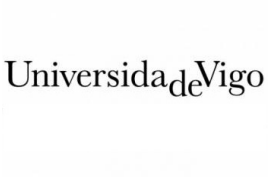

Estimación de la Potencia Estadística en Estudios de Asociación del Genoma Mitocondrial

Proyecto Fin de Máster

Máster en Técnicas Estadísticas

Jacobo José Pardo Seco

Wenceslao González Manteiga Antonio Salas Ellacuriaga

Universidade de Santiago de Compostela

Curso 2010-2011

Autorización de entrega

D. Wenceslao González Manteiga y D. Antonio Salas Ellacuriaga

Certifican

Que el proyecto titulado "**Estimación de la Potencia Estadística en Estudios de Asociación del Genoma Mitocondrial**" ha sido realizado por D. Jacobo José Pardo Seco, con D.N.I. 44819709-T, bajo la dirección de D. Wencesalo González Manteiga y D. Antonio Salas Ellacuriaga.

Esta memoria constituye la documentación que, con nuestra autorización, entrega dicho alumno como Proyecto Fin de Máster.

Firmado

D. Wenceslo González Manteiga

D. Antonio Salas Ellacuriaga

Santiago de Compostela, a 1 de Julio de 2011

Índice general

Capítulo 1

Introducción

Las mitocondrias son orgánulos celulares que se encuentran en el citoplasma de la célula, fuera del núcleo celular, y que están presentes en la mayoría de las células eucariotas, cuyo material genético está encerrado en el núcleo.

Las mitocondrias son de vital importancia, ya que producen la mayor parte del suministro de adenosín trifosfato (ATP) que se utiliza como fuente de energía química, y además están implicadas en otros procesos, como la diferenciación celular, muerte celular programada, el crecimiento celular, etc. La mitocondria está involucrada, directa o indirectamente, en todos los procesos fisicoquímicos que requieren el uso de energía para su ejecución.

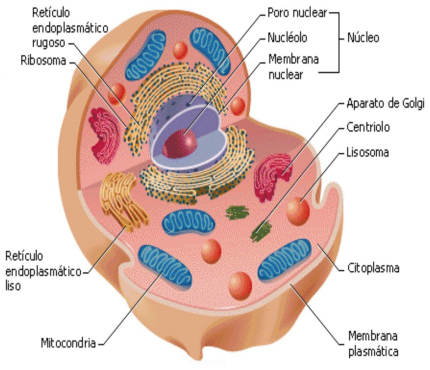

Figura 1.1: Estructura de una célula animal (extraído de http://lacelula2010.galeon.com).

El número de mitocondrias por célula varía ampliamente según el tipo de organismo o tejido, pudiendo contener desde unas pocas docenas a unos cuantos miles.

Estos orgánulos se componen de membrana mitocondrial externa, espacio intermembranoso, membrana mitocondrial interna, crestas y matriz mitocondrial. En cada una de estas partes se llevan a cabo funciones especializadas.

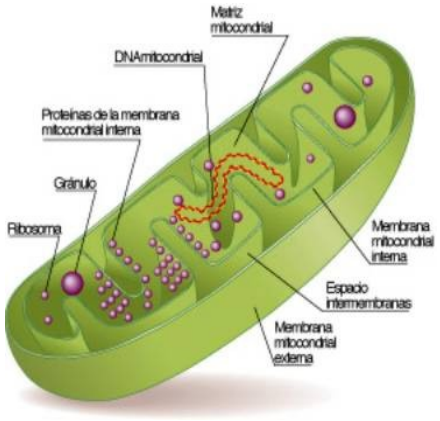

Figura 1.2: Estructura de una mitocondria (extraído de http://fondosdibujosanimados.com.es).

Aunque la mayor parte del ADN de la célula está en el núcleo celular, la mitocondria tiene su propio genoma, el ADN mitocondrial (ADNmt). Determinadas enfermedades u otras condiciones fenotípicas están relacionadas con las mitocondrias, y más concretamente con el ADNmt; sobre todo afectando a tejidos con mayor dependencia del metabolismo mitocondrial. A continución se muestran patologías que están relacionadas con el ADNmt:

- Alzheimer.
- Parkinson.
- Longevidad (más de 90 años).
- Motilidad espermática.
- Esclerosis lateral amiotrófica.
- Ceguera.
- Migrañas.
- Convulsiones.
- Retraso mental.

- Baja estatura.

Por ello, será de interés estudiar la relación entre determinadas variantes genéticas presentes en el ADNmt y ciertas enfermedades.

En un primer lugar se mostrarán las diferencias existentes entre el ADN nuclear y el ADNmt:

- El ADNmt se encuentra en las mitocondrias, fuera del núcleo.
- Se transmite sólo por vía materna, mientras que en el ADN nuclear cada progenitor aporta el 50 % del código genético. Esto es debido a que cuando el espermatozoide fecunda al óvulo, su cola y citoplasma (donde están las mitocondrias) se quedan fuera, permaneciendo dentro del óvulo sólo su núcleo con el ADN nuclear.
- La tasa de mutación promedio del ADNmt es 10 veces mayor que la del ADN nuclear, ya que el ADNmt está expuesto al daño oxidativo por las reacciones que se producen en la mitocondria, además del hecho de que el ADN nuclear está mejor protegido y de que los mecanismos de reparación de daños del ADN son poco eficientes en las mitocondrias.
- El ADNmt posee una estructura bicatenaria circular y cerrada, formada aproximadamente por 16569 bases nitrogenadas, mientras que el ADN nuclear tiene una estrucutura de doble hélice superenrollada formada por cerca de 25000 bases.

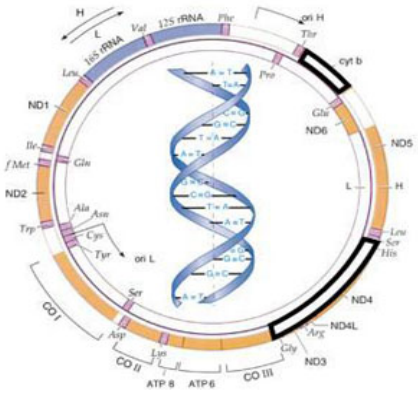

Figura 1.3: Estructuras de ADNmt y ADN nuclear (extraído de http://geezees.com)

1.1. Naturaleza de los datos

La variación genética en el ADNmt se origina a través de mutaciones que se acumulan sobre otras preexistentes. Ésta es la única fuente de variación dado que el ADNmt se hereda por vía materna y no existe recombinación (tal y como sucede con el ADN nuclear).

La forma en la que se genera la variabilidad junto con los modelos demográficos que distribuyen dicha variabilidad en las poblaciones y en el espacio, implica que determinadas secuencias son más prevalentes en determinados lugares o poblaciones.

Se define haplogrupo como el conjunto de secuencias mitocondriales que son filogenéticamente (evolutivamente) cercanas; desde el punto de vista de la variabilidad, esto significa que dichas secuencias comparten un conjunto de variantes ancestrales comunes. Cada haplogrupo está formado por lo tanto por un conjunto de secuencias mitocondriales muy próximas, denominadas haplotipos, término que hace mención al hecho de que todas las variantes genéticas de una misma secuencia serán trasmitidas como un bloque a la siguiente generación.

Los haplogrupos reciben nombres concretos que generalmente consisten en una o varias letras. Cuando uno de estos haplogrupos se subdivide en sub-haplogrupos, la denominación suele añadir un número (e.j. U1), y así de forma jerárquica a través de la estructura filogenética de la variabilidad mitocondrial.

El concepto de haplogrupo es sin embargo un tanto arbitrario, porque no existe un nivel jerárquico concreto a partir del cual se le asigne la categoría de haplogrupo a un conjunto dado de secuencias; el término es sin embargo funcional, y sirve para hacer estudios comparativos, que no podrían ser realizados a partir de los haplotipos, dado que la variabilidad estaría demasiado atomizada.

A lo largo de estos últimos años, se ha visto que determinados haplogrupos podrían estar implicados en la predisposición genética a determinadas enfermedades. La aproximación epidemiológica empleada para el estudio de la susceptibilidad genética explicada por variantes mitocondriales se realiza generalmente a través de estudios poblacionales tipo caso-control.

El objeto de estudio de este proyecto es realizar una revisión crítica de los métodos actuales para estimar la potencia estadística necesaria en un estudio caso-control y valorar y proponer aproximaciones nuevas.

1.2. Metodología

Para contrastar hipótesis sobre determinados haplogrupos se realizará un estudio tipo caso-control y se dispondrá de una tabla de contigencia de la siguiente forma:

Tipo paciente \ Haplogrupo	Haplogrupo 1	...	Haplogrupo k	Total filas
Control	N_{11}	...	N_{1k}	$N_{1\cdot}$
Caso	N_{21}	...	N_{2k}	$N_{2\cdot}$
Total columnas	$N_{\cdot 1}$	...	$N_{\cdot k}$	n

donde N_{1j} y N_{2j} denota al número de controles y casos, respectivamente, con el haplogrupo j.

Se podría contrastar si la distribuciones de casos y la de controles son la misma a lo largo de todos los haplogrupos, pero lo que realmente interesa es contrastar la homogeneidad sobre cada haplogrupo individididualmente, para así observar qué haplogrupos son más frecuentes en los casos con respecto a los controles. Además, a la hora de considerar un estudio mitocondrial, hay una estructura jerárquica, por lo que en general, existirá una gran dependencia entre las columnas (haplogrupos) y habrá clases que no sean disjuntas. Por ello se considerarán tablas marginales de la siguiente forma:

Tipo paciente \ Haplogrupo	Haplogrupo i	Resto de haplogrupos	Total filas
Control	N_{1i}	$\sum_{j\neq i} N_{1j}$	$N_{1\cdot}$
Caso	N_{2i}	$\sum_{j\neq i} N_{2j}$	$N_{2\cdot}$
Total columnas	$N_{\cdot i}$	$\sum_{j\neq i} N_{\cdot j}$	n

Para la realización del contraste se utilizará el estadístico χ^2 de Pearson, aunque en el próximo capítulo se plantearán distintas medidas para realizar el contraste.

El contraste χ^2 posee unas restricciones sobre las frecuencias esperadas para la convergencia del estadístico (regla de Cochran [6]); restricciones que, debido a la naturaleza de la variable a estudiar, serán violadas.

En el apartado de simulación de este proyecto (Capítulo 4), se considerará un escenario de una población con 10 haplogrupos: H, I, J, K, M, T, U, V, W y X. Alrededor del 95 % de la población europea pertenece a uno de estos haplogrupos; pero de éstos, 6 se encuentran en menos del 10 % de la población.

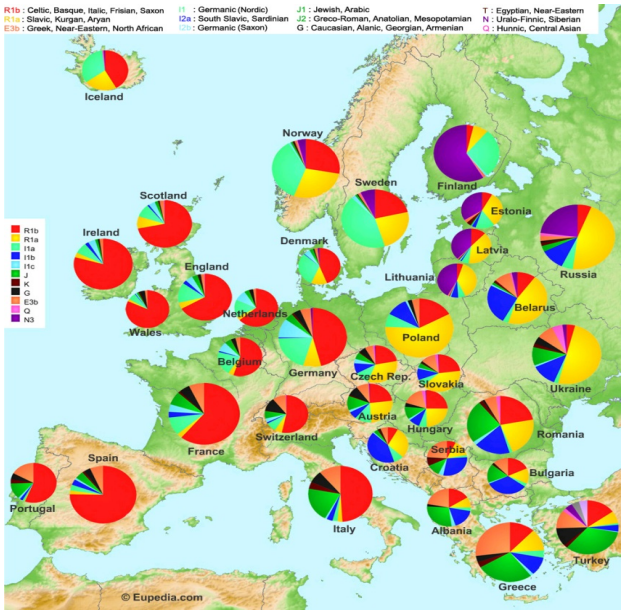

Figura 1.4: Mapa de la distribución de haplogrupos en Europa (extraído de http://www.disnorge.no).

Ante esta situación, el estadístico χ^2 de Pearson será artificialmente elevado, por lo que, si se acude a la tabla de la χ^2, estaremos aumentando el error de tipo I.

Por lo tanto, será necesario un método para calibrar la distribución de nuestro estadístico. Esto se estudiará en el Capítulo 3.

Capítulo 2

Medidas de discrepancia

Sean m poblaciones independientes e i.i.d.:

$$\begin{gathered} \textit{población } 1: \ X_{11}, \ldots, X_{1n_1} \\ \vdots \\ \textit{población } m: X_{m1}, \ldots, X_{mn_m}, \end{gathered}$$

con distribuciones $F_1, \ldots, F_m$ respectivamente, siendo todas discretas y con el mismo soporte.

Se pretende contrastar:

$$H_0 : F_1 = \ldots = F_m.$$

En la práctica se dispondrá de una tabla de contingencia de la siguiente forma:

	A_1	$\ldots$	A_k	Total filas
Población 1	N_{11}	$\ldots$	N_{1k}	$N_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
Población m	N_{m1}	$\ldots$	N_{mk}	$N_{m\cdot}$
Total columnas	$N_{\cdot 1}$	$\ldots$	$N_{\cdot k}$	n

donde N_{ij} denota el número de individuos de la población i-ésima, que poseen la característica A_j.

Si se denota por p_{ij} a la probabilidad de la clase A_j en la población i-ésima podría plantearse el contraste de la siguiente forma:

$$H_0 : p_{1j} = \ldots = p_{mj} \ para \ j = 1, \ldots k.$$

Bajo la hipotésis nula, se dispondrá de un único vector de probabilidades que denotaremos por $(p_1^0, \ldots, p_k^0)$.

2.1. Estadísticos tipo χ^2

La idea de este tipo de contrastes es comparar los valores observados de nuestra tabla con los valores esperados bajo la hipotésis nula mediante una distancia.

2.1.1. χ^2 de Pearson

Este estadístico, planteado por Pearson en 1900, tiene la siguiente expresión:

$$\chi^2 = \sum_{poblaciones} \sum_{clases} \frac{(observado - esperado)^2}{esperado}.$$

El problema es que se desconocen cuáles son los valores esperados bajo la hipótesis nula, por lo que estos han de ser estimados.

Sea $N_i = (N_{i1}, \ldots, N_{ik})$, se tiene que $N_i \sim Multinomial(n_i, (p_{i1}, \ldots, p_{ik}))$, o bajo la hipótesis nula $N_i \sim Multinomial(n_i, (p_1^0, \ldots, p_k^0))$.

Los valores esperados bajo la hipotésis nula para cada celda son de la forma $m_{ij} = n_i p_j^0$, y como las probabilidades p_j^0 son desconocidas, han de ser estimadas. La estimación por máxima verosimilitud (Apéndice A, Nota 1) resulta $\widehat{p}_j^0 = N_{\cdot j}/n$, por lo que el valor esperado estimado para la celda (i, j) es $\widehat{m}_{ij} = n_i N_{\cdot j}/n$.

Entonces el estadístico χ^2 se puede estimar de la siguiente forma [1]:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(N_{ij} - n_i N_{\cdot j}/n)^2}{n_i N_{\cdot j}/n}.$$

Bajo la hipotésis nula y ciertas condiciones sobre las frecuencias esperadas, se tiene:

$$\chi^2 \xrightarrow[n\to\infty]{} \chi^2_{(m-1)(k-1)}.$$

El problema de este estadístico es que si las poblaciones no son homogéneas, el estadístico será porporcional al tamaño muestral, n. Por lo tanto, no se podrán emplear para estudiar las diferencias entre poblaciones, ya que para grandes muestras, pequeños cambios sobre la hipotésis repercuten en grandes cambios en el estadístico. Por ello se planteó el siguiente estadístico (Cramer, 1946):

$$V^2 = \frac{\chi^2}{n\ min(m-1,\ k-1)},$$

que se mueve entre 0 y 1, tomando valores grandes en el caso de tener mucha heterogeneidad.

Este estadístico permite comparar el grado de homogeneidad en tablas de una misma dimensión.

