Test de hipótesis múltiples y métodos de ajuste del p-valor

Juan Pérez Rubio
Trabajo de fin de grado de Matemáticas
Universidad de Zaragoza

Director del trabajo: José Tomás Alcalá Nalváiz
12 de junio de 2023

Summary

Hypothesis testing is a procedure for judging whether a property we assume about a population is compatible with what is observed in a sample of that population. Hypothesis testing is used to test two hypotheses: a null hypothesis and an alternative hypothesis. Hypothesis testing gives us a procedure for rejecting or not the null hypothesis, based on the data we have.

Over the last decades, a large volume of data is being generated in different fields of study. The needs of disciplines such as genomics or biology have motivated the development of the theory of multiple hypothesis testing, which consists of testing a large number of hypotheses at the same time. For example, understanding the genetic basis of a disease consists of determining which genes are related to the development of the disease. To do this, we have to test a large number of hypothesis at the same time, one for each gene. In each test we check whether the gene is related to the development of the disease or not. Multiple hypothesis testing gives us a process for deciding which set of null hypotheses to reject.

However, performing a hypothesis test in the usual way on each individual hypothesis does not give satisfactory results. This is because when we do hypothesis testing, we set a quantity called the significance level, usually 0,05, which is the maximum probability we are willing to tolerate of falsely rejecting a true null hypothesis. If we test a large number of hypotheses, usually tens or hundreds of thousands in this context, we will end up rejecting too many true null hypotheses, which we want to avoid.

This is why, instead of controlling for the significance level of each individual hypothesis test, three new quantities are defined, the FWER, the FDP and the FDR, which play the analogous role to the significance level, but in the context of multiple hypothesis testing. It is the value of these three magnitudes that we want to control, usually at the 0,05 level again. Depending on the experiment, we are interested in controlling for one of the three, and the strategy for multiple hypothesis testing changes.

We also look for multiple hypothesis tests that are as powerful as possible, i.e., that once it is satisfied that we control the FWER, the FDP or the FDR at the pre-specified level, as many false null hypotheses as possible are rejected.

In hypothesis testing, the concept of p-value is central. It means that, depending on the data we have, we associate a number between 0 and 1 to the hypothesis test. If that value is less than or equal to the level of significance we are willing to tolerate, we reject the null hypothesis. In multiple hypothesis testing, we can define the adjusted p-value, which plays an analogous role. Thus, we associate an adjusted p-value to each hypothesis in a multiple testing procedure, and if this value is less than or equal to the value of the FWER, FDP or FDR that we are willing to tolerate, we reject that hypothesis. The method of calculating the adjusted p-values changes depending on whether we want to control for FWER, FDP or FDR.

In the first chapter, we try to formalise the problem, giving the first definitions and results. We describe the theoretical framework that helps us to understand the problem, and we set out our study situation. We define, among others, the magnitudes already mentioned: the FWER, the FDP, the FDR and the adjusted p-value.

In the second chapter, we look at the different p-value adjustment methods, and divide them according to whether they are designed to control the FWER, the FDP or the FDR. We also describe the algorithms for calculating the adjusted p-values for each method.

The third chapter presents a simulation study in which we check that all the properties outlined in the previous chapter are fulfilled. We simulate the data corresponding to the genetic basis of a disease, and test whether each gene is related to the development of the disease or not. We calculate the unadjusted

and adjusted p-values for the different methods. Finally, we give a brief scheme of use in which we explain which method is advisable to use depending on the characteristics of the experiment.

Two appendices are included. In the first one, we present the basic theory of hypothesis testing, as knowledge of this theory is necessary to define and understand multiple hypothesis testing. The second appendix presents and explains the R code used for the simulation in the third chapter.

In conclusion, this study describes several of the main p-value adjustment methods for multiple hypothesis testing, although there are many more. The development of new multiple hypothesis tests is a relatively recent and active area of study, so new p-value adjustment methods are emerging all the time. With the increasing volume of data currently being handled, multiple hypothesis testing and p-value adjustment methods are being used and refined more and more.

Resumen

Un test de hipótesis es un procedimiento para juzgar si una propiedad que suponemos de una población es compatible con lo observado en una muestra de dicha población. Los test de hipótesis nos sirven para contrastar dos hipótesis: una hipótesis nula y una hipótesis alternativa. Un test de hipótesis nos da un procedimiento para rechazar o no la hipótesis nula, en base a los datos que tengamos.

Durante las últimas décadas, se está generando un gran volumen de datos en distintos campos de estudio. Las necesidades de disciplinas como la genómica o la biología han motivado el desarrollo de la teoría de test de hipótesis múltiples, que consisten en contrastar una gran cantidad de hipótesis al mismo tiempo. Por ejemplo, comprender la base genética de una enfermedad consiste en determinar qué genes están relacionados con el desarrollo de la enfermedad. Para ello, tenemos que hacer un gran número de test de hipótesis a la vez, uno por cada gen. En cada test contrastamos si el gen está relacionado con el desarrollo de la enfermedad o no. Un contraste de hipótesis múltiples nos da un proceso para decidir el conjunto de hipótesis nulas que debemos rechazar.

Sin embargo, realizar un contraste de hipótesis del modo habitual a cada hipótesis individual no da resultados satisfactorios. Esto es debido a que cuando hacemos un test de hipótesis, fijamos una cantidad llamada nivel de significación, normalmente de 0,05, que es probabilidad máxima que estamos dispuestos a tolerar de rechazar falsamente una hipótesis nula verdadera. Si contrastamos un gran número de hipótesis, normalmente de decenas o cientos de miles en este contexto, acabaremos rechazando demasiadas hipótesis nulas verdaderas, cosa que queremos evitar.

Es por ello que, en lugar de controlar el nivel de signifiación de cada contraste de hipótesis individual, se definen tres nuevas magnitudes, el FWER, el FDP y el FDR, que hacen el papel análogo al nivel de significación, pero en el contexto de contrastes múltiples. Es el valor de estas tres magnitudes el que queremos controlar, normalmente a nivel 0,05 de nuevo. Dependiendo del experimento, estamos interesados en controlar una de las tres, y la estrategia para hacer el test de hipótesis múltiples cambia.

También buscamos test de hipótesis múltiples que sean lo más potentes posible, es decir, que una vez se cumpla que controlamos el FWER, el FDP o el FDR al nivel prefijado, se rechacen la mayor cantidad de hipótesis nulas falsas posible.

