000074	-0.000241	0.000994	0,001	0.9474
0.19	0.000065	0.000077	0.000077	-0.000690	0.000994	0,001	0.9462
0.20	-0.000117	0.000080	0.000080	0.000061	0.000996	0,001	0.9472
0.21	0.000076	0.000083	0.000083	-0.001200	0.000994	0,001	0.9414
0.22	0.000140	0.000086	0.000086	-0.000364	0.000995	0,001	0.9470
0.23	0.000003	0.000088	0.000089	-0.000303	0.000996	0,001	0.9454
0.24	0.000034	0.000091	0.000091	-0.000418	0.000996	0,001	0.9491
0.25	-0.000190	0.000094	0.000094	-0.000514	0.000996	0,001	0.9507
0.26	0.000217	0.000096	0.000096	0.000028	0.000997	0,001	0.9501
0.27	-0.000002	0.000098	0.000099	-0.000503	0.000997	0,001	0.9460
0.28	0.000066	0.000101	0.000101	-0.000685	0.000997	0,001	0.9497
0.29	-0.000070	0.000103	0.000103	-0.001179	0.000997	0,001	0.9503
0.30	-0.000086	0.000105	0.000105	-0.000629	0.000997	0,001	0.9499
0.31	0.000070	0.000107	0.000107	-0.000873	0.000997	0,001	0.9498
0.32	0.000077	0.000109	0.000109	-0.000340	0.000998	0,001	0.9469
0.33	0.000143	0.000110	0.000111	-0.000885	0.000998	0,001	0.9505
0.34	-0.000043	0.000112	0.000112	-0.000371	0.000998	0,001	0.9493
0.35	-0.000009	0.000114	0.000114	-0.000880	0.000998	0,001	0.9499
0.36	0.000094	0.000115	0.000115	0.000296	0.000999	0,001	0.9484
0.37	-0.000033	0.000116	0.000117	-0.000613	0.000998	0,001	0.9475
0.38	-0.000043	0.000118	0.000118	0.000164	0.000999	0,001	0.9519
0.39	0.000007	0.000119	0.000119	-0.000562	0.000999	0,001	0.9488
0.40	0.000018	0.000120	0.000120	-0.000397	0.000999	0,001	0.9472
0.41	0.000075	0.000121	0.000121	-0.000922	0.000999	0,001	0.9465
0.42	0.000035	0.000122	0.000122	-0.000312	0.000999	0,001	0.9527
0.43	-0.000043	0.000122	0.000123	-0.000469	0.000999	0,001	0.9493
0.44	-0.000038	0.000123	0.000123	-0.000582	0.000999	0,001	0.9578
0.45	-0.000160	0.000124	0.000124	-0.000540	0.000999	0,001	0.9463
0.46	-0.000108	0.000124	0.000124	-0.000236	0.000999	0,001	0.9478
0.47	-0.000069	0.000124	0.000125	-0.000841	0.000999	0,001	0.9505
0.48	0.000147	0.000125	0.000125	-0.000546	0.000999	0,001	0.9474
0.49	-0.000064	0.000125	0.000125	-0.000486	0.000999	0,001	0.9497
0.50	-0.000112	0.000125	0.000125	-0.000550	0.000999	0,001	0.9479

TABLA 1. Sesgos y Cobertura: n=1000

Bibliografía

[1] N.M. Laird C. Lange, *The fundamentals of modern statistical genetics* (2011)
[2] A.S. Foulkes, *Applied Statistical genetics with R* (2009)
[3] B.S. Weir, *Genetic Data Analisis II* (1996)
[4] P.W. Hedrick, *Genetics of Populations* (2005)
[5] G. Casella R.L. Berger, *Statistical Inference* (2002)
[6] J. Graffelman, *Graphical tests for Hardy-Weinberg equilibrium* (2013)
[7] D. Sarkar, *Lattice Graphics* (2014)