2.1.2. Razón de verosimilitudes

La hipótesis a contrastar es:

$$H_0 : \ p_{ij} = p_{i\cdot}p_{\cdot j}$$

$$H_1 : \ p_{ij} \neq p_{i\cdot}p_{\cdot j}$$

Sea la función de razón de verosimilitudes es:

$$\Lambda = \frac{\prod_{i=1}^{m}\prod_{j=1}^{k}\left(\frac{N_{i\cdot}}{n}\frac{N_{\cdot j}}{n}\right)^{N_{ij}}}{\prod_{i=1}^{m}\prod_{j=1}^{k}\left(\frac{N_{ij}}{n}\right)^{N_{ij}}} = \frac{\prod_{i=1}^{m}\prod_{j=1}^{k}\left(\frac{N_{i\cdot}N_{\cdot j}}{n}\right)^{N_{ij}}}{\prod_{i=1}^{m}\prod_{j=1}^{k}N_{ij}^{N_{ij}}}$$

Se define el estadístico de razón de verosimilitudes [1] como:

$$G^2 = -2\ log(\Lambda) = 2\sum_{i=1}^{m}\sum_{j=1}^{k}N_{ij}log\left(\frac{N_{ij}}{\widehat{m}_{ij}}\right).$$

La desventaja de este estadístico respecto al χ^2 de Pearson es que trabaja con logaritmos, por lo que todas las entradas han de ser positivas.

2.1.3. Transformación de Freeman-Tukey

Para muestras pequeñas, la convergencia asintótica del estadístico χ^2 de Pearson no sirve de mucho, por lo que algunos autores sugieren realizar una transformación para estandarizar las desviaciones de cada celda, de forma que estas desviaciones transformadas sean normales de media 0 y desviación típica 1. Freeman y Tukey (1950) presentaron el siguiente estadístico basado en lo dicho anteriormente:

$$K_{FT}^2 = 4\sum_{i=1}^{m}\sum_{j=1}^{k}(\sqrt{N_{ij}} - \sqrt{m_{ij}})^2.$$

Estos tres estadísticos convergen asintóticamente en distribución a una $\chi^2_{(m-1)(k-1)}$.

Esta convergencia se tiene cuando n es grande, pero la aproximación se considera acertada cuando se verifica la **regla de Cochran** (1954): No debe de haber frecuencias esperadas menores que 1, y menos del 20 % de las frecuencias esperadas pueden ser menores que 5.

2.2. Logit χ^2

En este caso se analizarán tablas de la forma $m \times 2$. Para la realización de este contraste se define el siguiente estadístico propuesto por Berkson en 1955 y 1968 [3]:

$$\chi^2_B = \sum_{i=1}^{m} \frac{N_{i1}N_{i2}}{N_{i\cdot}}(l_i - L_i)^2,$$

donde:

$$l_i = log\ \frac{N_{i1}}{N_{i2}} \quad L_i = log\ \frac{m_{i1}}{m_{i2}}$$

son los logits observados y esperados respectivamente.

Se puede reescribir el estadístico de otra forma:

$$\chi^2_B = \sum_{i=1}^{m} \frac{l_i - L_i}{\frac{1}{N_{i1}} + \frac{1}{N_{i2}}}.$$

Tanton este estadístico, como los tipo χ^2, utilizan de una forma u otra la estimación de m_{ij}, que depende solamente los totales por filas y columnas, sin importar cuál es el orden de filas o columnas. Por lo tanto, las permutaciones de filas o columnas no alterarán estos estadísticos, es decir la clasificación es tratada de una forma nominal. Para variables ordenadas estos métodos estarían ignorando información, por lo que habría que considerar otra clase de medidas de discrepancia.

2.3. Odds Ratio

Normalmente, a un investigador le gustaría saber algo más a parte del hecho de que las variables de estudio sean o no independientes. Sería de interés conocer el grado de dependecia existente, por muy pequeño que sea.

Para abordar este problema se estudiarán las tablas 2×2, aunque se puede generalizar a otro tipo de tablas.

Se define la *odds* para la fila 1 de la siguiente forma:

$$\Omega_1 = \frac{p_{12}}{p_{11}}.$$

y de la misma forma definimos la *odds* para la fila 2:

$$\Omega_2 = \frac{p_{22}}{p_{21}}.$$

Definimos la *odds ratio* [1] como:

$$\theta = \frac{\Omega_2}{\Omega_1} = \frac{p_{22}/p_{21}}{p_{12}/p_{11}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Notar que la *odd* de cada fila, Ω_i se puede escribir en función de las probabilidades condicionadas a las filas

$$\Omega_i = \frac{p_{2(i)}}{p_{1(i)}},$$

por lo que se tiene:

$$\theta = \frac{p_{2(2)}/p_{1(2)}}{p_{2(1)}/p_{1(1)}}.$$

Las distribuciones condicionales $(p_{1(1)}, p_{2(1)})$ y $(p_{1(2)}, p_{2(2)})$ serán idénticas, y por lo tanto las variables serán independientes, si y sólo si $\Omega_1 = \Omega_2$; es decir, si $\theta = 1$.

Si $1 < \theta < \infty$ los individuos de la segunda fila son más propensos a tomar la segunda respuesta que los de la primera fila, y viceversa si $0 \leq \theta < 1$.

En la práctica las probabilidades $\{p_{ij}\}$ son desconocidas, y por lo tanto también lo es θ; en consecuencia θ ha de ser estimada:

$$\widehat{\theta} = \frac{N_{11}N_{22}}{N_{12}N_{21}}.$$

Si denotamos por $\sigma(\widehat{\theta})$ a la desviación típica de $\widehat{\theta}$, se tiene:

$$\sigma(\widehat{\theta}) = \frac{\theta}{\sqrt{n}}\sqrt{\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}},$$

que se estima de la siguiente forma:

$$\widehat{\sigma}(\widehat{\theta}) = \widehat{\theta}\sqrt{\frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}}.$$

Se tiene que $\widehat{\theta}$ converge a una distribución normal con desviación típica la correspondiente.

2.3.1. Logaritmo de Odds Ratio

Notar que la odds ratio es una función multiplicativa de las celdas de nuestra tabla, por lo tanto su logaritmo será una función aditiva:

$$log\ \theta = log\ p_{11} - log\ p_{12} - log\ p_{21} + log\ p_{22}.$$

El logaritmo es simétrico respecto al 0, en el sentido de que dos valores opuestos indican un mismo grado de dependencia, pero de sentido opuesto.

Se tiene también la convergencia de $log\ \widehat{\theta}$ a una normal, aunque $log\ \widehat{\theta}$ lo hace más rápidamente que $\widehat{\theta}$. Si denotamos por $\sigma(log\ \widehat{\theta})$ a la desviación típica de $log\ \widehat{\theta}$ se tiene:

$$\widehat{\sigma}(log\ \widehat{\theta}) = \sqrt{\frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}},$$

lo que nos permitiría construir intervalos de confianza de significación α para $log\ \theta$ de la forma:

$$log\ \widehat{\theta} \pm z_{\alpha/2}\ \widehat{\sigma}(log\ \widehat{\theta}),$$

donde $z_{\alpha/2}$ denota el cuantil $\alpha/2$ de una $N(0,1)$.

2.3.2. Funciones generales de la Odds Ratio

Existen diversas medidas de asociación que son funciones monótonas de θ. Sea $f(\theta)$ una función monótona positiva de θ tal que $f(1) = 1$, entonces una medida normalizada basada en $f(\theta)$ será:

$$g(\theta) = \frac{f(\theta) - 1}{f(\theta) + 1}.$$

La desviación típica de $g(\widehat{\theta})$ es:

$$\sigma(g(\widehat{\theta})) = \frac{[1 - g(\theta)]^2 f'(\theta)}{2}\ \sigma(\widehat{\theta}).$$

A continuación se mostrarán dos medidas de interés, propuestas por Yule en 1900 y 1912 [3]:

$$Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{p_{11}p_{22}/p_{12}p_{21} - 1}{p_{11}p_{22}/p_{12}p_{21} + 1} = \frac{\theta - 1}{\theta + 1},$$

$$Y = \frac{\sqrt{p_{11}p_{22}} - \sqrt{p_{12}p_{21}}}{\sqrt{p_{11}p_{22}} + \sqrt{p_{12}p_{21}}} = \frac{\sqrt{\theta} - 1}{\sqrt{\theta} + 1},$$

Estas dos medidas toman valores entre -1 y 1, donde 0 indica no asociación.

Si una observación toma el mismo valor para cada una de las variables (primera respuesta o segunda) se dirá que es una observación concordante (discordante en caso contrario). La Q estima la diferencia entre las proporciones de observaciones concordantes y discordantes, siendo esta cercana a 1 cuando la mayoría son observaciones concordantes, y -1 en caso contrario.

La Y se interpreta como una medida de coligación.

2.4. Test exacto de Fisher

Dada una tabla 2×2:

	A_1	A_2	Total filas
Población 1	N_{11}	N_{12}	$N_{1\cdot}$
Población 2	N_{21}	N_{22}	$N_{2\cdot}$
Total columnas	$N_{\cdot 1}$	$N_{\cdot 2}$	n

Fisher argumentó que, bajo la hipotésis nula de homogeneidad, las redistribuciones de los sujetos bajo la restricción de los totales, es decir, cada una de las N_t tablas posibles con estos totales, son igualmente probables; donde:

$$N_t = \sum_{x=0}^{N_{1\cdot}} \begin{pmatrix} N_{\cdot 1} \\ x \end{pmatrix} \begin{pmatrix} N_{\cdot 2} \\ t-x \end{pmatrix}.$$

El test exacto de Fisher [8] consiste en calcular la probabilidad de la tabla original, que se obtiene mediante una distribución hipergeométrica:

$$p = \frac{\begin{pmatrix} N_{1\cdot} \\ N_{11} \end{pmatrix} \begin{pmatrix} N_{\cdot 2} \\ N_{21} \end{pmatrix}}{\begin{pmatrix} n \\ N_{\cdot 1} \end{pmatrix}} = \frac{N_{1\cdot}!N_{2\cdot}! + N_{\cdot 1}!N_{\cdot 2}!}{N_{11}!N_{12}!N_{21}!N_{22}!},$$

y ver el porcentaje de permutaciones cuya probabilidad es tan o más extrema que la de la tabla original.

El problema de este estadístico es que es discreto, por lo que no podremos trabajar con niveles de significación fijados de antemano.

Capítulo 3

Calibración de estadístico mediante permutación

Como ya se dijo antes, se desea contrastar la hipotésis de que la distribución de sanos y enfermos es la misma respecto a un determinado haplogrupo, para lo cual se realizará un contraste de homogeneidad utilizando el estadístico χ^2 de Pearson.

El estudio que se ha llevado a cabo para los resultados simulados de este proyecto se basa en cosiderar 10 haplogrupos a los que pertenece aproximadamente el 95 % de la población europea. El problema que se plantea, es que de los 10 haplogrupos, 6 se encuentran en menos del 10 % de la población, por lo que es probable que se incumpla la regla de Cochran.

En caso de incumplirse la regla de Cochran el estadístico será artificialmente elevado, por lo que se obtendrán falsos positivos (alta probabilidad de cometer error de tipo I). Los valores tabulados de la χ^2 no deben de ser tenidos en cuenta.

Para abordar este problema se plantean diversas soluciones:

- Agrupar las columnas con menores frecuencias esperadas en la tabla de nuestro estudio.
- Realizar el test exacto de Fisher con una corrección de la significación.
- Calibración del estadístico utilizando permutaciones.

La primera solución enmascara el efecto que pueden producir los haplogrupos poco frecuentes.

El test exacto de Fisher utiliza un estadístico discreto, por lo que no se puede realizar el contraste para una significación deseada.

En caso de que no se verifique la regla de Cochran se optará por calibrar el estadístico de contraste mediante permutación.

El método que aquí se verá fue propuesto por Roff y Bentzen en 1989 [14]. Se basa en la generación de permutaciones de la tabla original de tal forma:

1. Se calcula el valor de estadístico para la tabla original, que se denotará por χ_0.

2. Se genera una tabla permutando los valores de la tabla original, de forma que los totales por fila y columna permanezcan constantes.

3. Para esta permutación se calcula el valor del estadístico, χ_r.

4. Se repiten B veces los pasos 2 y 3.

5. El p-valor del estadístico original será estimado mediante la proporción de estadísticos simulados mayores o iguales que χ_0.

Una condición suficiente para que el contraste calibrado por permutación sea exacto e insesgado es la intercambialidad de las observaciones de la muestra bajo la hipótesis nula [8]. Las observaciones se dirán intercambiables si la probabilidad de cualquier resultado conjunto, es el mismo independientemente del orden en el que las observaciones se consideren.

En este caso, las observaciones serán intercambiables si, bajo la hipotésis nula, se verifican las siguientes condiciones [8]:

1. Las variables fila y columna son mutuamente independientes.

2. Las observaciones son independientes.

Estas condiciones se cumplen en nuestro estudio, la primera por la definición de la hipotésis nula de esa forma y la segunda por como se ha tomado la muestra.

Se tiene además la convergencia del estadístico calibrado por permutación a la distribución asintótica $\chi^2_{(k-1)(m-1)}$.

Ahora el problema radica en como calcular una tabla permutada. Dada una tabla:

	A_1	$\dots$	A_k	Total filas
Población 1	N_{11}	$\dots$	N_{1k}	$N_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
Población m	N_{m1}	$\dots$	N_{mk}	$N_{m\cdot}$
Total columnas	$N_{\cdot 1}$	$\dots$	$N_{\cdot k}$	n

Se define el vector:

$$A = \begin{pmatrix} 1 \\ \vdots \\ N_{\cdot 1} \\ \vdots \\ 1 \\ 2 \\ \vdots \\ N_{\cdot 2} \\ \vdots \\ 2 \\ 3 \\ \vdots \\ k \\ \vdots \\ N_{\cdot k} \\ \vdots \\ k \end{pmatrix} \xrightarrow{Permutación} A' = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 2 \\ k \\ 1 \\ 1 \\ \vdots \\ 2 \\ 1 \\ k \\ k \\ 1 \\ 1 \end{pmatrix}.$$

La tabla permutada resultará de la siguiente forma:

	A_1	$\dots$	A_k	Total filas
Población 1	N'_{11}	$\dots$	N'_{1k}	$N_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
Población m	N'_{m1}	$\dots$	N'_{mk}	$N_{m\cdot}$
Total columnas	$N_{\cdot 1}$	$\dots$	$N_{\cdot k}$	n

siendo $N'_{1j} = \sharp\{A'_k = j,\ 1 \leq k \leq N_{1\cdot}\}$ y $N'_{ij} = \sharp\{A'_k = j,\ \sum_{l=1}^{i-1} N_{l\cdot} < k \leq \sum_{l=1}^{i} N_{l\cdot}\}$ para $i > 1$.

Notar que al definir A de esa forma se fijan los totales por columna, y al construir la tabla a partir de la permutación, A' se fijan los totales por filas.

Este algoritmo, aunque intuitivo, es muy farragoso; ya que conlleva la numeración de todos los individuos de la tabla, y esto puede lastrar mucho la eficiencia del algoritmo.

Para solucionar este problema se puede considerar la genereración de permutaciones viendo cuál es la distribución de probabilidades de las tablas cuyas marginales coinciden con las de la tabla original.

Si denotamos $r = (N_{1\cdot}, \ldots, N_{m\cdot})$ y $c = (N_{\cdot 1}, \ldots, N_{\cdot k})$ se define $T_X = \{Y \ / \ Y \ es \ m \times k \ \ con \ totales \ r \ y \ c\}$. La probabilidad de cada tabla de T_X bajo la hipotésis nula es:

$$\mathbb{P}(Y) = \frac{\prod_{i=1}^{m} N_{i\cdot}! \prod_{j=1}^{k} N_{\cdot j}!}{n \prod_{i=1}^{m} \prod_{j=1}^{k} y_{ij}!},$$

donde y_{ij} denota los valores de la tabla Y [9].

Se tiene que el p-valor de la tabla X es:

$$p_X = \sum_{Y \in s(x)} P(Y),$$

con $s(X) = \{Y \ / \ Y \in T_X, \ P(Y) \leq P(X)\}$.

Se tiene que $p_X = \prod_{i=1}^{m-1} \prod_{j=1}^{k-1} p_{ij}$ donde

$$p_{ij} = \frac{\begin{pmatrix} N_{\cdot j} - s_{ij} \\ y_{ij} \end{pmatrix} \begin{pmatrix} n - v_{ij} - N_{\cdot j} + s_{ij} \\ N_{i\cdot} - t_{ij} - y_{ij} \end{pmatrix}}{\begin{pmatrix} n - v_{ij} \\ N_{i\cdot} - t_{ij} \end{pmatrix}}$$

con

$$s_{ij} = \sum_{g=1}^{i-1} y_{gj}, \quad t_{ij} = \sum_{h=1}^{j-1} y_{ih}$$

$$v_{ij} = \sum_{g=1}^{i-1} N_{g\cdot} + \sum_{h=1}^{j-1} N_{\cdot h} - \sum_{g=1}^{i-1} \sum_{h=1}^{j-1} y_{gh}.$$

Para la demostración ver [11].