En test de hipótesis, el concepto de p-valor es central. Consiste en que, dependiendo de los datos que tengamos, asociamos un número entre 0 y 1 al contraste de hipótesis. Si ese valor es menor o igual que el nivel de significación que estamos dispuestos a tolerar, rechazamos la hipótesis nula. En test de hipótesis múltiples, podemos definir el p-valor ajustado, que juega un papel análogo. De este modo, a cada hipótesis de un contraste múltiple le asociamos un p-valor ajustado, y si este valor es menor o igual que el valor del FWER, FDP o FDR que estamos dispuestos a tolerar, rechazamos esa hipótesis. El método de cálculo de los p-valores ajustados cambia dependiendo de si queremos controlar el FWER, el FDP o el FDR.

En el primer capítulo, tratamos de formalizar el problema, dando las primeras definiciones y resultados. Describimos el marco teórico que nos ayuda a entender el problema, y exponemos nuestra situación de estudio. Definimos, entre otras, las magnitudes ya mencionadas: el FWER, el FDP, el FDR y el p-valor ajustado.

En el segundo capítulo, vemos los distintos métodos de ajuste del p-valor, y los dividimos según estén diseñados para controlar el FWER, el FDP o el FDR. También describimos los algoritmos de cálculo de los p-valores ajustados para cada método.

En el tercer capítulo se presenta un estudio de simulación en el que comprobamos que todas las

propiedades expuestas en el capítulo previo se cumplen. Simulamos los datos correspondientes a la base genética de una enfermedad, y contrastamos si cada gen está relacionado con el desarrollo de la enfermedad o no. Calculamos los p-valores sin ajustar y ajustados para los distintos métodos. Finalmente, damos un breve esquema de uso en el que explicamos qué método es recomendable usar dependiendo de las características del experimento.

Se incluyen dos anexos. En el primero de ellos exponemos la teoría básica de test de hipótesis, ya que conocer esta teoría es necesario para definir y entender los test de hipótesis múltiples. En el segundo anexo se presenta y explica el código de R utilizado para la simulación del tercer capítulo.

Como conclusión, señalar que en este trabajo se describen varios de los principales métodos de ajuste del p-valor para test de hipótesis múltiples, aunque hay muchos más. El desarrollo de nuevos contrastes de hipótesis múltiples es un área de estudio relativamente reciente y activa, por lo que están apareciendo nuevos métodos de ajuste del p-valor continuamente. Con el creciente volumen de datos que se está manejando actualmente, los test de hipótesis múltiples y métodos de ajuste del p-valor se están utilizando y perfeccionando cada vez más.

Índice general

Capítulo 1

Introducción

1.1. Motivación

Históticamente, ha habido necesidad de contrastar la igualdad de medias entre dos o más muestras de datos. Para ello, los procedimientos ANOVA *(Analysis of Variance)* han sido los más utilizados. En ellos, enfrentamos una hipótesis nula, de que todas las medias son iguales, contra una hipótesis alternativa, de que al menos un par de medias son distintas. Es decir, suponiendo que tenemos k muestras, estamos enfrentando

$$H_0 : \mu_1 = \ldots = \mu_k \quad \text{vs.} \quad H_1 : \exists i, j \ni \mu_i \neq \mu_j$$

El procedimiento para estudiar este tipo de contrastes ya ha sido ampliamente descrito, ver por ejemplo [20]. La desventaja es que para poder ejecutar este tipo de procedimientos ANOVA necesitamos que se cumplan estas tres hipótesis:

- Muestras aleatorias independientes.
- Las k muestras vienen de distribuciones normales.
- Las varianzas de las k muestras deben ser iguales.

Por ello, si se quieren contrastar otro tipo de hipótesis o no se cumplen estas condiciones, ANOVA no nos ofrece una solución. Además, cuando se rechaza la hipótesis nula, surge la necesidad de un procedimiento que permita comparar los $\binom{k}{2}$ pares de medias para determinar qué pares de medias son iguales y cuáles no. Esto es algo que en los últimos años se ha ido necesitando más y más, sobre todo debido al desarrollo de la genómica. Los test de hipótesis múltiples que estudiamos en este trabajo tratan de dar respuesta a esta necesidad de ampliar los clásicos contrastes de hipótesis.

Durante las últimas décadas, se están generando datos a una escala masiva en distintos campos de estudio. La biología es probablemente una de las que más volumen de datos produce, y ramas como la genómica o la farmacogenómica requieren contrastar una gran cantidad de hipótesis al mismo tiempo, incluso a escala de cientos de miles o millones. Por ejemplo, comprender la base genética de una determinada enfermedad implica analizar la expresión de miles de genes entre distintos grupos de pacientes. En este contexto, la expresión de un gen es una forma de matematizar su valor para poder comparar los genes de distintos individuos. Las necesidades de estas disciplinas han exigido el desarrollo matemático de diversos métodos para el testeo de una gran cantidad de hipótesis al mismo tiempo, tratando de reducir lo máximo posible los errores que cometemos a la hora de rechazar o no un número tan elevado de hipótesis.

En este trabajo, explicamos la necesidad de introducir los test de hipótesis múltiples y vemos los problemas que surgen a la hora de elegir el nivel de error que podemos tolerar cuando tratamos con un número tan elevado de hipótesis, ya que podemos definir el error de tipo I de varias maneras en este contexto. Definiremos el concepto de p-valor ajustado, que, al igual que el p-valor en los test de hipótesis simples nos da una regla de decisión para rechazar la hipótesis nula o no, el p-valor ajustado nos permite

decidir si rechazar o no cada hipótesis de un test de hipótesis múltiples. Posteriormente, veremos los principales métodos que hay para determinar los p-valores ajustados según cómo hayamos definido el error de tipo I, y compararemos estos métodos para determinar cuál es recomendable usar según el contexto. Utilizaremos el software R para ver cómo funciona cada uno de los métodos expuestos.

1.2. Formalización del problema

El problema de los contrastes de hipótesis en estadística ha sido ampliamente estudiado y desarrollado por diversos autores. Un buen modo de introducirse a los contrastes de hipótesis es el libro *Testing Statistical Hypothesis*, de E. L. LEHMANN Y JOSEPH P. ROMANO [1], referente en la materia. En el Anexo A hemos descrito los test de hipótesis simples y sus principales características. Comprender los test de hipótesis simples es necesario para poder introducir los test de hipótesis múltiples.