De esta forma, la probabilidad de una tabla se puede calcular de una forma recursiva, y permite generar tablas de T_X ya que da cotas para los valores y_{ij}.

$$0 \leq y_{ij} \leq c_j - s_{ij}$$

$$0 \leq r_i - t_{ij} - y_{ij} \leq n - v_{ij} - c_j + s_{ij},$$

o reescrito de otra forma:

$$l_{ij} \leq y_{ij} \leq u_{ij},$$

donde

$$l_{ij} = max\{0, v_{ij} + c_j - s_{ij} + r_i - t_{ij} - n\}$$

y

$$u_{ij} = min\{c_j - s_{ij}, r_i - t_{ij}\}.$$

(Para más detalles ver [11]).

Volviendo a la idea del primer algoritmo, se puede plantear un método más sencillo por el hecho de trabajar con tablas 2×2 (si se genera una entrada de la tabla, el resto vienen determinadas por las restricciones de fila y columna).

No tendría más que tomarse $N'_{11} = X$, con $X \sim Hiper(N_{\cdot 1}, N_{\cdot 2}, N_{1\cdot})$.

	A_1	A_2	Total filas
Población 1	x	$N_{1\cdot} - x$	$N_{1\cdot}$
Población 2	$N_{\cdot 1} - x$	$N_{\cdot 2} - (N_{1\cdot} - x)$	$N_{2\cdot}$
Total columnas	$N_{\cdot 1}$	$N_{\cdot 2}$	n

Capítulo 4

Simulación

En este capítulo se hará una comparación entre la calibración del estadístico en base a permutaciones con la aproximación a una χ^2_1. También realizará un estudio del comportamiento de la potencia estadística en este tipo de contrastes, lo que proporcionará interesantes aplicaciones.

Se fijará un escenario de 11 haplogrupos (10 haplogrupos y 1 residual) cuyas frecuencias en controles son conocidas [18] y en casos son fijadas.

Se realizarán simulaciones para obtener valores de potencia para distintos haplogrupos y tamaños muestrlaes bajo distintas hipotésis.

4.1. Diferencia entre calibraciones

En una primera aproximación se han simulado 5000 tablas de contingencia para cada haplogrupo y cada desviación de la hipotésis nula en las que el número de casos es igual al número de controles, y se obtiene un valor de potencia para cada método de calibrado (para la calibración por permutación se han considerado 5000 permutaciones para cada tabla). Se ha tomado com osignificación $\alpha = 5\,\%$.

En las gráficas de la página siguiente están representadas curvas de potencia simuladas en las que se observa que la aproximación asintótica es más potente que la obtenida a través de permutación, pero esto es debido a la violación de la regla de Cochran.

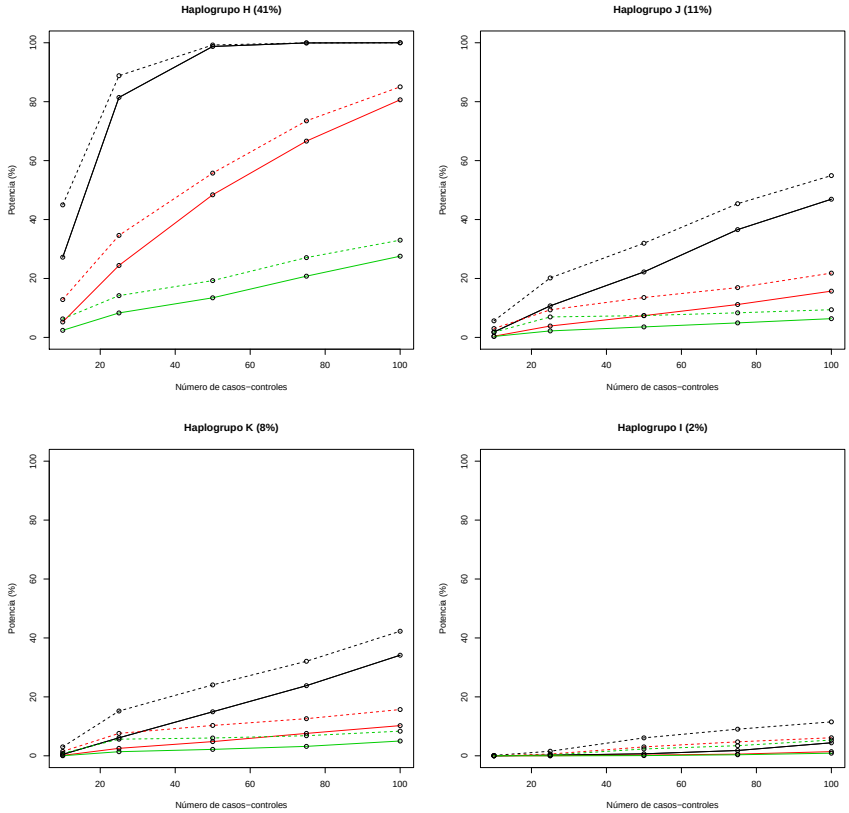

Figura 4.1: Curvas de potencia para los haplogrupos H, J, K e I para distintos métodos de calibrado. El trazo continuo denota la potencia obtenida mediante permutación y el trazo discotinuo la asintótica. El color negro denota un incremento de la frecuencia en los casos del 100 % respecto a los controles, el rojo del 50 % y el verde del 25 %.

A continuación se muestra el porcentaje de tablas en cada caso que no cumplen la regla de Cochran:

Haplogrupo H					
nº de casos/controles / Incremento frecuencia	10	25	50	75	100
↑ 100 %	0.99	0.53	0.04	0	0
↑ 50 %	0.96	0.01	0.	0	0
↑ 25 %	0.95	0	0	0	0

Haplogrupo J					
nº de casos/controles / Incremento frecuencia	10	25	50	75	100
↑ 100 %	0.99	0.91	0.34	0.07	0.01
↑ 50 %	0.99	0.91	0.34	0.07	0.01
↑ 25 %	1	0.96	0.44	0.09	0.01

Haplogrupo K					
nº de casos/controles / Incremento frecuencia	10	25	50	75	100
↑ 100 %	1	0.98	0.66	0.28	0.09
↑ 50 %	1	0.9	0.72	0.31	0.09
↑ 25 %	1	0.99	0.79	0.35	0.11

Haplogrupo I					
nº de casos/controles / Incremento frecuencia	10	25	50	75	100
↑ 100 %	1	1	1	1	0.98
↑ 50 %	1	1	1	1	0.99
↑ 25 %	1	1	1	1	1

A continuación se muestran los *boxplots* de los p-valores obtenidos para ambas calibraciones para los 4 haplogrupos. En ellos se aprecia una diferencia entre los dos métodos de calibrado, diferencia que parece acentuarse según disminuye la frecuencia poblacional.

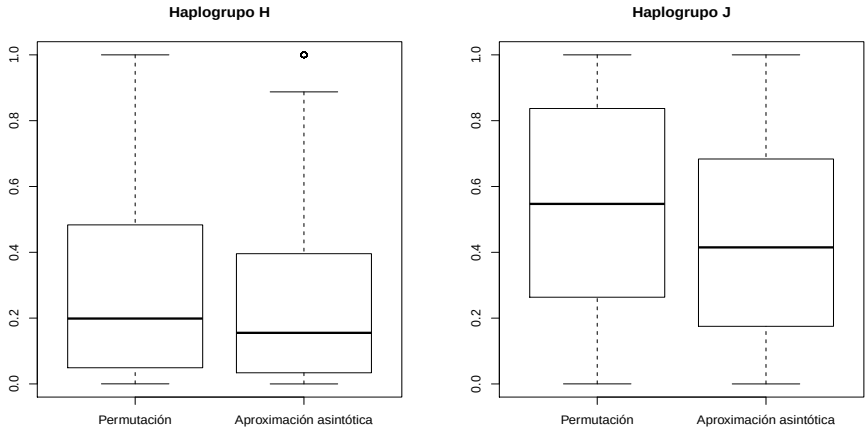

Figura 4.2: *Boxplot* para los p-valores considerando un incremento del 25 % y 100 casos-controles y los haplogrupos H y J.

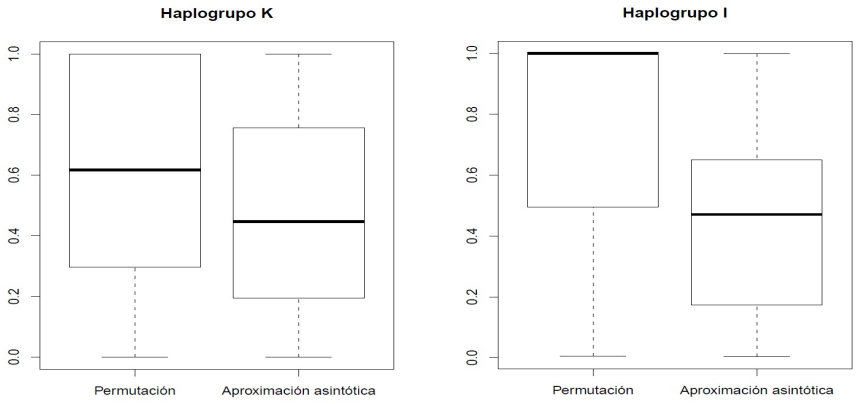

Figura 4.3: *Boxplot* para los p-valores considerando un incremento del 25 % y 100 casos-controles y los haplogrupos K e I.

Pero este hecho ha de comprobarse. Para ello se ha realizado el contraste de los signos (Apéndice B.1) para comprobar si el efecto de considerar una u otra calibración tiene un efecto significativo sobre el p-valor. En la siguiente tabla se muestra el valor del estadístico, su p-valor y la diferencia media entre ambas calibraciones para el caso de un tamaño de 100 casos-controles y una frecuencia en casos un 25 % mayor que la frecuencia en controles. Como hipotésis alternativa se considera que la calibración asinótica proporciona p-valores estocásticamente mayores que los obtenidos por permutación:

Haplogrupo	Estadístico	p-valor	Promedio de las diferencias
Haplogrupo H	4873	0	0.04
Haplogrupo J	4631	0	0.10
Haplogrupo K	4552	0	0.12
Haplogrupo I	4049(∗)	0	0.24

(En el caso del haplogrupo I, el estadístico se construye con 4954 muestras, en lugar de las 5000 del resto de haplogrupos, ya que hay tablas cuyas marginales son nulas y se han excluido).

Entonces se tiene que el método de calibración es influyente en el p-valor.

Las diferencias observadas entre los distintos p-valores parece que son mayores cuando la frecuencia en los controles es menor, por lo que ahora se analizará este hecho.

Para ello se considerarán otra vez los haplogrupos H, J, K e I. Se generarán 5000 tablas de contingencia y se calcularán el p-valor promedio (Apéndice D) y la potencia obtenida.

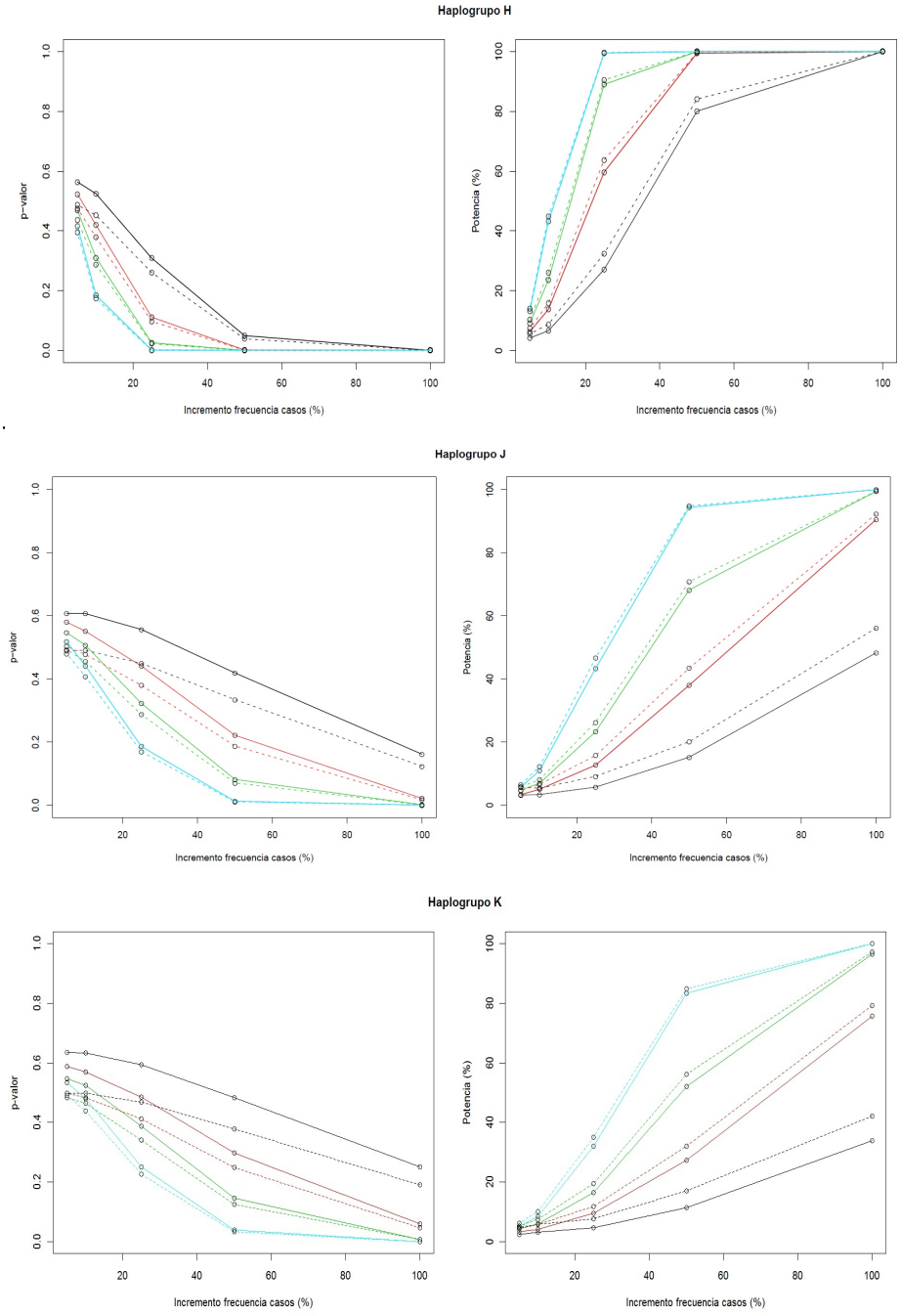

Figura 4.4: Comparación de los p-valores bootstrap y valores de potencia estimados para las dos calibraciones (linea continua la de permutación y discontinua la asintótica), en función del incremento de la frecuencia del haplogrupo de casos respecto a los controles y para diversos tamaños muestrales (negro 100, rojo 250, verde 500 y azul 1000 casos-controles). Todo esto haplogrupos H, J y K

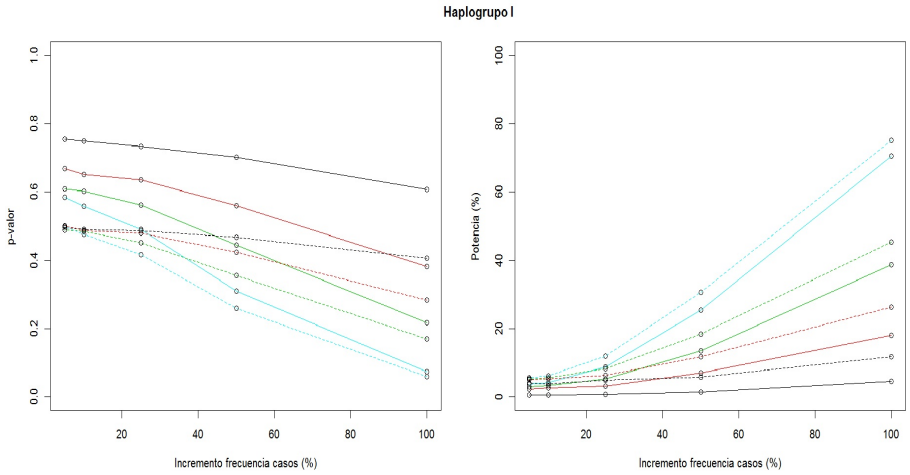

Figura 4.5: Comparación de los p-valores bootstrap y valores de potencia estimados para las dos calibraciones (linea continua la de permutación y discontinua la asintótica) para el haplogrupo I, en función del incremento de la frecuencia del haplogrupo de casos respecto a los controles y para diversos tamaños muestrales (negro 100, rojo 250, verde 500 y azul 1000 casos-controles).

A la vista de estas gráficas se pueden sacar tres conclusiones:

- Para un tamaño muestral mayor, la diferencia entre ambas calibraciones disminuye. No hay que olvidar el hecho de que la distribución límite de estadístico calibrado por permutación coincide con la distribución asintótica teórica.
- La diferencia entre las calibraciones es menor cuanto mayor sea la desviación de la hipotésis nula, ya que los p-valores tienden a ser más pequeños y están acotados inferiormente por 0.
- La diferencia entre las calibraciones es mayor (tanto en el p-valor como en la potencia) cuanto menor sea la frecuencia del haplogrupo en los controles.

Para corroborar este último hecho se simularán 5000 tablas para cada haplogrupo considerado (de nuevo, H, J, I, K), un tamaño muestral de 100 casos-controles y un incremento de la frecuencia en los casos del 25 % respecto a la frecuencia en los controles; y se diferenciarán los p-valores obtenidos para cada tabla.

En primer lugar se muestran en la página siguiente los boxplots de las diferencias obtenidas para cada haplogrupo. Se aprecia una tendencia creciente en la diferencia de p-valores a medida que la frecuencia en controles del haplogrupo en cuestión disminuye.

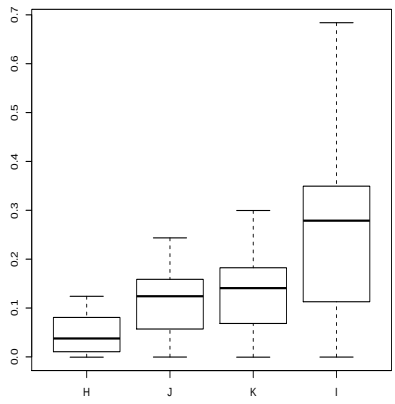

Figura 4.6: Boxplots de las diferencias entre p-valores obtenidos por permutación y por aproximación asintótica para distintos haplogrupos.

Para comprobarlo se ha realizado el contraste de Kruskal-Wallis (Apéndice B.3), obteniendo un p-valor practicamente 0, con lo que hay indicios suficientes como para asegurar que la frecuencia poblacional es influyente.

Finalmente se utilizará el test Wilcoxon-Mann-Whitney (Apéndice B.2) para contrastar la homogeneidad de las diferencias de p-valores haplogrupo a haplogrupo. En la tabla siguiente se muestran los p-valores obtenidos de considerar el test unilateral correspondiente (tomando como hipotésis alternativa que la diferencia de p-valores es estocásticamente mayor para el haplogrupo con menor frecuencia poblacional):

	Haplogrupo J	Haplogrupo K	Haplogrupo I
Haplogrupo H	0	0	0
Haplogrupo J	–	0	0
Haplogrupo K	–	–	0

Todas las comparaciones han dado como resultado un p-valor significativo.

Una explicación al hecho de que la frecuencia en controles sea tan influyente en la diferencia de p-valores, es que el estadístico χ^2 será artificialmente mayor cuando la frecuencias en controles sean menores; y en estos casos la diferencia entre p-valores se acentúa.

4.2. Análisis de la potencia

Centrándose ahora en la estimación de curvas de potencia, en las cuatro primeras gráficas de este capítulo se ha visto, además de la diferencia entre ambas calibraciones, que hay cierta heterogeneidad de las curvas de potencia, que se aprecia mejor en las siguientes gráficas.

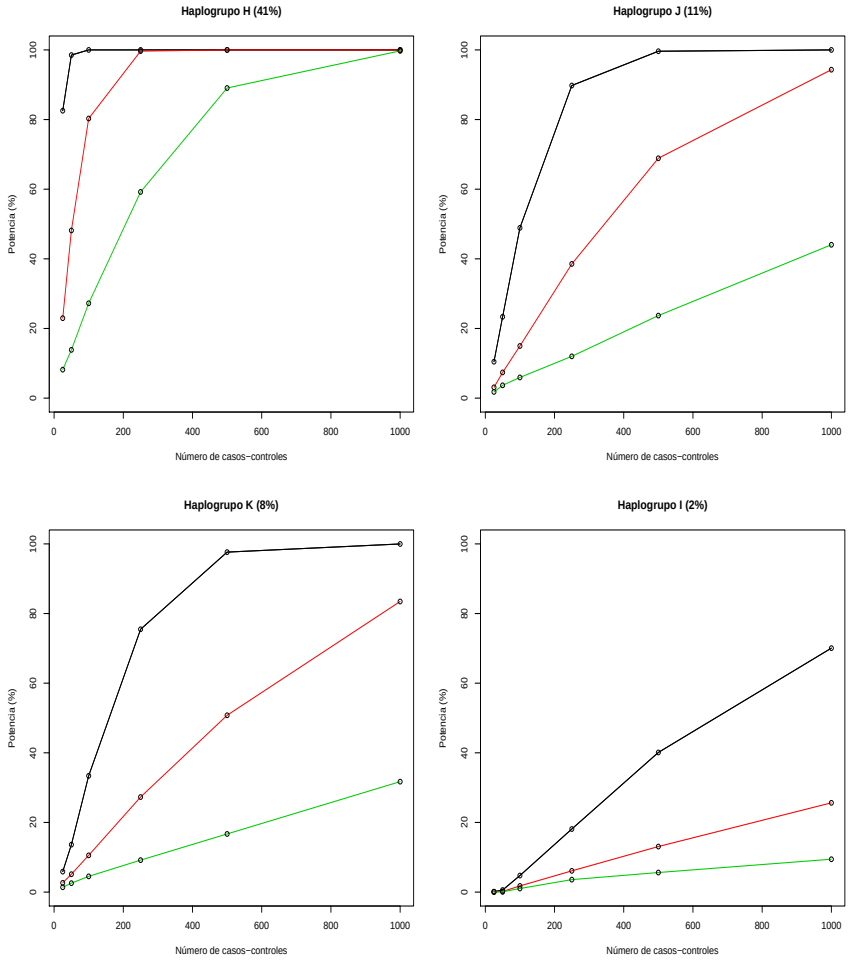

Figura 4.7: Curvas de potencia simuladas para distintas hipotésis. Las curvas negras representan un incremento del 100 % de la frecuencia en los casos respecto a la de los controles, las rojas un incremento del 50 % y las verdes del 25 %

En ellas se ve que la potencia no depende tan sólo del número de casos y controles, si no que también influye la proporción del haplogrupo en la población (cuando la proporción en la población es grande, una pequeña desviación de la hipotésis nula se detecta más facilmente) y de la magnitud de cambio entre la frecuencia del haplogrupo en los casos y en los controles (a mayor diferencia, se tiene una desviación mayor de la hipotésis nula, y por lo tanto mayor potencia).

Notar que se ha considerado que el número de controles y el de casos es el mismo, pero una diferencia entre estos dos valores también alterará la curva de potencia, como se puede ver en la gráfica de la página siguiente.

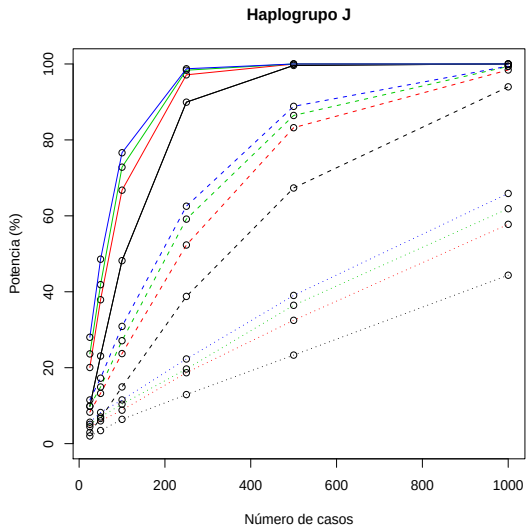

Figura 4.8: Curvas de potencia para el happlogrupo J, para distintas *odds* control-caso: negro, 1; rojo, 2; verde, 3; azul, 4; y distintas frecuencias del haplogrupo en los casos: trazo continuo, incremento del 100 %, discontinuo , incremento del 50 %; punteado, incremento del 25 %.

Fijado el número de casos, para un mayor número de controles se tiene una mayor potencia.

Esto se debe a que, si se aumenta el número de controles, la población "control" estará mejor representada, por lo que se dispondrá de más información para contrastar si las dos poblaciones son homogéneas o no.

Por todo esto la relación entre la potencia y el número de casos va a ser compleja. Sería de interés obtener una medida que involucrase los términos antes mecionados, para obtener curvas de potencia homogéneas en función de ese parámetro.

4.3. Estandarización de las curvas de potencia

Si se denota por p_0 a la frecuencia del haplogrupo en la población, p_1 a la frecuencia del haplogrupo en los casos y N_{ca} el número de controles (que en este caso coincide con el número de controles), se define [17]:

$$N_{sc} = \frac{N_{ca}(p_1 - p_0)^2}{p_1(1-p_1) + p_0(1-p_0)}.$$

En la gráfica de la izquierda se observa como los valores de potencia son muy distintos entre si, mientras que en la gráfica de la derecha, considerando la medida estandarizadora antes definida, los valores de potencia tienden a agruparse en torno a una curva.

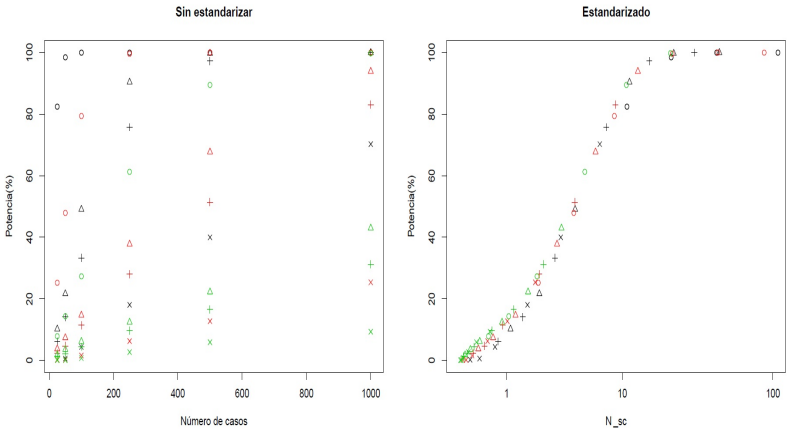

Figura 4.9: Estandarización de la potencia, donde los círculos denotan la potencia para el haplogrupo H, los triángulos para el haplogrupo J, las cruces para el haplogrupo K y las aspas para el haplogrupo I. Los colores negro, rojo y verde denotan un incremento del 100 %, 50 % y 25 % respectivamente

Notar que esta medida se ha definido cuando el número de controles y el de casos es el mismo. El artículo en el que esta inspirado este proyecto, [17], se limita a abordar esta situación, pero podemos generalizar este parámetro a situaciones en las que el número de controles difiera del número de casos.

Por tanto se definirá una nueva medida (Apéndice A, Nota 2):

$$N'_{sc} = \frac{(p_1 - p_0)^2}{\frac{p_1(1-p_1)}{N_{ca}} + \frac{p_0(1-p_0)}{N_{co}}},$$

siendo N_{co} el número de controles.

En las gráficas siguientes se ve como, con esta nueva medida, los distintos valores de la potencia tienden a agruparse en una misma curva.

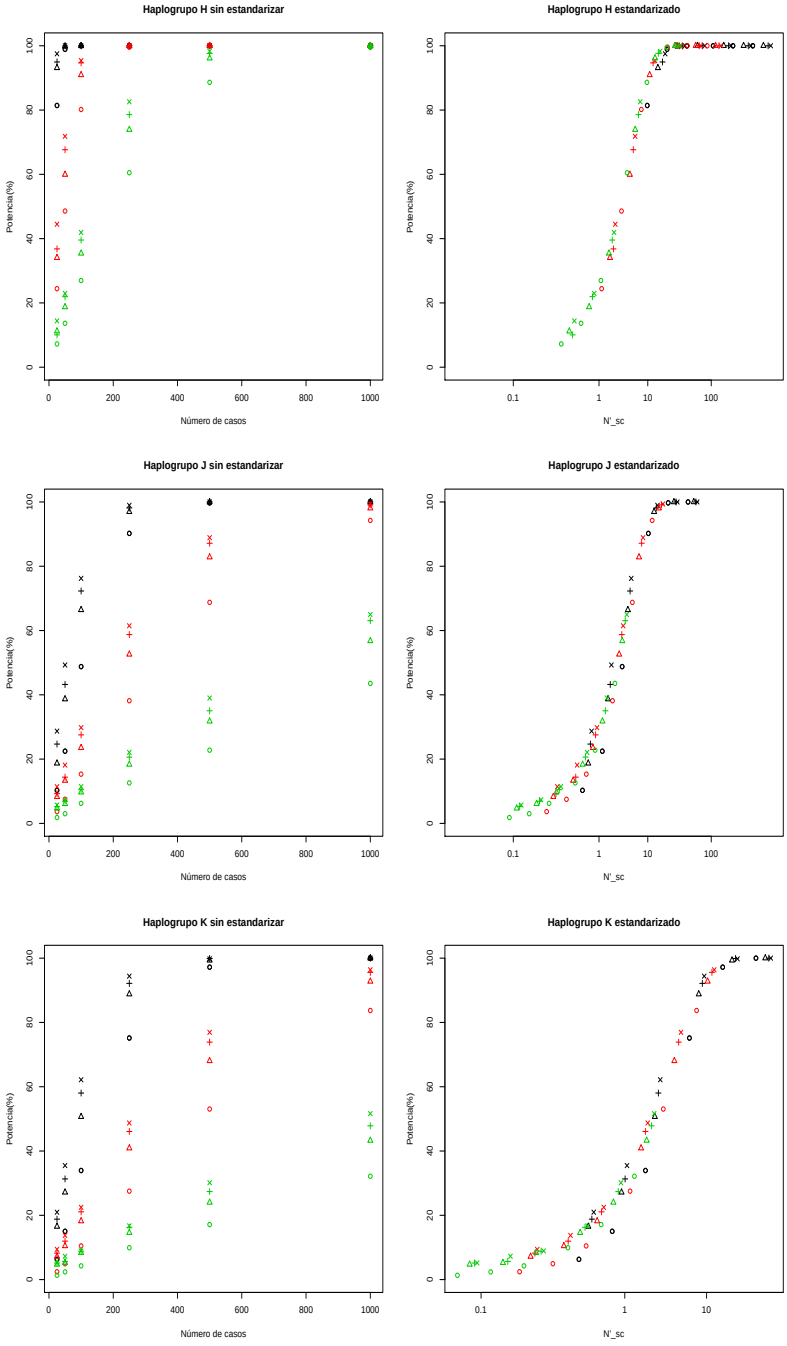

Figura 4.10: Estandarización de valores de potencia para los haplogrupo H, J y K. Los círculos denotan una *odds* control-caso de 1, triángulos para 2, 3 para las cruces y 4 para las aspas. Los cambios de frecuencias en los casos dados por el color negro, rojo y verde para incrementos de 100 %, 50 % y 25 % respectivamente

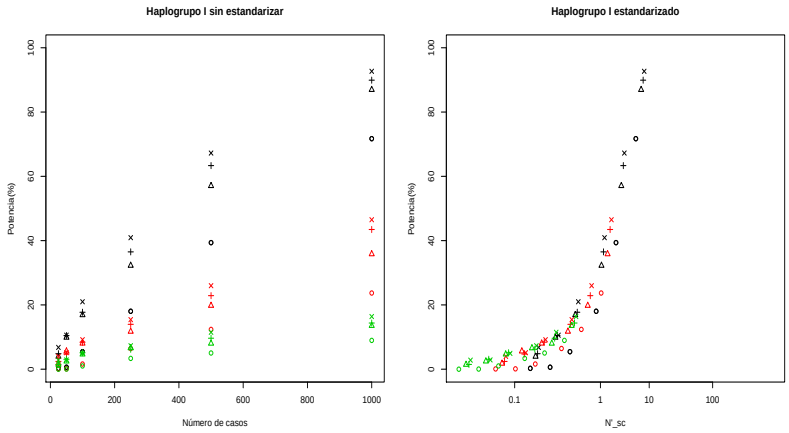

Figura 4.11: Estandarización de valores de potencia para el haplogrupo I. Los círculos denotan una *odds* control-caso de 1, triángulos para 2, 3 para las cruces y 4 para las aspas. Los cambios de frecuencias en los casos vienen dados por el color negro, rojo y verde para incrementos de 100 %, 50 % y 25 % respectivamente

Se ha encontrado una medida que tiene una relación clara con la potencia del contraste (al contrario de la medida "número de controles"), pero ahora cabría pensar en su posible utilidad.