Las definiciones y notación introducidas en este capítulo están basadas en las dadas en [2].

1.2.1. Primeras nociones

Comenzamos definiendo el concepto de modelo estadístico:

Definición 1. Un modelo estadístico $\mathscr{M}$ es un par $(S, \mathscr{P})$ donde S es el conjunto de todas las observaciones posibles, y $\mathscr{P}$ es un conjunto de distribuciones de probabilidad sobre S.

La intuición detrás de esta defincón es que hay una distribución de probabilidad verdadera que genera los datos observados. Elegimos $\mathscr{P}$ de modo que esté formado por un conjunto de distribuciones de probabilidad que contenga una distribución que aproxime de manera adecuada la distribución de probabilidad verdadera.

Suponemos que el conjunto $\mathscr{P}$ es *paramétrico*, es decir $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, donde $\Theta \subseteq \mathbb{R}^m$ es el conjunto de parámetros del modelo.

Nuestra situación de estudio

Sean $X_1, \ldots, X_n$ vectores aleatorios independientes, cada uno de tamaño d, $X = (X(j) : j = 1, \ldots, d) \sim P_\theta \in \mathscr{M}$, donde la distribución que genera los datos, P_θ pertenece a un modelo estadístico $\mathscr{M}$. Particularizando al caso de la genómica, $(X_i(1), \ldots, X_i(d))$ es un vector de tamaño d con medidas de expresiones de genes de un paciente $i = 1, \ldots, n$. Normalmente, el número de expresiones de genes estudiadas para cada paciente, d, suele ser del orden de decenas o cientos de miles, mientras que el del número de pacientes, n, es mucho mas pequeño, normalmente de cientos.

Un modo habitual de realizar los experimentos es separar a los pacientes en dos grupos. Al primer grupo lo denominamos el grupo de *control*, formado por individuos sanos, que no tienen la enfermedad o condición que estamos estudiando. Al segundo grupo lo llamamos el grupo de *casos*, formado por individuos que tienen la enfermedad o condición objeto de estudio. El grupo de control es de tamaño n_1 mientras que el grupo de casos es de tamaño n_2. Obviamente, $n_1 + n_2 = n$.

El objetivo del estudio es identificar los genes que causan una enfermedad. Para ello lo que hacemos es comparar la media de la expresión de cada gen en el grupo de control y en el grupo de casos. Si rechazamos que las medias son iguales, podremos decir que tenemos evidencia estadística de que ese gen está relacionado con el desarrollo de la enfermedad.

Parámetros

Los parámetros que caracterizan la distribución P_θ son $\theta \in \mathbb{R}^m$. Como la función de distribución P_θ genera cada vector $(X_i(1), \ldots, X_i(d))$, podemos poner $\theta = \{\theta_j : j = 1, \ldots, d\}$, donde cada elemento de este conjunto caracteriza la distribución generadora de $X_i(j)$, con $i = 1, \ldots, n$ y $j = 1, \ldots, d$. Es decir, la distribución generadora de $X_i(j)$ está caracterizada por θ_j, que a su vez puede estar formado por más de un elemento (por ejemplo, una distribución normal está caracterizada por dos valores, su media y su

varianza), es decir, $\theta_j = (\theta_{j_1}, \ldots, \theta_{j_{k(j)}})$. Por tanto, podemos expresar θ como un vector formado por los siguientes elementos, $\theta = \{\theta_1, \ldots, \theta_j\} = (\theta_{1_1}, \ldots, \theta_{1_{k(1)}}, \ldots, \theta_{d_1}, \ldots, \theta_{d_{k(d)}})$. Asumiendo que el tamaño de este vector es m, por simplificar, podemos escribir $\theta = (\theta_1, \ldots, \theta_m)$, donde ahora cada elemento es un número. Nuestro objetivo es plantear un test de hipótesis para estos parámetros unidimensionales, por lo que tenemos que plantear m test de hipótesis. Los parámetros más habituales son medias, diferencias de medias, varianzas, coeficientes de correlacion, covarianzas, etc.

Para el caso de estudio experimental definido en el párrafo anterior, en el que distinguimos entre el grupo de casos y el de control, se tiene $m = d$, pues para cada gen estudiamos solamente la diferencia de medias entre los dos grupos. Definimos pues los parámetros $\mu_i(j)$, la media de la variable aleatoria $X_i(j)$, que representa la medida del gen j-ésimo para el paciente i-ésimo, con $i = 1, \ldots, n$ y $j = 1, \ldots, m$.

Por lo tanto, en el caso de que el gen j-ésimo no esté relacionado con el desarrolo de la enfermedad que estamos estudiando, se tiene $\mu_1(j) = \ldots = \mu_{n_1}(j) = \mu_{n_1+1}(j) = \ldots = \mu_{n_2}(j)$, y en el caso de que sí lo esté, $\mu_1(j) = \ldots = \mu_{n_1}(j) \neq \mu_{n_1+1}(j) = \ldots = \mu_{n_2}(j)$. Definimos por lo tanto $\mu^1(j) = \mu_{\text{control}}(j) = \mu_1(j) = \ldots = \mu_{n_1}(j)$ y $\mu^2(j) = \mu_{\text{casos}}(j) = \mu_{n_1+1}(j) = \ldots = \mu_{n_2}(j)$.

De este modo, queremos comparar si los parámetros $\mu^1(j)$ y $\mu^2(j)$ son iguales o no para cada j.

Hipótesis nulas generales

Definimos m hipótesis nulas en función del espacio paramétrico $\Theta \subseteq \mathbb{R}^m$. Como hemos visto, $\theta = (\theta_1, \ldots, \theta_m)$, donde cada elemento de este vector es un número (unidimensional), luego se tiene $\Theta = \Theta_1 \times \ldots \times \Theta_m$, con $\Theta_j \subseteq \mathbb{R}$ y $\theta_j \in \Theta_j$ para cada $j = 1, \ldots, m$. Estamos por tanto en un caso igual al descrito en el Anexo A para cada $j = 1, \ldots, m$. Las m hipótesis nulas son definidas como $H_{0j} : \theta_j \in \Theta_{0j}$, y las correspondientes hipótesis alternativas son $H_{1j} : \theta_j \in \Theta_{1j}$, donde Θ_{0j} y Θ_{1j} son una partición del espacio paramétrico $\Theta_j \subseteq \mathbb{R}$. Por lo tanto, H_{0j} es verdadera si $\theta_j \in \Theta_{0j}$ y falsa si $\theta_j \notin \Theta_{0j}$.