4.4. Utilidades de la medida N'_{sc}

Esta medida puede ser utilizada para calcular el número mínimo de casos-controles necesarios para rechazar una hipotésis con una potencia deseada, despejando N_{ca} en la definición de N_{sc}:

$$N_{ca_{min}} = \left[N_{sc} \frac{p_1(1-p_1) + p_0(1-p_0)}{(p_1-p_0)^2} \right] + 1,$$

donde $[x]$ denota la parte entera de x.

Pero en un estudio real será de mayor utilidad obtener el número mínimo de controles, fijado el de casos, para rechazar una determinada hipotésis con cierta probabilidad y fijado el número de casos, con lo cual resultaría:

$$N_{co_{min}} = \left[\frac{p_0(1-p_0)}{\frac{(p_1-p_0)^2}{N'_{sc}} - \frac{p_1(1-p_1)}{N_{ca}}} \right] + 1 \quad si \quad N_{ca} > \frac{p_1(1-p_1)N'_{sc}}{(p_1-p_0)^2}.$$

Esto es una interesante aplicación, ya que en un estudio tipo caso-control, el número de casos será limitado y se tendrán en cuenta todo los indivduos con la etiqueta "caso", especialmente en enfermedades poco frecuentes. En cambio un investigador no se encuentra con esta limitación a

la hora de considerar los controles, ya que la población control puede ser tan grande como uno desee.

Notar que la restrición sobre el número de casos indica que no se puede aumentar la potencia tanto como se quiera, si no que debemos de tener un número suficiente de casos, que ha de ser mayor cuanto mayor sea la potencia que deseamos alcanzar y menor si se tiene una mayor desviación de la hipotésis nula.

Otra aplicación de interés sería, dados el número de casos y controles de nuestro estudio y la frecuencia del haplogrupo en la población ver cuáles son las desviaciones mínimas de la hipotésis nula que se pueden detectar con una potencia deseada, para un nivel de significación.

Para analizar esta situación, se deben de analizar los casos del incremento y decrecimiento de la frecuencia por separado, ya que la variación en torno a la frecuencia poblacional no es simétrica (e.j. es posible un incremento del 200 % para el haplogrupo I, pero no un decrecimiento). Sea entonces:

$$p_1 = p_0 + c\ a\ p_0,$$

siendo $a > 0$, $c = -1$ *ó* 1 (dependiendo si se considerá un incremento o decrecimiento respecto a la frecuencia poblacional) y tal que $0 < p_1 < 1$.

Tras unas sencillas cuentas (Apéndice A, Nota 3), se tiene:

$$a = \frac{\frac{c(1-2p_0)}{N_{ca}} + \sqrt{\left(\frac{1-2p_0}{N_{ca}}\right)^2 + 4(1-p_0)p_0\left(\frac{1}{N_{ca}} + \frac{1}{N_{co}}\right)\left(\frac{1}{N_{ca}} + \frac{1}{N'_{sc}}\right)}}{2p_0\left(\frac{1}{N_{ca}} + \frac{1}{N'_{sc}}\right)},$$

siendo N'_{sc} el correspondiente valor que proporciona la potencia buscada para el nivel de significación deseado.

En resumen, si se denota $a_- = c\ a$ si $c = -1$ y $a_+ = c\ a$ si $c = 1$,para un nivel de significación α y una potencia β se tiene:

$$\mathbb{P}(\textit{Rechazar hipotésis nula} \mid p_1 = p_0 + a\ p_0 \ \ \textit{con} \ \ a \geq \ a_+ \ \ \textit{ó} \ \ -a \geq a_-) \geq \beta.$$

Éstas son interesantes aplicaciones del parámetro que sirven para estudiar la capacidad discriminatoria de nuestro contraste, pero se plantea un problema, ¿qué valor de N'_{sc} necesitamos para significación y potencia dadas?

Para responder a esta pregunta, se simularán 5000 tablas, para cada tabla se harán 5000 permutaciones, y todo esto para 8 haplogrupos, tres desviaciones de la hipotésis nula, 4 *odds* caso-control y 6 valores del número de controles; para obtener curvas de potencia, que serán estandarizadas, para posteriormente realizar regresión no paramétrica tomando como variable explicativa la potencia y como variable respuesta N'_{sc}.

El método aquí considerado será el de Nadaraya-Watson con selector de ventana validación cruzada, función núcleo $N(0,1)$, y cuyo funcionamiento se puede consultar en el Apéndice B. Se podrían considerar otros métodos como el lineal local u otros selectores, como el plug-in [4].

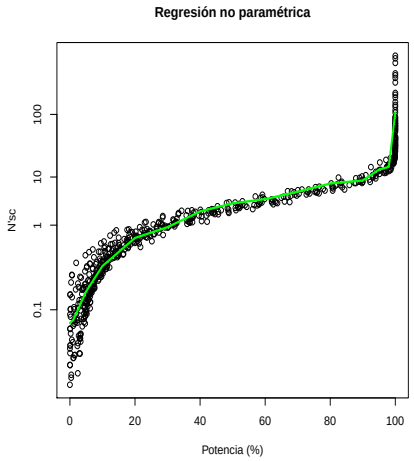

Figura 4.12: Regresión paramétrica tomando como variable explicativa la potencia, y variable dependiente N'_{sc}

A continuación mostraremos los valores N'_{sc} tabulados para distintos valores de significación y potencia.

Significación \ Potencia	95 %	90 %	80 %	70 %	60 %	50 %
0.1 %	23.04	19.57	16.20	14.87	13.82	9.32
0.5 %	17.94	13.67	14.12	10.44	7.07	6.82
1 %	14.66	13.66	10.05	8.88	7.63	6.37
5 %	14.31	9.75	8.74	6.53	5.03	4.33
10 %	9.26	8.29	6.04	5.53	3.52	2.64

Esto permitiría, por ejemplo, graficar el incremento de la frecuencia en casos respecto a la de controles en función del número de casos y controles, dadas significación y potencia (en este casos se ha tomado un nivel de significación $\alpha = 5\,\%$ y una potencia $\beta = 80\,\%$); y así conocer las limitaciones del contraste de antemano.

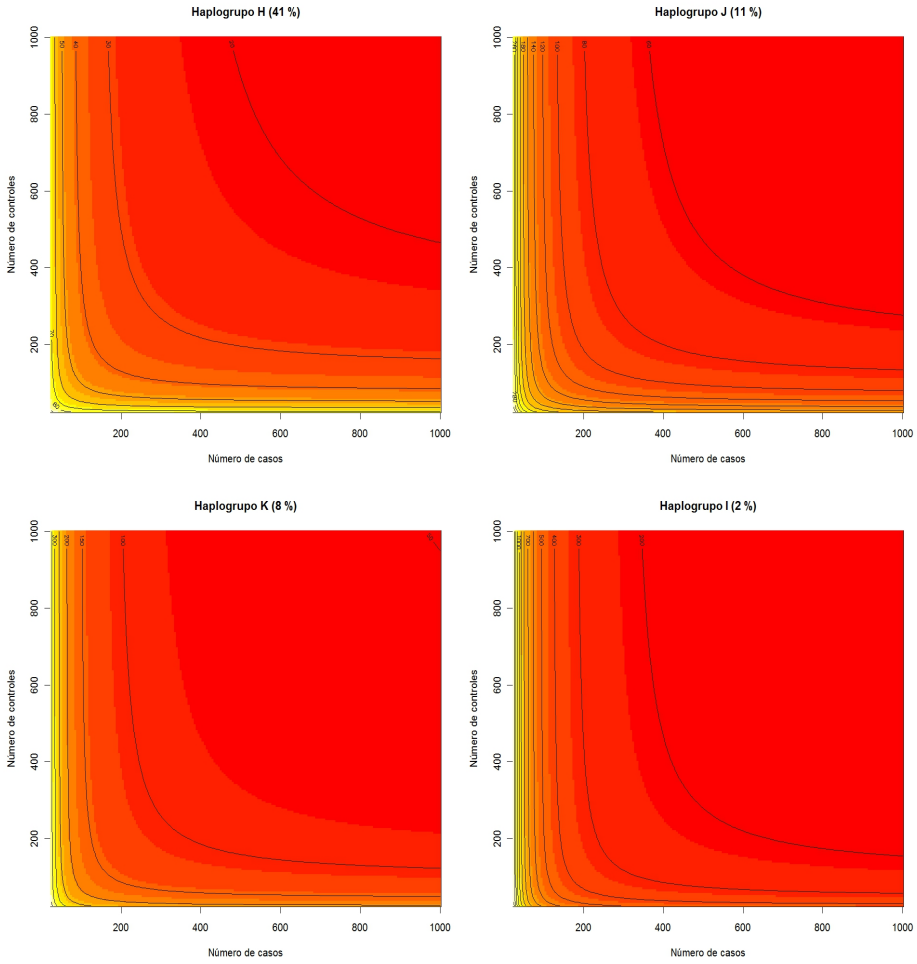

Figura 4.13: Curvas de nivel del incremento detectado con una potencia del 80 % en función del número de casos y controles

En esta Figura se aprecia un hecho que pasa desapercibido en el resto de gráficas, y es que la relación entre los números de casos y controles con respecto a la desviación detectable fijada una potencia no es lineal. Si se quiere pasar de detectar una desviación a detectar otra con una cierta potencia, los números de casos y controles no dependerán tan sólo de la diferencia entre ambas desviaciones, si no que tambíen dependerá de las magnitiud de las desviaciones en si.

Por ejemplo, en el caso del haplogrupo H, pasando de considerar 190 casos-controles a considerar 300 casos y controles (una diferencia de 110 casos-controles) , se pasaría a detectar un incremento del 40 % a detectar un incremento del 30 %; mientras que para detectar un incremento del 20 % se necesitarían 600 casos-controles (300 más que para detectar in incrmento del 30 %).

También se podría generar una tabla de la siguiente forma, que serían de interés para un inves-

tigador. En ella se muestran las desviaciones detectables para distintas frecuencias en contrales y tamaños muestrales:

N° casos-controles	100		200		300		400		500	
Frec. poblacional	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑
0.5	40.0 %	40.0 %	28.9 %	28.9 %	23.7 %	23.7 %	20.6 %	20.6 %	18.5 %	18.5 %
0.4	47.1 %	51.1 %	34.4 %	36.5 %	28.4 %	29.8 %	24.7 %	25.8 %	22.2 %	23.13 %
0.3	56.1 %	66.8 %	41.5 %	47.0 %	34.5 %	38.2 %	30.1 %	33.0 %	27.1 %	29.4 %
0.2	69.0 %	93.1 %	51.9 %	64.4 %	43.5 %	52.0 %	38.2 %	44.6 %	34.5 %	39.7 %
0.1	92.3 %	156.6 %	71.6 %	105.1 %	60.9 %	83.5 %	54.0 %	71.1 %	49.1 %	62.9 %
0.05	–	261.4 %	93.9 %	169.3 %	81.3 %	132.2 %	72.9 %	111.4 %	66.8 %	97.7 %

Esta tabla se ha construido para una potencia del 80 % y un nivel de significación $\alpha = 5\,\%$.

Notar que para el caso de una frecuencia poblacional de 0.05 y un estudio con 100 casos y 100 controles el decrecimiento mínimo no se ha mostrado, ya que en esta situación el decrecimiento tendría que ser de más del 100 %; un decrecimiento menor no sería detectado con la potencia deseada.

Capítulo 5

Aplicación a datos reales

En este capítulo se analizarán datos referentes a la meningococcemia [16], producida por la bacteria *Neisseria meningitidis*, que es frágil pero altamente peligrosa. Esta bacteria se encuentra frecuentemente en la nariz y la garganta de los individuos sanos y la meningococcemia ocurre cuando este organismo invade el torrente sanguíneo.

Por lo general, su sintomatología consiste en el sangrado y muerte de las áreas afectadas. Las tasa de mortalidad de los pacientes con meningococcemia puede llegar hasta el $40\,\% - 50\,\%$, y los que sobreviven sueler tener secuelas graves como amputaciones o necesidad de injertos de piel.

Esta enfermedad es una de las infecciones mortales más importantes en los países industrializados, y cuya incidencia es mayor en las zonas con menos recursos y en los niños.

Se cree que factores genéticos influyen en el hecho de tener la *Neisseria meningitidis*.

Para contrastar la posible influencia de determinados haplogrupos en la posibilidad de padecer una enfermedad, se estudiarán de 268 casos recogidos en 41 hospitales españoles, y como controles se disponen de dos grupos distintos que serán denotados por CG1 y CG2, con 917 y 216 controles respectivamente. La primera muestra de control ha sido utilizada como muestra piloto para el contraste, y la segunda se ha utilizado para analizar si los resultados obtenidos con la primera muestra son realmente significativos.

Para el análisis de resultados se estudiará el comportamiento del contraste con el grupo de controles CG1. Se han generado 50000 tablas y 50000 permutaciones para cada tabla.

Para la estimación de la potencia se ha fijado un nivel de significación $\alpha = 5\,\%$.

En la siguiente tabla se muestran el p-valor Bootstrap y la potencia estimada (Apéndice D) para cada uno de los haplogrupos considerados y los tamaños muestrales antes citados:

Haplogrupo	Frec. CG1	Frec. Casos	Variación relativa (%)	p-valor nominal	p^{Boots}	Potencia (%)
H	0.41	0.41	↑ 0 %	0.321	0.5525	4.7 %
H1	0.20	0.20	↑ 0 %	0.826	0.526	4.5 %
H3	0.09	0.07	↓ 22.2 %	0.292	0.383	14.0 %
HV	0.53	0.47	↓ 11.3 %	0.116	0.206	38.8 %
I	0.02	0.02	↑ 0 %	0.831	0.572	3.7 %
J	0.08	0.06	↓ 25 %	0.275	0.370	16.3 %
J1	0.06	0.04	↓ 33.3 %	0.311	0.327	19.6 %
K	0.06	0.08	↑ 33.3 %	0.224	0.357	21.7 %
K1	0.05	0.07	↑ 40 %	0.219	0.336	24.0 %
R	0.93	0.87	↓ 6.4 %	0.005	0.044	83.0 %
R0	0.58	0.50	↓ 13.7 %	0.038	0.101	63.4 %
T	0.09	0.10	↑ 11.1 %	0.750	0.491	7.6 %
TJ	0.17	0.16	↓ 5.8 %	0.600	0.505	6.0 %
U	0.18	0.17	↓ 5.5 %	0.876	0.506	5.5 %
U5	0.06	0.04	↓ 33.3 %	0.408	0.327	19.6 %
V	0.04	0.06	↑ 50 %	0.093	0.310	27.8 %
W	0.01	0.01	↑ 0 %	0.593	0.599	2.7 %

Los p-valores obtenidos mediante la aproximación asintótica (el p-valor nominal) y mediante Bootstrap son muy distintos. Cabría pensar que el p-valor Bootstrap sería mayor que el de la aproximación χ^2, pero si se analiza la metodología se conluye que esto no tiene porque ser así.

El p-valor asintótico ha sido obtenido para un tabla concreta de la que no se dispone de todos los datos (faltan las totales por haplogrupos), mientras que el método Bootstrap lo que analiza es una población, y no un caso en particular.

En esta tabla se observa que las frecuencias de los haplogrupos no son complementarias. Esto se debe a la estructura jerárquica que se mencionó en el Capítulo 1.

Esta tabla se ha construido *a posteriori*, tomando las frecuencias muestrales en casos y controles del estudio como las frecuencias “reales". Para un investigador sería de interés conocer la potencia con la que se rechazarán ciertas desviaciones de la hipotésis nula previamente a la realización del estudio, fijado número de casos y controles.

A continuación se mostrarán los valores de potencia en cada una de las situaciones consideradas.