Los casos de estudio más habituales son de la siguiente forma:

- Contrastes unilaterales: $H_{0j} : \theta_j \leq \theta_{0j}$ vs. $H_{1j} : \theta_j > \theta_{0j}$ o
  $H_{0j} : \theta_j \geq \theta_{0j}$ vs. $H_{1j} : \theta_j < \theta_{0j}$
- Contrastes bilaterales: $H_{0j} : \theta_j = \theta_{0j}$ vs. $H_{1j} : \theta_j \neq \theta_{0j}$

donde $\theta_{0j} \in \mathbb{R}$.

En el caso del estudio experimental descrito anteriormente, los contrastes de hipótesis comparan la difernecia de medias de dos muestras independientes, y tienen la forma del siguiente contraste bilateral:

$$H_{0j} : \mu^1(j) - \mu^2(j) = 0 \quad \text{vs.} \quad H_{1j} : \mu^1(j) - \mu^2(j) \neq 0$$

Estadísticos de contraste

La decisión de rechazar o no cada hipótesis nula se basa en un vector T_n de tamaño m formado por estadísticos de contraste, $T_n = (T_n(j) : j = 1, \ldots, m)$, que son funciones de los datos $X_1, \ldots, X_n$. Dependiendo del valor de cada estadístico de contraste $T_n(j)$ rechazamos o no la hipótesis nula correspondiente H_{0j}. Asumimos de ahora en adelante que valores crecientes de $T_n(j)$ proporcionan evidencia en contra de la hipótesis nula H_{0j}, por lo que si se cumple $T_n(j) \geq k_j$ para un valor límite k_j, rechazamos la hipótesis H_{0j}. Para valores decrecientes, se procedería de manera análoga cambiando las desigualdades correspondientes, y para contrastes bilaterales, basta con tomar el valor absoluto del estadístico de contraste.

En el ejemplo particular de la diferencia de medias, el estadístico de contraste es el que se utiliza en una prueba de *t de Student* (ver Sección 5.3 de [1]), que nos sirve para comparar si las dos medias que vienen de dos muestras independientes son iguales o no. Para poder utilizar este estadístico, nuestros datos tienen que venir de dos distribuciones normales, pero en el caso de la expresión de genes esto se cumple.

En el caso de que las dos distribuciones tengan varianzas iguales, el estadístico de contraste está definido como

$$T_n(j) = \frac{\overline{X}_{\text{control}}(j) - \overline{X}_{\text{casos}}(j)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde $\overline{X}_{\text{control}}$ es la media muestral de la expresión del gen j-ésimo del grupo de control, mientras que $\overline{X}_{\text{casos}}$ es la media muestral de la expresión del gen j-ésimo del grupo de casos. Como ya hemos dicho, n_1 y n_2 son los respectivos tamaños de cada grupo, y S es

$$S = \sqrt{\frac{(n_1-1)S^2_{\text{control}} + (n_2-1)S^2_{\text{casos}}}{n_1+n_2-2}}$$

donde S_{control} y S_{casos} son simplemente la cuasi-desviación típica muestral de la muestra de cada grupo.

En el caso de que las dos distribuciones no tengan varianzas iguales, el estadístico de contraste es el que se utiliza en un *test de Welch* [1, pág. 447],

$$T_n(j) = \frac{\overline{X}_{\text{control}}(j) - \overline{X}_{\text{casos}}(j)}{S_W}$$

donde $S_W = \sqrt{\frac{S^2_{\text{control}}}{n_1} + \frac{S^2_{\text{casos}}}{n_2}}$

1.2.2. Test de hipótesis múltiples

Antes de dar la definición de test de hipótesis múltiples, introducimos la notación $S_0 = \{j : H_{0j} \text{ es verdadera}\} = \{j : \theta_j \in \Theta_{0j}\}$, que es el conjunto de tamaño $m_0 = |S_0|$ de hipótesis nulas verdaderas. Análogamente, $S_0^c = \{j : H_{0j} \text{ es falsa}\} = \{j : \theta_j \notin \Theta_{0j}\}$ es el conjunto de tamaño $m_1 = m - m_0$ de hipótesis nulas falsas. El objetivo de los test de hipótesis múltiples es estimar lo más precisamente posible el conjunto S_0, y por tanto su complementario S_0^c, a la vez que controlamos el error de tipo I (que definiremos más adelante para este contexto) a un nivel α proporcionado por el usuario.

Un test de hipótesis múltiples nos proporciona un conjunto S_n de hipótesis nulas rechazadas, que estima S_0^c, el conjunto de hipótesis nulas falsas:

$$S_n = \{j : H_{0j} \text{ es rechazada}\} \subseteq \{1, \ldots, m\}$$

Errores de tipo I y tipo II

Como se puede ver en el Anexo A, cuando tomamos la decisión de rechazar o no una hipótesis H_{0j}, estamos sujetos a error, que puede ser de tipo I (rechazarla si es verdadera) o de tipo II (no rechazarla si es falsa).

En el caso de test de hipótesis múltiples, con m hipótesis a contrastar, el número de errores de tipo I que cometemos es $V_n = |S_n \cap S_0|$, es decir, el número hipótesis nulas que son ciertas y rechazadas. Análogamente, el número de errores de tipo II que cometemos es $U_n = |S_n^c \cap S_0^c|$, el número de hipótesis nulas que son falsas y no rechazadas. También definimos $R_n = |S_n|$ el número de hipótesis rechazadas. La situación se puede resumir en la siguiente tabla, que respresenta el número de errores de tipo I y II cuando contrastamos m hipótesis:

	Número de H_{0j} no rechazadas	Número de H_{0j} rechazadas	Total
Número de H_{0j} ciertas	$m_0 - V_n$	V_n	m_0
Número de H_{0j} falsas	U_n	$m_1 - U_n$	m_1
Total	$m - R_n$	R_n	m

Tabla 1.1: Número de errores cometidos al contrastar m hipótesis

Idealmente, nos gustaría minimizar el número de errores de tipo I, V_n, y el número de errores de tipo II, U_n, simultáneamente. Sin embargo, el enfoque que tomaremos en este tipo de problemas es parecido al que tomamos en el caso en el que solo tenemos un contraste de hipótesis. Para explicar el procedimiento, necesitamos definir los análogos al error de tipo I y a la potencia que hemos definido para test de hipótesis simples en el Anexo A, pero en el contexto de test de hipótesis múltiples.

La idea para definir la potencia es que, a más hipótesis falsas rechazadas, más potencia tiene un test de hipótesis múltiples.

Definición 2. Definimos la *potencia* de un test de hipótesis múltiples de un modo análogo al que se define la potencia de un test de hipótesis simples. Recordar que en este caso es $1-\Pr$(cometer un error de tipo II), es decir, es la probabilidad de rechazar una hipótesis nula falsa. En el caso de los test de hipótesis múltiples la definimos como

$$E\left(\frac{m_1 - U_n}{m_1}\right) = 1 - E\left(\frac{U_n}{m_1}\right)$$

Es decir, es el valor esperado de la proporción de hipótesis nulas falsas rechazadas sobre el total de hipótesis nulas falsas.

Es importante señalar que esta definición de potencia es una definición *"ad hoc"*, ya que está adaptada al modo en el que se ha estimado la potencia en la simulación del Capítulo 3, pero no sirve para otros fines.

En el contexto de test de hipótesis múltiples, varias definiciones son posibles para el error de tipo I "global" para el conjunto de todas las hipótesis, pero la idea es la misma: tratamos de reducir este error de tipo I sea cual sea su definición con distintos métodos, y veremos en qué situaciones es más adecuado utilizar cada uno de ellos. Las definiciones del error de tipo I en test de hipótesis múltiples son:

Definición 3. *Family-wise error rate* (FWER), es la probabilidad de cometer al menos un error de tipo I para alguna hipótesis H_{0j},

$$\text{FWER} = \Pr(V_n \geq 1)$$

Definición 4. *False discovery proportion* (FDP), es la proporción entre falsos rechazos de la hipótesis nula y el total de rechazos,

$$\text{FDP} = \frac{V_n}{R_n}$$

con el convenio de FDP$= 0$ si $R_n = 0$

Definición 5. *False discovery rate* (FDR), es el promedio de todas las réplicas posibles para el cálculo de FDP. En otras palabras, es la esperanza de FDP independientemente del número de rechazos, es decir, FDR$= E[\text{FDP}]$. Por tanto, FDR $= E\left[\frac{V_n}{R_n}\right]$ si $R_n \neq 0$ y FDR $= 0$ si $R_n = 0$.

Además, el FDR se puede expresar como

$$\text{FDR} = E\left[\frac{V_n}{R_n} | R_n > 0\right] \Pr(R_n > 0)$$

Análogamente a lo que hacíamos en el caso de hipótesis simples, el modo de proceder en test de hipótesis múltiples es elegir un valor α, llamado nivel de significación, y consideramos todos los test tales que la probabilidad de cometer un error de tipo I (el test cambia dependiendo de cuál de las tres definiciones dadas consideremos) sea menor o igual que α. Una vez se cumpla esta condición, tratamos de buscar el test que minimice lo máximo posible el número de errores de tipo II cometidos, U_n, o dicho de otro modo, aumentamos lo máximo posible la potencia del test.

Este enfoque se debe a que el papel de la hipótesis nula y alternativa no es el mismo. En la práctica, la hipótesis nula tiene el papel de *"status quo"*, es decir, lo aceptado o aceptable en ausencia de evidencia.

Por tanto, queremos que la probabilidad de cometer el error de rechazar la hipótesis nula cuando es cierta sea baja, por eso se suele elegir un valor $\alpha = 0{,}05$ o $\alpha = 0{,}01$, aunque esto solo se trata de un convenio.

Nuestro objetivo es diseñar test de hipótesis múltiples que controlen uno de los errores de tipo I que acabamos de definir al nivel α que consideremos. Este control se puede dar en dos sentidos:

Definición 6. Hablaremos de *control débil* del error de tipo I que consideremos si el test de hipótesis múltiples solo nos puede garantizar el control de dicho error de tipo I a nivel α si se cumple que todas las hipótesis nulas son verdaderas, es decir, si $m = m_0$.

Definición 7. Hablaremos de *control fuerte* del error de tipo I que consideremos si el test de hipótesis múltiples nos garantiza el control de dicho error de tipo I a nivel α sea cual sea la combinación de hipótesis nulas falsas y verdaderas.

Test de hipótesis múltiples basados en estadísticos de contraste

Como ya hemos dicho, un test de hipótesis múltiples nos proporciona un conjunto S_n de hipótesis nulas rechazadas, que estima S_0^c, el conjunto de hipótesis nulas falsas.

$$S_n = S(T_n, Q_0, \alpha) = \left\{ j : H_{0j} \text{ es rechazada} \right\} \subseteq \{1, \ldots, m\}$$

Como hemos indicado en la notación, S_n depende de:

1. Los datos $X_1, \ldots, X_n$ a través del vector de estadísticos de contraste T_n.
2. El conjunto de distribuciones nulas Q_0, que es el conjunto de funciones de distribución de cada componente del vector de estadísticos de contraste, T_n, cuando cada hipótesis nula es cierta. Es decir, $Q_0 = \left\{ Q_{0j}\,,\, j = 1, \ldots, m \right\}$, donde Q_{0j} es la distribución de $T_n(j)$ si H_{0j} es cierta.
3. Un valor α que es una cota superior para el error de tipo I que estamos dispuestos a cometer. Dependiendo de si usamos el FWER, el FDP o el FDR (ver Definiciones 3, 4 y 5 respectivamente) como error de tipo I, el conjunto S_n cambia.

Los test de hipótesis múltiples con la hipótesis de que valores crecientes de $T_n(j)$ proporcionan evidencia en contra de la hipótesis nula H_{0j} se pueden representar como

$$S_n = S(T_n, Q_0, \alpha) = \left\{ j : T_n(j) \geq k_j \right\}$$

donde $k_j = k_j(T_n, Q_0, \alpha),\ j = 1, \ldots, m$, son los valores críticos que determinan si rechazamos o no una hipótesis nula. Si el valor del estadístico $T_n(j)$ es mayor o igual que el valor crítico k_j, rechazamos la hipótesis H_{0j}. Además, elegimos este valor para que la probabilidad de cometer un error de tipo I (recordar que hay tres definiciones de error de tipo I en este contexto) sea $\leq \alpha$.