Haplogrupo	Frec. CG1	Variación relativa (%)				
		↑ 100 %	↑ 50 %	↑ 25 %	↓ 25 %	↓ 50 %
H	0.41	100 %	100 %	83.5 %	85.7 %	100 %
H1	0.20	100 %	91.0 %	41.0 %	43.7 %	98.1 %
H3	0.09	97.0 %	54.4 %	18.8 %	18.1 %	67.8 %
HV	0.53	–	100 %	96.9 %	96.9 %	100 %
I	0.02	42.8 %	16.1 %	7.6 %	3.7 %	10.3 %
J	0.08	95.1 %	49.4 %	17.6 %	15.9 %	61.9 %
J1	0.06	86.7 %	38.7 %	14.1 %	11.9 %	47.1 %
K	0.06	86.8 %	38.8 %	14.3 %	11.9 %	46.3 %
K1	0.05	79.3 %	33.2 %	12.6 %	9.5 %	37.1 %
R	0.93	–	–	–	100 %	100 %
R0	0.58	–	100 %	99.2 %	98.5 %	100 %
T	0.09	97.0 %	53.8 %	19.2 %	18.0 %	67.9 %
TJ	0.17	100 %	85.1 %	34.2 %	37.0 %	95.2 %
U	0.18	100 %	87.6 %	36.7 %	38.9 %	96.3 %
U5	0.06	87.0 %	38.4 %	14.4 %	11.2 %	46.5 %
V	0.04	70.5 %	28.0 %	11.1 %	7.9 %	27.3 %
W	0.01	23.9 %	10.4 %	5.8 %	1.6 %	1.5 %

Notar que hay casos en los que no se muestran valores de potencia, debido a que la frecuencia en casos sería mayor que 1.

Ahora se procederá a mostrar la funcionalidad de la medida N'_{sc}.

En un primer lugar se estudiará cuál es el número de controles mínimo necesario para detectar una diferencia entre las frecuencias de casos y controles, fijado el número de casos con la potencia y significación deseadas.

Se fijara un nivel de significación $\alpha = 0.05$, y una potencia del 80 %.

Como ya se dijo antes, para calcular el número de mínimo de controles existe una restricción sobre el número de casos, condición que no se verifica en ninguna de los haplogrupos tenidos en cuenta. En este caso se calculará el número mínimo de controles para el número mínimo de casos que verifique la condición.

Sólo se mostrarán aquellos haplogrupos cuyas frecuencias en casos y controles sean distintas. En caso contrario, se estaría bajo la hipotésis nula y la potencia deseada nunca podría ser alcanzada.

Haplogrupo	Frec. CG1	Frec. Casos	Variación relativa (%)	Nº casos	Nº controles
H3	0.09	0.07	$\downarrow 22.2\,\%$	1423	$4.5\ 10^6$
HV	0.53	0.47	$\downarrow 11.3\,\%$	605	$1.5\ 10^6$
J	0.08	0.06	$\downarrow 25\,\%$	1233	$3.0\ 10^6$
J1	0.06	0.04	$\downarrow 33.3\,\%$	840	$1.0\ 10^6$
K	0.06	0.08	$\uparrow 33.3\,\%$	1609	$2.3\ 10^6$
K1	0.05	0.07	$\downarrow 40\,\%$	1423	$2.6\ 10^6$
R	0.93	0.87	$\downarrow 6.4\,\%$	275	$1.0\ 10^5$
R0	0.58	0.50	$\downarrow 13.7\,\%$	342	$1.9\ 10^5$
T	0.09	0.10	$\uparrow 11.1\,\%$	7866	$4.3\ 10^{18}$
TJ	0.17	0.16	$\downarrow 5.8\,\%$	$1.1\ 10^4$	$3.2\ 10^8$
U	0.18	0.17	$\downarrow 5.5\,\%$	$1.2\ 10^3$	$1.8\ 10^8$
U5	0.06	0.04	$\downarrow 33.3\,\%$	840	$1.0\ 10^6$
V	0.04	0.06	$\uparrow 50\,\%$	1233	$1.5\ 10^6$

Observar que el número de controles necesarios en cada caso es altísimo, pero esto se debe a que el número de casos que se ha considerado es muy pequeño (se ha considerado como número de casos el menor número entero tal que el denominador que aparece en el cálculo de los controles no se anule).

En el caso del haplogrupo R, el número de casos necesarios para calcular el número mínimo de controles no es lo suficientemente grande, pero en la primera tabla de este capítulo en la que estimaba la potencia se observa que se alcanza una potencia del 83.0 %. Cabría pensar que el número de casos es suficiente, y que el numero mínimo de controles necesario será menor que el número de controles disponibles.

Ante esta siuación cabe preguntarse qué es lo que ocurre. No hay que olvidar que a la hora de explicar la relación entre la potencia y el parámetro N'_{sc} a través del estimador no paramétrico se está obviando la variabilidad de la muestra. Por eso, sería de interés construir intervalos de confianza para el valor N'_{sc} condicionado a un valor potencia, y así poder dar intervalos de confianza para el número mínimo de controles. Éste sería el siguiente paso a seguir en la investigación de este tema.

La tabla anterior es una tabla *a posteriori*, ya que las desviaciones de la hipótesis nula se suponen conocidas.

Ahora se estudiará como varía el número de controles para diversas desviaciones de la hipóteis nula.

De nuevo, habrá haplogrupos que, bajo ciertas condiciones, no verifiquen la restricción del número de casos, por lo que se ha procedido a calcular el número mínimo de controles requeridos para el número mínimo de casos que que verifique la condición. En los casos en los que no se cumple la condición se ha puesto en número mínimo de casos entre paréntesis.

Haplogrupo	Frec. CG1	Variación de la frec. en casos				
		$\uparrow 100\,\%$	$\uparrow 50\,\%$	$\uparrow 25\,\%$	$\downarrow 25\,\%$	$\downarrow 50\,\%$
H	0.41	13	62	897	594	58
H1	0.20	44	444	$7.3\ 10^5$(656)	$9.5\ 10^5$ (446)	198
H3	0.09	218	$1.7\ 10^5$ (505)	$8.7\ 10^6$ (1724)	$4.7\ 10^6$ (1087)	1149
HV	0.53	–	34	213	224	35
I	0.02	$3.7\ 10^5$ (840)	$6.6\ 10^6$ (2544)	$1.1\ 10^8$ (8522)	$5.3\ 10^7$ (5166)	$2.0\ 10^6$(866)
J	0.08	319	$1.4\ 10^6$ (577)	$6.3\ 10^6$ (1967)	$3.0\ 10^6$ (1233)	1851
J1	0.06	3157	$6.6\ 10^5$ (796)	$3.5\ 10^7$ (2695)	$5.5\ 10^6$ (1670)	$3.8\ 10^5$ (283)
K	0.06	3157	$6.6\ 10^5$ (796)	$3.5\ 10^7$ (2695)	$5.5\ 10^6$ (1670)	$3.8\ 10^5$ (283)
K1	0.05	$1.4\ 10^3$ (315)	$7.4\ 10^5$ (971)	$1.7\ 10^7$ (3278)	$8.9\ 10^7$ (2019)	$1.61\ 10^6$ (341)
R	0.93	–	–	–	13	3
R0	0.58	–	–	147	164	28
T	0.09	218	$1.7\ 10^5$ (505)	$8.7\ 10^6$ (1724)	$4.7\ 10^6$ (1087)	1149
TJ	0.17	58	1198	$2.0\ 10^6$ (810)	$4.7\ 10^6$ (539)	1149
U	0.18	52	772	$1.2\ 10^6$ (753)	$3.2\ 10^5$ (505)	238
U5	0.06	3157	$6.6\ 10^5$ (796)	$3.5\ 10^7$ (2695)	$5.5\ 10^6$ (1670)	$3.8\ 10^5$ (283)
V	0.04	$8.8\ 10^4$ (403)	$1.5\ 10^6$ (1233)	$2.7\ 10^4$ (4152)	$1.2\ 10^7$ (2544)	$4.8\ 10^5$ (429)
W	0.01	$1.5\ 10^6$ (1714)	$2.7\ 10^7$ (5166)	$4.7\ 10^8$ ($1.7\ 10^4$)	$2.1\ 10^8$ (10^4)	$8.1\ 10^6$ (1740)

Esta tabla puede ser de gran utilidad para un investigador, ya que le permitirá decidir si es productivo, o no, aumentar el número de controles para alcanzar una determinada potencia.

Por último, se analizarán cuales son las menores desviaciones de la hipotésis nula que podrán ser detectadas con una potencia del 80 %. Se denotará por $freq_{ca}^{-}$ a la frecuencia máxima detectable menor que la frecuencia en los controles, y $freq_{ca}^{+}$ a la mínima frecuencia detectable mayor que la frecuencia en los controles. En la página suigiente se muestrta la tabla con los resultados obtenidos.

También se puede graficar las desviaciones mínimas de la hipotésis nula detectables con una potencia del 80 %, en función de la frecuencia en controles para el tamaño de estudio considerado.

Haplogrupo	Frec. CG1	$freq_{ca}^{-}$	$freq_{ca}^{+}$
H	0.41	0.313	0.512
H1	0.20	0.128	0.290
H3	0.09	0.043	0.162
HV	0.53	0.428	0.629
I	0.02	0.003	0.067
J	0.08	0.036	0.149
J1	0.06	0.023	0.123
K	0.06	0.023	0.123
K1	0.05	0.017	0.110
R	0.93	0.863	0.969
R0	0.58	0.477	0.677
T	0.09	0.043	0.162
TJ	0.17	0.103	0.257
U	0.18	0.111	0.268
U5	0.06	0.023	0.123
V	0.04	0.012	0.096
W	0.01	10^{-4}	0.05

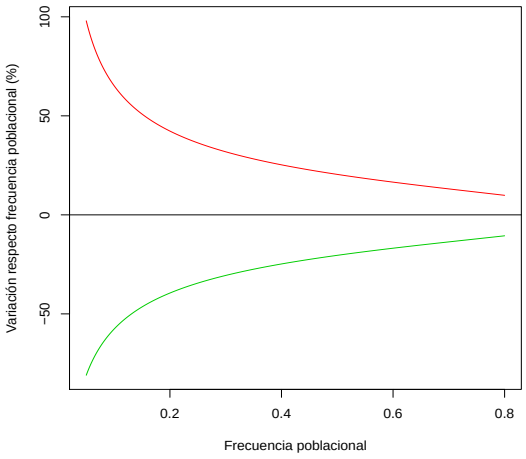

Figura 5.1: Relación entre la frecuencia poblacional y las variaciones detectables. El trazo rojo denota el incremento mínimo detectable y el verde el decrecimiento mínimo detectable. La linea negra, denota la hipótesis nula

Capítulo 6

Conclusión

A lo largo de este proyecto, se han analizado principalmente dos problemas enmarcados en el análisis del genoma mitocondrial y los estudios tipo caso-control:

- Calibración del estadístico χ^2.
- Estimación de potencia estadística.

El primero de estos problemas ha sido abordado con la calibración mediante permutación.

Se ha visto que, a medida que la frecuencia poblacional de un haplogrupo disminuye, la diferencia entre los p-valores de la aproximación asintótica y la calibración por permutación aumenta. Esto es debido a que se producirá un incumplimiento reiterado de la regla de Cochran y por lo tanto el p-valor nominal será artificialmente bajo.

Posteriormente, se ha realizado un análisis de la potencia del contraste empleado. Se ha visto que la potencia depende de las frecuencias en casos y controles, y de los tamaños muestrales para cada población.

Se ha definido un parámetro que tiene en cuenta todas estas variables y a partir del cual se pueden extraer distintas conclusiones de interés sobre la capacidad de un estudio tipo caso-control.

Apéndice A

Cálculos

En esta sección se mostrarán lo cálculos realizados para la consecución de alguna de las conclusiones de este proyecto.

Nota 1

Sea $X = (\xi_1, \ldots, \xi_k) \sim Multinomial(n, (p_1, \ldots, p_k))$, donde $\sum_{i=1}^{k} p_i = 1$. Se pretende estimar por máxima verosimilitud el vector $(p_1, \ldots, p_k)$.

Se tiene que

$$\mathbb{P}(\xi_1 = n_1, \ldots, \xi_k = n_k) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i} \quad con \quad \sum_{i=1}^{k} n_i = n,$$

y se construye la función de verosimilitud de la siguiente forma:

$$l(p_1, \ldots, p_k) = \sum_{i=1}^{k} n_i \; log \; p_i \quad sujeto \; a \quad \sum_{i=1}^{k} p_i = 1.$$

Para maximizar esta función, sujeto a la restricción, utilizaremos el método de los multiplicadores de Lagrange, con los cual debemos de definir la función:

$$L(p_1, \ldots, p_k, \lambda) = l(p_1, \ldots, p_k) + \lambda(1 - \sum_{i=1}^{k} p_i) = \sum_{i=1}^{k} n_i \; log \; p_i + \lambda(1 - \sum_{i=1}^{k} p_i)$$

$$\frac{\partial L}{\partial p_i}(p_1, \ldots, p_n, \lambda) = \frac{n_i}{p_i} - \lambda = 0 \Leftrightarrow p_i = \frac{n_i}{\lambda} \quad i = 1, \ldots, k$$

y puesto que $\sum_{i=1}^{k} p_i = 1$, se tiene $\lambda = n$ con lo cual resulta:

$$\widehat{p_i} = \frac{n_i}{n} \quad i = 1 \ldots, k.$$

es el estimador de máxima verosimilitud, ya que:

$$\frac{\partial^2 L}{\partial p_i^2}(\widehat{p}_1, \ldots, \widehat{p}_k, n) = -\frac{n_i}{\widehat{p}_i} < 0.$$

Nota 2

En la sección 4 de este proyecto se definió un parámetro que unificaba las curvas de potencia, N_{sc}, o de forma más general N'_{sc}. En esta nota daremos la interpretación del segundo parámetro (el primero es un caso particular).

Dadas dos variables aleatorias $X \sim Bi(p_0, N_{co})$ e $Y \sim Bi(p_1, N_{ca})$ independientes, se dispondrá de una tabla de contingencia de la siguiente forma:

	A_1	A_2	Total filas
Población 1	N_{11}	N_{12}	N_{co}
Población 2	N_{21}	N_{22}	N_{ca}
Total columnas	$N_{\cdot 1}$	$N_{\cdot 2}$	n

Los estimadores de máxima verosimilitud de p_0 y p_1 son $\widehat{p_0} = \frac{N_{11}}{N_{co}}$ y $\widehat{p_1} = \frac{N_{21}}{N_{ca}}$ respectivamente.

Se define el coeficiente de variación de Pearson [2] para $(\widehat{p_1} - \widehat{p_0})$ como

$$CV = \frac{\sigma(\widehat{p_1} - \widehat{p_0})}{\mathbb{E}(\widehat{p_1} - \widehat{p_0})},$$

donde $\sigma(\widehat{p_1} - \widehat{p_1})$ denota la desviación típica de $\widehat{p_1} - \widehat{p_0}$.