1.2.3. P-valores sin ajustar y ajustados

Recordar la definición de p-valor para test de hipótesis simples dada en (A.2), o la dada en (A.3), basada en estadísticos de contraste y en la que asumimos que valores crecientes de éstos apoyan la hipótesis alternativa, como ya hemos supuesto. De ahora en adelante, nos referimos a esta definición del p-valor como el *p-valor sin ajustar*, y para cada hipótesis nula H_{0j} tenemos su correspondiente p-valor sin ajustar, al que denotamos por $P_n(j)$. Recordar que, al igual que los estadísticos de contraste, los p-valores sin ajustar sirven para tener una regla de decisión a la hora de decidir si rechazamos o no H_{0j}, ya que la rechazamos si $P_n(j) \leq \alpha$. Por la definición original de p-valor sin ajustar, si usamos la regla de decisión que acabamos de describir, la probabilidad de cometer un error de tipo I es $\leq \alpha$ a la hora de decidir si rechazamos o no H_{0j}.

Los p-valores ajustados nos sirven para lo análogo en el caso de test de hipótesis múltiples. Es decir, a cada hipótesis H_{0j} le asociamos un número, el p-valor ajustado, y en el caso de que el p-valor ajustado sea $\leq \alpha$, rechazamos la hipótesis correspondiente. Los calculamos de tal forma que, con esta regla de

decisión, mantengamos el error de tipo I "global" menor o igual que α. Dependiendo la definición del error de tipo I que apliquemos (Definiciones 3, 4 o 5), el cálculo de los p-valores ajustados cambia.

La introducción del concepto de p-valor ajustado es necesaria porque si utilizaramos el p-valor sin ajustar para decidir si rechazamos H_{0j} o no con un nivel de significación α para cada hipótesis, no podríamos asegurar un nivel de significación global α para el conjunto de hipótesis, que es lo que queremos.

Definimos los p-valores ajustados basándonos en la definición de los p-valores sin ajustar dada en (A.3), y para la que necesitamos de un estadístico de contraste y que valores crecientes del estadístico apoyen la hipótesis alternativa (aunque si esto último no es así basta con cambiar las desigualdades o añadir valores absolutos para adaptarlo).

Definición 8. Dado un test de hipótesis múltiples

$$S_n = S(T_n, Q_0, \alpha) = \left\{ j : T_n(j) \geq k_j(T_n, Q_0, \alpha) \right\}$$

basado en los valores límite $k_j = k_j(T_n, Q_0, \alpha)$, definimos el **p-valor ajustado** para la hipótesis nula H_{0j} como

$$\begin{aligned} \widetilde{P}_n(j) &= \inf\left\{ \alpha \in [0,1] : H_{0j} \text{ es rechazada en el test de hipótesis múltiples a nivel } \alpha, \text{ dado } T_n(j) \right\} \\ &= \inf\left\{ \alpha \in [0,1] : j \in S(T_n, Q_0, \alpha) \right\} \\ &= \inf\left\{ \alpha \in [0,1] : T_n(j) \geq k_j(T_n, Q_0, \alpha) \right\} \end{aligned}$$

Es decir, el p-valor ajustado es el nivel de significación más pequeño que resulta en el rechazo de la hipótesis nula a ese nivel de significación para el test de hipótesis múltiples, dado $T_n(j)$. Recordar que con rechazar una hipótesis en un test de hipótesis múltiples con nivel de significación α nos podemos referir a que queremos que el FWER, FDP o FDR sean menores o iguales que ese valor α. Dependiendo de la elección de una de estas tres definiciones, y de la estrategia para minimizar su valor, la defínción de $k_j(T_n, Q_0, \alpha)$ cambia.

Esta definición no es práctica a la hora de calcular los p-valores ajustados. Lo que haremos realmente será aplicar una serie de algoritmos o métodos a la lista de p-valores sin ajustar $(P_n(1), \ldots, P_n(m))$ para obtener los p-valores ajustados $(\widetilde{P}_n(1), \ldots, \widetilde{P}_n(m))$ que cumplan que si aplicamos la regla de decisión de rechazar H_{0j} si $\widetilde{P}_n(j) \leq \alpha$ entonces se cumpla que el error de tipo I que queremos controlar (FWER, FDP o FDR) se mantenga $\leq \alpha$.

Al igual que pasaba en test de hipótesis simples, tenemos dos reglas de decisión para rechazar una hipótesis de un test de hipótesis múltiples mientras mantenemos el nivel de significación α. La primera, basada en los valores límite $k_j = k_j(T_n, Q_0, \alpha)$ para el estadístico de contraste $T_n(j)$,

$$S_n = S(T_n, Q_0, \alpha) = \left\{ j : T_n(j) \geq k_j(T_n, Q_0, \alpha) \right\}$$

y la segunda basada en los p-valores ajustados $\widetilde{P}_n(j)$,

$$S_n = S(T_n, Q_0, \alpha) = \left\{ j : \widetilde{P}_n(j) \leq \alpha \right\}$$

Es decir, si queremos un nivel de significación α, la hipótesis H_{0j} es rechazada si $\widetilde{P}_n(j) \leq \alpha$. La ventaja de utilizar los p-valores ajustados como regla de decisión es que el nivel de significación no tiene que ser fijado previamente, si no que podemos calcular los $\widetilde{P}_n(j)$ y luego elegir si rechazamos o no cada hipótesis según el nivel de significación que estemos dispuestos a tolerar. Además, los p-valores nos sirven para ver cuánta evidencia tenemos para rechazar una hipótesis. Cuanto más pequeño sea el p-valor, más evidencia.

1.2.4. Los distintos tipos de test de hipótesis múltiples: *single-step* y *stepwise*

Distinguimos dos tipos de test de hipótesis múltiples: los procedimientos *single-step* y los procedimientos *stepwise*.

Definición 9. En los procedimientos *single-step*, el vector de valores límite $k = (k_j : j = 1, \ldots, m)$ para el estadístico de contraste T_n es constante, es decir, el vector k no depende de T_n.

Por tanto, cada hipótesis H_{0j} es evaluada usando un valor crítico $k_j = k_j(Q_0, \alpha)$, que no depende del resultado de los test para otras hipótesis y no es función de los datos $X_1, \ldots, X_n$. Aplicando la definición de p-valor ajustado, vemos que en los procedimientos *single-step*, el valor del p-valor ajustado $\widetilde{P}_n(j)$ no depende del resultado del test para otras hipótesis distintas a H_{0j}.

Los siguientes procedimientos se diseñan para aumentar la potencia de un test de hipótesis múltiples mientas que se sigue manteniendo el control del error de tipo I:

Definición 10. Los procedimientos *stepwise* rechazan o no una hipótesis dependiendo del resultado del test para otras hipótesis. Esto es, los valores límite $k_j = k_j(T_n, Q_0, \alpha)$ pueden depender de los datos $X_1, \ldots, X_n$ a través del estadístico de contraste T_n, y por lo tanto el valor del p-valor ajustado $\widetilde{P}_n(j)$ depende del resultado del test para las demás hipótesis nulas.

Generalmente, estos procedimientos consideran los p-valores ajustados de forma ordenada. Resulta entonces que, los procedimientos *stepwise* se dividen a su vez en dos tipos:

Definición 11. Los procedimientos *step-down*, son aquellos que consideran sucesivamente las hipótesis correspondientes a los estadísticos de contraste más significativos, y por tanto a los p-valores sin ajustar más pequeños, desde el p-valor sin ajustar más pequeño al más grande. En estos procedimientos, el rechazar o no una hipótesis depende del resultado de las anteriores, pues una vez llegamos a no rechazar una hipótesis nula H_{0j}, tampoco son rechazadas las hipótesis posteriores.

Definición 12. Los procedimientos *step-up*, son aquellos que consideran sucesivamente las hipótesis correspondientes a los estadísticos de contraste menos significativos, y por tanto a los p-valores sin ajustar más grandes, desde el p-valor sin ajustar más grande al más pequeño. En estos procedimientos, el rechazar o no una hipótesis depende del resultado de las anteriores, pues una vez llegamos a rechazar una hipótesis nula H_{0j}, también son rechazadas las hipótesis posteriores.

Capítulo 2

Métodos para controlar los distintos errores de tipo I

2.1. Métodos para controlar el FWER

Ejemplo introductorio: Supongamos que queremos comparar la expresión de 1000 genes independientes para dos grupos de 10 individuos cada uno. El primer grupo será el grupo de control formado por individuos sanos y el grupo de casos estará formado por individuos enfermos. Queremos ver si los valores medios de cada expresión de un gen son iguales entre los dos grupos o no. Tenemos que contrastar por lo tanto 1000 hipótesis. Supongamos que estamos en el caso en el que los valores medios de cada expresión son iguales para cada gen entre los dos grupos, es decir, la enfermedad no está relacionada con la expresión de ninguno de los 1000 genes que estamos considerando, por lo que $m = m_0$. Con la notación del Capítulo 1, se tiene $n_1 = n_2 = 10$, $n = 20$, $d = m = 1000$. Suponer que para cada hipótesis realizamos un test con nivel de significación $\alpha = 0{,}05$. En este caso, para cada test, la probabilidad de no cometer un error de tipo I es de $1 - 0{,}05 = 0{,}95$. Por tanto, para el conjunto de 1000 test, la probabilidad de no cometer ningún error de tipo I es de $(1-\alpha)^{1000} = 0{,}95^{1000} = 5{,}3 \times 10^{-23}$, por lo que la probabilidad de cometer al menos un error de tipo I, es decir, el FWER, es FWER$= \Pr(V_{20} \geq 1) = 1 - (1-\alpha)^{1000} = 1 - 5{,}3 \times 10^{-23} \approx 1$, ya que como todas las hipótesis nulas son verdaderas, cada hipótesis que rechazamos es un falso rechazo.

Con este sencillo ejemplo hemos visto que si contrastamos cada hipótesis con nivel de significación α, no podemos asegurar que el FWER sea menor o igual que α. De hecho, si el número de test es elevado como en este caso, su valor se acerca mucho a 1. Por eso es necesario introducir los test de hipótesis múltiples y los p-valores ajustados, para garantizar que podemos mantener el FWER por debajo de un nivel prefijado.

Supongamos además que los p-valores sin ajustar correspondientes a las hipótesis nulas verdaderas cumplen lo siguiente:

$$\Pr(P_n(j) \leq u) \leq u \ \ \forall u \in [0,1] \tag{2.1}$$

La mayoría de los procedimientos que se van a exponer requieren que los p-valores cumplan esta propiedad para poder demostrar su validez.

Si en particular se cumple

$$\Pr(P_n(j) \leq u) = u \ \ \forall u \in [0,1] \tag{2.2}$$

para los p-valores sin ajustar correspondientes a las hipótesis nulas ciertas, tenemos que en este caso los p-valores siguen una distribución uniforme.

Las condiciones para que se cumplan estas dos propiedades están dadas en el Teorema A.1.

2.1.1. Método de Bonferroni

El método de Bonferroni [3, pág. 345, 352, 353] [4] fue uno de los primeros métodos desarrollados para los test de hipótesis múltiples. Es capaz de controlar el FWER a nivel α. El procedimiento consiste

en rechazar las hipótesis tales que $P_n(j) \leq \alpha/m$, siendo $P_n(j)$ el p-valor sin ajustar para la hipótesis H_{0j}. Es decir, rechazamos las hipótesis tales que $mP_n(j) \leq \alpha$, por lo que los p-valores ajustados son:

$$\widetilde{P}_{n,bonf}(j) = \min\{mP_n(j), 1\} \tag{2.3}$$

Se trata por lo tanto de un método *single-step*, ya que la decisión de rechazar o no cada hipótesis nula no depende de si rechazamos o no las demás.

Una de las ventajas de este método, aparte de su sencillez, es que tiene un control fuerte del FWER a nivel α. Recordar que esto quiere decir que controla el FWER para cualquier combinación de hipótesis nulas falsas y verdaderas. También sirve para cualquier estructura de dependencia de los p-valores, siempre que se cumpla (2.1). Veremos que en otros métodos esto no tiene porqué ser así.

Teorema 2.1. *El método de Bonferroni dado en* (2.3) *controla el FWER a nivel* α.

Demostración. Sea I_0 el conjunto de indices correspondientes a las hipótesis nulas verdaderas. Sabemos que $|I_0| = m_0$.