Ahora se realizarán unos sencillos cálculos teniendo en cuenta que que $N_{11} \sim Bi(N_{co}, p_0)$ y $N_{21} \sim Bi(N_{ca}, p_1)$ y que son variables independientes:

$$\mathbb{E}(\widehat{p_1} - \widehat{p_0}) = \mathbb{E}(\widehat{p_1}) - \mathbb{E}(\widehat{p_0}) = \mathbb{E}\left(\frac{N_{21}}{N_{ca}}\right) - \mathbb{E}\left(\frac{N_{11}}{N_{co}}\right) = \frac{\mathbb{E}(N_{21})}{N_{ca}} - \frac{\mathbb{E}(N_{11})}{N_{co}} =$$

$$= \frac{p_1 N_{ca}}{N_{ca}} - \frac{p_0 N_{co}}{N_{co}} = p_1 - p_0$$

$$Var(\widehat{p_1} - \widehat{p_0}) = Var(\widehat{p_1}) + Var(\widehat{p_0}) = Var\left(\frac{N_{21}}{N_{ca}}\right) + Var\left(\frac{N_{11}}{N_{co}}\right) = \frac{Var(N_{21})}{N_{ca}^2} + \frac{Var(N_{11})}{N_{co}^2} =$$

$$= \frac{p_1(1-p_1)N_{ca}}{N_{ca}^2} + \frac{p_0(1-p_0)N_{co}}{N_{co}^2} = \frac{p_1(1-p_1)}{N_{ca}} + \frac{p_0(1-p_0)}{N_{co}}.$$

Entonces resulta:

$$CV = \frac{\sqrt{\frac{p_1(1-p_1)}{N_{ca}} + \frac{p_0(1-p_0)}{N_{co}}}}{p_1 - p_0}.$$

Podría pensarse en definir N'_{sc} de esa forma, pero ésta no estaría definida si $p_0 = p_1$, por lo que, en lugar de utilizar CV, debería utilizarse $1/CV$; y para que esta medida sea independiente del signo de $p_1 - p_0$, se considerará

$$N'_{sc} = \frac{(p_1 - p_0)^2}{\frac{p_1(1-p_1)}{N_{ca}} + \frac{p_0(1-p_0)}{N_{co}}}.$$

Nota 3

A continuación se muestran los cálculos de la obtención de las desviaciones mínimas detectadas dados número de controles, número de casos y frecuencia del haplogrupo en la población. Sea N'_{sc} el correspondiente para un nivel de significación y potencia, y $p_1 = p_0 + c\, a p_0$ con $a > 0$ y $c = -1$ *ó* 1:

$$N'_{sc} = \frac{(p_1 - p_0)^2}{\frac{p_1(1-p_1)}{N_{ca}} + \frac{p_0(1-p_0)}{N_{co}}} = \frac{((p_0 + c\ a p_0) - p_0)^2}{\frac{(p_0 + c\ a\ p_0)(1-(p_0+c\ a\ p_0))}{N_{ca}} + \frac{p_0(1-p_0)}{N_{co}}} = \frac{(c\ a p_0)^2}{\frac{p_0(1+c\ a)(1-(1+c\ a)p_0)}{N_{ca}} + \frac{p_0(1-p_0)}{N_{co}}} =$$

$$= \frac{a^2 p_0}{\frac{(1+c\ a)(1-(1+c\ a_+)p_0)}{N_{ca}} + \frac{(1-p_0)}{N_{co}}} \Rightarrow \frac{(1 + c\ a)(1 - (1 + c\ a_+)p_0)}{N_{ca}} + \frac{1 - p_0}{N_{co}} - \frac{a^2 p_0}{N'_{sc}} =$$

$$= (1 - p_0)\left(\frac{1}{N_{ca}} + \frac{1}{N_{co}}\right) + c\ a\left(\frac{1 - 2p_0}{N_{ca}}\right) - a^2 p_0\left(\frac{1}{N_{ca}} + \frac{1}{N'_{sc}}\right) = 0 \Rightarrow$$

$$\Rightarrow a = \frac{\frac{c(1-2p_0)}{N_{ca}} \pm \sqrt{\left(\frac{c(1-2p_0)}{N_{ca}}\right)^2 + 4(1 - p_0)\left(\frac{1}{N_{ca}} + \frac{1}{N_{co}}\right) p_0 \left(\frac{1}{N_{ca}} + \frac{1}{N'_{sc}}\right)}}{2p_0\left(\frac{1}{N_{ca}} + \frac{1}{N'_{sc}}\right)},$$

y dado que $a > 0$ se tiene:

$$a = \frac{\frac{c(1-2p_0)}{N_{ca}} + \sqrt{\left(\frac{1-2p_0}{N_{ca}}\right)^2 + 4(1 - p_0)\left(\frac{1}{N_{ca}} + \frac{1}{N_{co}}\right) p_0 \left(\frac{1}{N_{ca}} + \frac{1}{N'_{sc}}\right)}}{2p_0\left(\frac{1}{N_{ca}} + \frac{1}{N'_{sc}}\right)},$$

Nota 4

Sea p_0 la probabilidad del haplogrupo al que se quiere aplicar el contraste (esta probabilidad se supone conocida y se facilitará por el usuario). En este proyecto se han expresado las desviaciones respecto a la hipotésis nula como variaciones relativas a la frecuencia en la población, de la siguiente forma: $p_1 = p_0 + ap_0$, siendo p_1 la frecuencia del haplogrupo en los casos y a el incremento de la proporción del número de casos con ese haplogrupo respecto a la proporción del número de controles con ese haplogrupo.

Es de interés expresar está proporción, a, en términos de la *odds ratio*, una medida más usual.

$$OR = \frac{p_0/(1 - p_0)}{p_1/(1 - p_1)}$$

Para solucionar este problema se utilizará la expresión de p_1.

$$OR = \frac{p_0/(1-p_0)}{p_1/(1-p_1)} = \frac{p_0/(1-p_0)}{(1+a)p_0/(1-(1+a)p_0)} = \frac{1/(1-p_0)}{(1+a)/(1-(1+a)p_0)} =$$
$$= \frac{1-(1+a)p_0}{(1+a)(1-p_0)} = \frac{1-p_0-ap_0}{1-p_0+a-ap_0} = 1 - \frac{a}{1-p_0+a-ap_0} \Rightarrow$$

$$\Rightarrow (1-OR)(1-p_0+a-ap_0) = a \Rightarrow (1-OR)(1-p_0) = a(1-(1-OR)(1-p_0)) \Rightarrow a = \frac{(1-OR)(1-p_0)}{1-(1-OR)(1-p_0)}.$$

La razón por la que no se ha empleado la *odds ratio* en las simulaciones de este proyecto es que una misma *odd ratio* no significa lo mismo para dos haplogrupos con distinta frecuencia poblacional. Por ejemplo, para el haplogrupo H ($p_0 = 0.41$) una *odds ratio* de 0.5 equivaldría a un valor $a = 0.41$; mientras que para el haplogrupo I ($p_0 = 0.2$) equivaldría a $a = 0.96$.

En este proyecto, se pretendía mostrar la diferencia entre curvas de potencia para distintos haplogrupos en situaciones equivalentes.

Apéndice B

Contrastes no paramétricos

En el Capítulo 4 se han planteado tres contrastes no paramétricos: el de signos, para ver si existían diferencia significativas entres los p-valores obtenidos mediante dos métodos de calibración (para una misma tabla); el contraste de Kruskal-Wallis, para comprobar si el haplogrupo considerado era significativo a la hora de explicar la diferencia entre los p-valores de cada calibración; y el el Wilcoxon-Mann-Whitney, para hacer la hacer un contraste de homogeneidad sobre la diferencia de p-valores para cada par de haplogrupos

En esta sección se recordará el funcionamiento de estos dos contrastes.

B.1. Contrastes de signos

Dada una muestra aleatoria simple $\{(X_i, Y_i)\}_{i=1}^n$ de una variable bidimensional (X, Y), se desea contrastar si X e Y estan distribuidos de la misma forma.

$$H_0 : F_X(z) = F_Y(z)$$

versus

$$H_1 : F_X(z) \geq F_Y(z) \;\; (Y \textit{ es estocásticamente mayor que } X)$$

Para contrastar este hecho en muestras pareadas, Wilcoxon (1945) propuso el siguiente estadístico [15]:

$$D = \sum_{i=1}^{n} \mathbb{I}_{\{Y_i > X_i\}}.$$

Si Y es estocásticamente mayor que X, el estadístico D tomará un valor grande; mientras que tomará un valor pequeño si X es estocásticamente mayor que Y, y valores intermedios (en torno a $n/2$) si las dos poblaciones son homogéneas.

Bajo la hipotésis nula de igualdad de distribuciones, se tiene que $D \sim Bi(n, 0.5)$.

B.2. Contraste Wilcoxon-Mann-Whitney

Dadas dos variables aleatorias X e Y con distribuciones F_X y F_Y, se desea contrastar la homogeneidad de estas dos poblaciones, para lo que se dispondrá de dos muestras independientes $\{X_i\}_{i=1}^n$ e $\{Y_j\}_{j=1}^m$ de X e Y respectivamente. Se pretende contrastar

$$H_0 : \ F_X(z) = F_Y(z)$$

versus

$$H_1 : \ F_X(z) \leq F_Y(z) \ \ (Y \textit{ es estocásticamente mayor que } X).$$

Wilcoxon (1945) y Mann y Whitney (1947) paralelamente, propusieron un método para contrastar la homogeneidad de dos poblaciones de este tipo.

Para la realización de este contraste se han planteado dos estadísticos que son equivalentes: el estadístico de Mann-Whitney y el de Wilcoxon [15].

El estadístico propuesto por Mann y Whitney se basa en la comparación de las dos muestras a través del siguiente estadístico [15]:

$$D_{MW} = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}_{\{Y_j > X_i\}}.$$

El estadístico cuenta el número de pares (X_i, Y_j) tal que $Y_j > X_i$ de entre los nm, pares posibles.

Cuando Y sea estocásticamente mayor que X, el estadístico tomará un valor alto. Por contra, si X es estocásticamente mayor que Y, el estadístico tomará un valor pequeño. Bajo la hipotésis nula el D_{MW} tomará un valor intermedio.

La distribución de D_{MW} bajo la hipotésis nula puede ser calculada. Además tenemos la convergencia asintótica:

$$\frac{D_{MW} - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \xrightarrow[n,m\to\infty]{} N(0,1).$$

El estadístico propuesto por Wilcoxon es el siguiente [15]:

$$D_W = \sum_{j=1}^{m} r(Y_j),$$

donde

$$r(Y_j) = \sharp\{X_i \ / \ X_i \leq Y_j\} + \sharp\{Y_l \ / \ Y_l \leq Y_j\}.$$

Se verifica:

$$D_W = D_{MW} + \frac{m(m+1)}{2},$$

por lo que los estadístico son equivalentes (aunque D_W es más facilmente calculable que D_{MW}) y por ello, el contraste que se basa en este estadístico se denomina el contraste de Wilcoxon-Mann-Whitney [15].

B.3. Contraste Kruskal-Wallis

Este contraste, propuesto por Kruskal y Wallis en 1952, es la extensión del contraste Wilcoxon-Mann-Whitney para 3 o más poblaciones.

Dadas k muestras $\{X_{ij}\}_{i=1}^{n_i}$ con $i = 1, \ldots, m$, se define el estadístico de Kruskal-Wallis [15]:

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{m} \left(\frac{r_i}{n_i} - \frac{n-1}{2} \right)^2$$

siendo $n = \sum_{i=1}^{m} n_i$ y $r_i = \sum_{j=1}^{n_i} r(X_{ij})$.

Bajo la hipotésis nula y para n_i no demasiado pequeños (n_i>5) se tiene la convergencia asintótica del estadístico a una χ^2_{k-1}.

Apéndice C

Regresión no paramétrica

Dadas dos variables, X e Y, se pretende estudiar la forma en la que la variable dependiente, Y, se puede explicar a partir de la variable X mediante una función desconocida

$$y = m(x).$$

Para la estimación de esta función dispondremos de una muestra aleatoria simple $\{(x_i,\ y_i)\}_{i=1}^{n}$ y se tendrá:

$$y_i = m(x_i) + \epsilon_i,\ i = 1, \ldots, n,$$

siendo

$$m(x) = \mathbb{E}(Y \mid X = x),$$

y donde ϵ_i representa la variabilidad de y_i respecto a x_i.

Esta función puede ser estimada de forma paramétrica, fijando un modelo, pero este modelo puede ser restrictivo. Los modelos de regresión no paramétrica serán más flexibles.

Si X e Y son dos variables aleatorias continuas, con densidad conjunta $f(x, y)$, se define la esperanza condicional de Y dado $X = x$ como:

$$\mathbb{E}(Y \mid X = x) = \int y\ f(y|x)\ dy = \frac{\int yf(x,y)\ dy}{f_X(x)},$$

donde

$$f(y \mid x) = \frac{f(x,y)}{f_X(x)}$$

es la densidad condicional de y dado $X = x$, y f_X es la densidad marginal de X.

Se procederá estimando de forma no paramétrica $f(x, y)$ y $f_X(x)$ mediante el estimadores tipo

núcleo correspondientes propuestos por Parzen (1962) y Rosenblatt (1956) [4]:

$$\widehat{f}_{n,K}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

$$\widehat{f}_{n,K}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x - X_i)K_{h_2}(y - Y_i),$$

siendo h, h_1, h_2 los paramétros de suavizados y $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$, siendo K una densidad simétrica con una única moda en 0.

El estimador del numerador resulta:

$$\int y\widehat{f}_{n,K}(x,y)\, dy = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x - X_i)Y_i,$$

con lo cual, se tiene el siguiente estimador para la función de regresión propuesto por Nadaraya y Watson (1964) [4]:

$$\widehat{m}_{n,K}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x - X_i)Y_i}{\frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j)}.$$

(Notar que por comodidad tomaremos $h_1 = h_2 = h$).

El efecto de la elección del núcleo será poco relevante, pero no así la de la ventana.

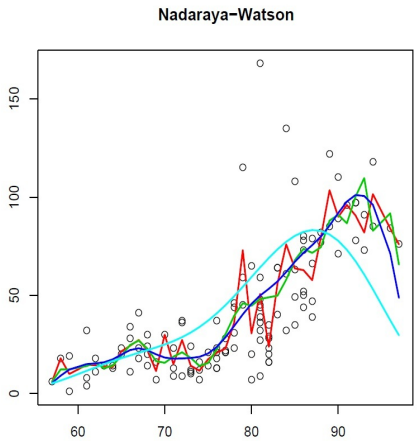

Figura C.1: Efecto de la ventana. La curva roja es la estimación de la función de regresión para $h = 0.1$, la verde para $h = 1$, la azul oscuro para $h =$ y la azul claro para $h = 6$

En esta gráfica se aprecia que para un paramétro de suavizado se tiene un sobreajuste de los datos, mientras que con ventanas grandes se produce una sobresuavización.

Para subsanar el problema de la ventana se debe elegir algún criterio de selección de ventana como puede ser el de validación cruzada [4]. Para estudiar la bondad de nuestro modelo se podría plantear el error medio:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_{n,K}(X_i))^2.$$

El problema de esta medida es que puede resultar demasiado optimista, ya que utiliza la muestra para construir el estimador y para calcular el error cometido.

Para evaluar el error de predicción sería preferible promediar los errores resultantes de evaluar el modelo construido sin el dato i-ésimo evaluado en ese mismo dato. La función de validación cruzada se define como:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_{-i,K}(X_i))^2,$$

donde $\widehat{m}_{-i,K}$ es el estimador de Nadaraya-Watson construido sin considerar el par (X_i, Y_i).

El criterio de elección de la ventana consistirá en elegir h_{CV} tal que:

$$h_{CV} = arg\ min(CV(h)).$$

Apéndice D

Metodología Monte Carlo y Bootstrap

En este Apéndice se explicará brevemente como se ha procedido a la hora de simular datos bajo distintas hipotésis, y como se ha trabajado con datos reales.

D.1. Monte Carlo

En una primera parte del trabajo, se han simulado curvas de potencia mediante Monte Carlo [7]. Se ha supuesto que, dado el número de controles y casos (denotados por N_{co} y N_{ca} respectivamente), la frecuencias en controles (p_0) es conocida [18] y la frecuencia en casos (p_1) es fijada según la hipotésis que se desee contrastar. Se procederá simulando tablas de la forma:

Clase / Población	Haplogrupo i	No haplogrupo i	Total filas
Controles	X	$N_1 - X$	N_{co}
Casos	Y	$N_2 - Y$	N_{ca}
Total columnas	$X + Y$	$N_1 + N_2 - (X + Y)$	$N_{co} + N_{ca}$

donde $X \sim Bi(p_0, N_{co})$ e $Y \sim Bi(p_1, N_{ca})$.

Para simular un valor de potencia para un nivel de significación α se procederá de la siguiente forma:

1. Se generan B tablas, $T_1, \ldots, T_B$.
2. Para cada tabla, T_b, se calcula el estadístico χ^2, que se denotará por χ_b.
3. Mediante el calibrado correspondiente (permutación o aproximación asintótica) se obtiene el p-valor para χ_b, denotado por p_b.

4. Se toma como estimación de la potencia $\widehat{\beta} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}_{\{p_b<\alpha\}}$.

D.2. Bootstrap paramétrico

A continuación se resume el modo de proceder cuando se disponen de una tabla de datos [7]:

Población \ Clase	Haplogrupo i	No haplogrupo i	Total filas
Controles 1	N_{11}	N_{12}	N_{co}
Casos 2	N_{21}	N_{22}	N_{ca}
Total columnas	$N{\cdot}1$	$N_{\cdot 2}$	n

se tiene que $N_{11} \sim Bi(N_{co}, p_0)$ y $N_{21} \sim Bi(N_{ca}, p_1)$, donde p_0 y p_1 son las frecuencias en controles y casos respectivamente y que son desconocidas.

Estas frecuencias se estimarán de forma natural, $\widehat{p}_0 = \frac{N_{11}}{N_{co}}$ y $\widehat{p}_1 = \frac{N_{21}}{N_{ca}}$.

Entonces se procederá simulando tablas de contingencia de la siguiente forma:

Población \ Clase	Haplogrupo i	No haplogrupo i	Total filas
Controles	X	$N_1 - X$	N_{co}
Casos	Y	$N_2 - Y$	N_{ca}
Total columnas	$X+Y$	$N_1 + N_2 - (X+Y)$	$N_{co} + N_{ca}$

donde $X \sim Bi(\widehat{p}_0, N_{co})$ e $Y \sim Bi(\widehat{p}_1, N_{ca})$.

Entonces se procederá de la siguiente forma:

1. Se generan B tablas, $T_1, \ldots, T_B$.

2. Para cada tabla, T_b, se calcula el estadístico χ^2 de Pearson, que se denotará por χ_b.

3. Se obtiene el p-valor para χ_b (o bien por la calibración a través de permutación o por la distribución asintótica).

4. Se toma como p-valor bootstrap, $p^{Boots} = \frac{1}{B}\sum_{i=1}^{B} p_i$.

También se puede dar un valor de potencia, procediendo por la metodología Monte Carlo una vez han sido estimadas las frecuencias de casos y controles.

Apéndice E

Código R

A la hora de calibrar el contraste mediante permutación se ha utilizado la función *chisq.test* implementada en R e incluida en el paquete *stats*:

```
chisq.test(x,simulate.p.value=TRUE,B=cte)
```

donde el parámetro *simulate.p.value* indica si se quiere realizar o no la calibración mediante permutación, y B el número de permutaciones que se desean.

Este algoritmo es de rápida ejecución, porque se utiliza código de C.

A continuación se muestran funciones de programación propia que utilizan la metodología de la que se ha hablado en el proyecto:

```
chisq.test.permuta_1<-function(x,b){
# x <- Tabla con los datos
# b <- Número de permutaciones
r<-rowSums(x)
c<-colSums(x)
expected=r%*%t(c)/sum(x)
if(any(expected==0)){stop("Hay valores esperados nulos")}
statistic<-sum((x-expected)^2/expected)
# Cálculo del estadístico para x
nr<-nrow(x)
nc<-ncol(x)
Ac<-rep(seq(1:nc),c)
#Generación del vector A
Ar<-rep(seq(1:nr),r)
```

```
    # Hacemos lo propio con las filas (vector auxiliar para contruir la tabla)
    table<-array(dim=c(2,2,B))
    for(i in 1:b){
    Ac<-sample(Ac,length(Ac),replace=F)
    # Permutación de Ac
    table[,,i]<-as.matrix(table(Ar,Ac),nr=2,ncol=2)}
    table<-table[,,which(apply(table,3,function(z){all(colSums(z)!=0)}))]
    chip<-apply(table,3,function(z){sum((z-expected)^2/expected)})
    # Cálculo del estadístico para las tablas permutadas
    p.value=mean(statistic<=chip)
    # Aproximación del p-valor de x
    names(statistic)<-"X-squared"
    names(p.value)<-"p-value"
    structure(list(statistic = statistic, p.value = p.value))}
```

Notar que esta función realiza el contraste para tablas bidimensionales de cualquier tamaño.

Se ha planteado una función alternativa para este calibrado utilizando el hecho de que estamos trabajando con tablas 2×2.

```
chisq.test.permuta_2<-function(x,b){
# x <- Tabla con los datos
# b <- Número de permutaciones
r<-rowSums(x)
c<-colSums(x)
expected<-r%*%t(c)/sum(x)
if(any(expected==0)){stop("Hay valores esperados nulos")}
statistic<-sum((x-expected)^2/expected)
# Cálculo del estadístico para x
tablepr<-array(dim=c(2,2,b))
tablepr[1,1,]<-rhyper(b,c[1],c[2],r[1])
tablepr[1,2,]<-r[1]-tablapr[1,1,]
tablepr[2,1,]<-c[1]-tablapr[1,1,]
tablepr[2,2,]<-r[2]-tablapr[2,1,]
tablepr<-tablepr[,,which(apply(tablepr,3,function(z){all(colSums(z)!=0)}))]
# Generación de tablas permutadas
chip<-apply(tablepr,3,function(z){sum((z-expected)^2/expected)})
# Cálculo del estadístico para la tablas permutadas
p.value<-mean(statistic<=chip)
# Aproximación del p-valor de x
```

```
names(statistic)<-"X-squared"
names(p.value)<-"p-value"
structure(list(statistic = statistic, p.value = p.value))}
```

De estas tres funciones, la más lenta es la chisq.test.permuta_1, ya que el bucle que se considera en la primera función lastra mucho su funcionamiento y además hay que definir un vector de de longitud el número de individuos de la tabla, cuya permutación acarreará un coste computacional grande.

Todas las gráficas y salidas mostradas en este proyecto se han obtenido utilizando la función chisq.test.

A continuación, se mostrarán funciones de interés para estudios tipo caso-control.

Dado un estudio de este tipo, sería interesante estimar la potencia, dados número de casos, número de controles, frecuencia poblacional, frecuencia en los casos y significación:

```
est_power<-function(N_ca,N_co,freq_ca,freq_co,alpha,M,b){
# N_ca <- Número de casos
# N_co <- Número de controles
# freq_ca <- Frecuencia del haplogrupo en los casos
# freq_co <- Frecuencia del haplogrupo en los controles
# alpha <- Nivel de significación
# M <- Número de réplicas deseadas para el cálculo de la potencia
# b <- Número de permutaciones para cada tabla
if((freq_ca<0)|(freq_ca>1)){stop("La frecuencia en casos debe de estar entre 0 y 1")}
if((freq_co<0)|(freq_co>1)){stop("La frecuencia en controles debe de estar entre 0 y 1")}
if(N_ca!=floor(N_ca)){N_ca=floor(N_ca)}
if(N_co!=floor(N_co)){N_co=floor(N_co)}
if((alpha<0)|(alpha>1)){stop("El nivel de significación debe de estar entre 0 y 1")}
co<-rbinom(M,N_co,freq_co)
ca<-rbinom(M,N_ca,freq_ca)
table<-array(dim=c(2,2,M))
table[1,1,]<-rbinom(M,N_co,freq_co)
table[2,1,]<-rbinom(M,N_ca,freq_ca)
table[,2,]<-c(N_co,N_ca)-table[,1,]
# Generación de las tablas
table<-table[,,which(apply(table,3,function(z){all(colSums(z)!=0)}))]
# Eliminamos las simulaciones con marginales nulas
pvalue<-apply(table,3,function(z){chisq.test(z,simulate.p.value=TRUE,B=b)$p.value})
```

```
# Calculamos el p-valor calibrando por permutación
power<-100*mean(pvalue<alpha)
# Tomamos como potencia el promedio de p-valores menores que alpha
d<-M-dim(table)[3]
names(power)<-"Potencia"
names(d)<-"Número de tablas eliminadas (con alguna marginal nula)"
structure(list(power = power,d=d))}
```

donde N_ca denota el número de casos, N_co el número de controles, freq_ca la frecuencia del haplogrupo en los casos, freq_co la frecuencia del haplogrupo en la población, alpha el nivel de siginificación, M el número de simulaciones y b el número de permutaciones para cada simulación.

La potencia estimada viene expresada en %, y el parámetro d es el número de simulaciones que no se han tenido en cuenta a la hora de calcular la potencia por tener alguna entrada de la frecuencias esperadas nula.

Como ya se ha dicho, es de interés saber cuál es el número de controles necesarios para detectar una desviación de la hipotésis nula con determinada probabilidad, dado un número de casos, las frecuencias en casos y controles y nivel de significación. Para ello tendremos que hacer uso de la matriz con valores N'_{sc} que se ha mostrado en el trabajo:

```
est_controls<-function(N_ca,freq_ca,freq_co,alpha,beta){
# N_ca <- Número de casos
# freq_ca <- Frecuencia del haplogrupo en los casos
# freq_co <- Frecuencia del haplogrupo en los controles
# alpha <- Nivel de significación
# beta <- Potencia deseada
sig<-c(.001,.005,.01,.05,.1)
pow<-c(0.95,.90,.8,.70,.60,.50)
if((freq_ca<=0)|(freq_ca>=1)){stop("La frecuencia en casos debe de estar entre 0 y 1")}
if((freq_co<=0)|(freq_co>=1)){stop("La frecuencia en controles debe de estar entre 0 y 1")}
if(N_ca!=floor(N_ca)){N_ca=floor(N_ca)}
if((freq_ca-freq_co)==0){stop("Hipótesis erronéa")}
# Bajo la hipotésis nula la potencia coincidirá con el nivel de significación
if(all(sig!=alpha)){
stop("El nivel de significación debe de ser 0.001,0.005,0.01,0.05 o 0.1")}
i<-which(sig==alpha)
if(all(pow!=beta)){stop("La potencia debe de ser 0.95, 0.90, 0.80, 0.70, 0.60 o 0.50")}
j<-which(pow==beta)
# Seleccionamos el elemento de la matriz N'sc de interés
```

```
t<-(freq_co-freq_ca)^2/NSC[i,j]-freq_ca*(1-freq_ca)/N_ca
a="Sí"
if(t<0){N_ca<-floor(NSC[i,j]*freq_ca*(1-freq_ca)/(freq_co-freq_ca)^2)+1
a<-"No"}
# Si no se cumple la condición sobre el número de casos, refijamos este número
# para que así sea
# Si hay que cambiar el número de casos, el parámetro "a" toma el valor "No"
t<-(freq_co-freq_ca)^2/NSC[i,j]-freq_ca*(1-freq_ca)/N_ca
N_co=floor(freq_co*(1-freq_co)/t)+1
names(N_co)<-"Número mínimo de controles requerido"
names(N_ca)<-"Número de casos final"
names(a)<-"Número de casos suficientes"
structure(list(a=a,N_ca=N_ca,N_co=N_co))}
```

La matriz NSC es de la siguiente forma:

```
NSC<-rbind(c(23.04,19.57,16.20,14.87,13.82,9.32),c(17.94,13.67,14.12,10.44,7.07,6.82),
c(14.66,13.66,10.05,8.88,7.63,6.37),c(14.31,9.75,8.74,6.53,5.03,4.33),
c(9.26,8.29,6.04,5.53,3.52,2.64))
```

Otra función de interés será la que permita calcular las desviaciones mínimas de la hipotésis que podrán ser detectadas para una significación y potencia dadas, fijados número de casos, número de controles controles y frecuencia poblacional:

```
est_min_des<-function(N_ca,N_co,freq_co,alpha,beta,OR){
# N_ca <- Número de casos
# N_co <- Número de controles
# freq_co <- Frecuencia del haplogrupo en los controles
# alpha <- Nivel de significación
# beta <- Potencia deseada
# OR <- Parámetro lógico. Si OR=TRUE los resultados son expresados en téminos de odds ratio
sig<-c(.001,.005,.01,.05,.1)
pow<-c(0.95,.90,.8,.70,.60,.50)
if((freq_co<=0)|(freq_co>=1)){stop("La frecuencia en controles debe de estar entre 0 y 1")}
if(N_ca!=floor(N_ca)){N_ca=floor(N_ca)}
if(N_ca!=floor(N_co)){N_ca=floor(N_co)}
if(all(sig!=alpha)){
stop("El nivel de significación debe de ser 0.001,0.005,0.01,0.05 o 0.1")}
i<-which(sig==alpha)
if(all(pow!=beta)){stop("La potencia debe de ser 0.95, 0.90, 0.80, 0.70, 0.60 o 0.50")}
```

```
j<-which(pow==beta)
# Seleccionamos el elemento de la matriz N'sc de interés
a<-(1-2*freq_co)/N_ca
b<-sqrt(((1-2*freq_co)/N_ca)^2+4*(1-freq_co)*(1/N_ca+1/N_co)*freq_co*(1/N_ca+1/NSC[i,j]))
c<-2*freq_co*(1/N_ca+1/NSC[i,j])
t<-c((-a+b)/c,(a+b)/c)
# Cálculo de las desviaciones mínimas detectables
t1<-t[1]*100
t2<-t[2]*100
names(t1)<-"Mínimo decrecimiento detectado (%)"
names(t2)<-"Mínimo incremento detectado (%)"
if(OR==TRUE){t1<-1+(t1/100)/(1-freq_co-(t1+t1*freq_co)/100)
t2<-1-(t2/100)/(1-freq_co+(t2-t2*freq_co)/100)
names(t1)<-"Mínima OR detectada (>1)"
names(t2)<-"Máxima OR detectada (<1)"}
if(t[1]>1){t1="El decreciento mínimo no puede ser aproximado (es mayor del 100%)"}
if((freq_co+t[2]*freq_co)>1){
t2="Incremento no puede ser aproximado (freq_co sería mayor que 1)"}
freq_ca_di<-NaN
freq_ca_in<-NaN
if((is.numeric(t1))&(t[1]<1)){freq_ca_di<-freq_co-t[1]*freq_co}
if((is.numeric(t2))&((freq_co+t[2]*freq_co)<1)){freq_ca_in<-freq_co+t[2]*freq_co}
names(freq_ca_di)="Frecuencia máxima detectable (<freq_co)"
names(freq_ca_in)="Frecuencia mínima detectable (>freq_co)"
structure(list(t1=t1,t2=t2,freq_ca_di=freq_ca_di,freq_ca_in=freq_ca_in))}
```

El parámetro OR deberá de tomarse el valor TRUE si se quieren obtener las desviaciones mínimas detectables con una determianda potencia en términos de *odds ratio*. En caso contrario, se mostrará términos de la desviación relativa respecto a la frecuencia en controles.

Bibliografía

[1] Agresti, A (1984). *Analysis of Ordinal Categorical Data.* John Wiley & Sons.

[2] Armitage, P; Berry, G; Matthews, J N S (2002). *Statistical Methods in Medical Research.* Blackwell, Oxford, United Kingdom

[3] Bishop, Y M M; Fienberg, S E; Holland, P W (1975). *Discrete Multivariate Analysis. Theory and practice.* Cambridge, Massachusetts : MIT Press.

[4] Bowman, A W; Azzalini, A (1997). *Applied Smootthing Techniques for Data Analysis. The Kernel Approach with S-Plus Illustrations.* Oxford, Claderon Press

[5] Boyett, J M (1979) *Algorithm AS 144. Random R x C Tables with Given Row and Column Totals.* Appl. Statist., vol. 28, 329-332.

[6] Cochran, W G (1954). *Some Methods for Strengthening the Common χ^2 Tests.* Biometrics, vol. 10, No. 4, 417-451.

[7] Efron, B (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Society of Industrial and Applied Mathematics CBMS-NSF Monographs.

[8] Good, P (2000). *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses.* Springer.

[9] Halton, J H (1969). *A Rigorous Derivation of the Exact Contingency Formula.* Proc. Camb. Phil. Soc., vol. 65, 527-530.

[10] Jobling, M A; Hurles, M E, Tyler-Smith, C (2004). *Human Evolutionary Genetics. Origins, Peoples & Disease.* Garland Science. Taylor & Francis Group.

[11] Pagano, M; Halvorse, K T (1981). *An Algorithm for Finding the Exact Significance Levels of $r \times c$ Contingency Tables.* Journal of the American Statistical Association, vol. 76, 931-934.

[12] Parzen, E (1962). *On Estimation of a Probability Density Function and Mode.* Ann. Math. Statist., vol. 33, 1065.1076.

[13] Patefield, W M (1981). *Algorithm AS 159: An Efficient Method of Generating Random R×C Tables with Given Row and Column Totals.* Journal of the Royal Statistical Society. Series C , vol. 30, 91-97.

[14] Roff, D A, Bentzen, P (1989). *The Statistical Analysis of Mithocondrial DNA Polymorphisms: Chi 2 an the Problem of Small Samples.* Molo. Biol. Evol., vol. 6, 564-567.

[15] Rohatgi, V K (1984). *Statistical Inference.* John Wiley & Sons.

[16] Salas, A; Fachal, L; Marcos-Alonso, S; Vega, A; Martinón-Torres, F (2009).*Investigating the Role of Mitochondrial Haplogroups in Genetic Predisposition to Meningococcal Disease.* PLoS ONE 4(12): e8347. doi:10.1371/journal.pone.0008347

[17] Samuels,D C; Carothers, A D; Horton, R; Chinnery, P F (2006). *The Power to Detect Disease Associations with Mitochondrial DNA Haplogroups.* The American Journal of Human Genetics, vol. 78, 713-720.

[18] Torroni, A; Huoponen, K; Francalacci, P; Petrozzi, M; Morelli, L; Scozzari, R; Obinu, D; Savontaus, M L; Wallace, D C (1996). *Classification of European mtDNAs from an Analysis of Three European Populations.* Genetics Society of America, vol. 144, 1835-1850.

[19] Wasserman, L. (2005). *All of Nonparametric Statistics.* Springer.