Recordar que el FWER es la probablilidad de rechazar al menos un H_{0j} verdadero, luego se tiene:

$$\text{FWER} = \Pr\left\{\bigcup_{j\in I_0}\left(P_n(j) \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{j\in I_0}\Pr\left\{P_n(j) \leq \frac{\alpha}{m}\right\} \leq m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha$$

Donde hemos utilizado la propiedad (2.1) que deben cumplir los p-valores sin ajustar correspondientes a hipótesis nulas verdaderas. □

También podemos decir que el método de Bonferroni controla el FWER a nivel $\pi_0\alpha$, siendo $\pi_0 = \dfrac{m_0}{m}$ la proporción de hipótesis nulas verdaderas sobre el total de hipótesis. Por lo tanto este método controla el FWER a un nivel menor o igual que α siempre, por lo que diremos que se trata de un método conservador. Además, si hay muchas hipótesis nulas falsas, el FWER es controlado a un nivel significativamente menor que α, es decir, el método es muy conservador. En general, el método de Bonferroni es muy conservador si los p-valores están positivamente correlacionados y menos conservador si los p-valores son independientes [5, pág. 1955].

El precio a pagar por que este método sea aplicable en cualquier situación es que no es un método muy potente en general, es decir, la probabilidad de cometer muchos errores de tipo II es alta. Esto es porque al tener un gran número de hipótesis, m, la condición de rechazar cada hipótesis H_{0j} si $P_n(j) \leq \dfrac{\alpha}{m}$ es una condición relativamente dificil de cumplir, ya que el valor α/m es muy pequeño, por lo que no rechazamos un gran número de hipótesis, entre las que hay hipótesis falsas, es decir, cometemos errores de tipo II.

Algoritmo 1 Método de Bonferroni con p-valores ajustados

Input: m, $P_n(j) \quad j = 1,\ldots,m$
$\widetilde{P}_{n,bonf}(j) = \min\{mP_n(j), 1\}$
Output: $\widetilde{P}_{n,bonf}(j) \quad j = 1,\ldots,m$

Finalmente, aplicamos la regla de decisión de rechazar las hipótesis H_{0j} con $\widetilde{P}_{n,bonf}(j) \leq \alpha$.

2.1.2. Método de Holm

El método de Holm [6] es un procedimiento *step-down* basado en el clásico método de Bonferroni, y al igual que éste, es aplicable en cualquier situación, sea cual sea la estructura de dependencia de los p-valores. Para controlar el FWER a nivel α, el método de Holm propone lo siguiente:

Sean $\mathscr{O}_n(j)$ los índices de los p-valores sin ajustar ordenados de menor a mayor, de modo que $P_n(\mathscr{O}_n(1)) \leq \ldots \leq P_n(\mathscr{O}_n(m))$, con sus respectivas hipótesis nulas asociadas $H_{0\mathscr{O}_n(j)}$. Entonces la hipótesis $H_{0\mathscr{O}_n(j)}$ es rechazada si su p-valor sin ajustar correspondiente cumple $P_n(\mathscr{O}_n(j)) \leq \dfrac{\alpha}{m-j+1}$. Además,

una vez llegamos a no rechazar una de las hipótesis nulas, las hipótesis nulas posteriores tampoco son rechazadas.

Por lo tanto, podemos expresar los p-valores ajustados según el método de Holm de la siguiente manera:

$$\widetilde{P}_{n,holm}(\mathscr{O}_n(j)) = \max_{k=1,\dots,j} \{\min\{(m-k+1)P_n(\mathscr{O}_n(k)), 1\}\} \tag{2.4}$$

Veamos que efectivamente este método tiene un control fuerte del FWER a nivel α.

Teorema 2.2. *El método de Holm dado en* (2.4) *controla el FWER a nivel* α.

Demostración. Sea I_0 el conjunto de indices correspondientes a las hipótesis nulas verdaderas. Sabemos que $|I_0| = m_0$.

Veamos que si rechazamos alguna hipótesis nula verdadera, entonces hay una hipótesis nula verdadera $H_{0\mathscr{O}_n(j)}$ para la cual $P_n(\mathscr{O}_n(j)) \leq \dfrac{\alpha}{m_0}$:

Primeramente, notar que como en este caso hay al menos una hipótesis nula verdadera, se tiene $m_0 \geq 1$. Sea j tal que $H_{0\mathscr{O}_n(j)}$ es la primera hipótesis nula verdadera rechazada, siendo el orden de las hipótesis definido para el método de Holm. Entonces, $H_{0\mathscr{O}_n(1)}, \dots, H_{0\mathscr{O}_n(j-1)}$ son todas hipótesis falsas rechazadas. Por lo tanto, $j-1 \leq m-m_0$, luego $\dfrac{1}{m-j+1} \leq \dfrac{1}{m_0}$. Como $H_{0\mathscr{O}_n(j)}$ es rechazada, tiene que ser $P_n(\mathscr{O}_n(j)) \leq \dfrac{\alpha}{m-j+1}$, por definición del método de Holm, y aplicando la desigualdad que acabamos de ver, se tiene $P_n(\mathscr{O}_n(j)) \leq \dfrac{\alpha}{m_0}$. Por tanto, se tiene que

$$\text{FWER} \leq \Pr\left\{\bigcup_{j\in I_0}\left(P_n(j) \leq \frac{\alpha}{m_0}\right)\right\} \leq \sum_{j\in I_0} \Pr\left\{P_n(j) \leq \frac{\alpha}{m_0}\right\} \leq m_0\frac{\alpha}{m_0} = \alpha$$

□

Proposición 2.3. *El método de Holm tiene una potencia mayor o igual que la del método de Bonferroni.*

Demostración. En el método de Bonferroni los p-valores sin ajustar deben cumplir $P_n(\mathscr{O}_n(j)) \leq \dfrac{\alpha}{m}$ para que la hipótesis correspondiente $H_{0\mathscr{O}_n(j)}$ sea rechazada (notar que hemos reordenado las hipótesis para hacerlas comparables, pues el orden no influye en el método de Bonferroni), mientras que en Holm, si queremos que esa hipótesis sea rechazada, se debe de cumplir $P_n(\mathscr{O}_n(j)) \leq \dfrac{\alpha}{m-j+1}$. Como se tiene que $m-j+1 \leq m$, entonces es más probable rechazar una hipótesis si utilizamos el método de Holm que si utilizamos el de Bonferroni, o lo que es lo mismo, es más probable no rechazar una hipótesis si utilizamos el método de Bonferroni que si utilizamos el de Holm. Notar que cada vez que no rechacemos una hipótesis utilizando el método de Holm, tampoco rechazamos esa hipótesis con el método de Bonferroni, pero no al revés. Suponer que $H_{0\mathscr{O}_n(j)}$ es falsa, entonces, por lo que hemos dicho antes, si el método de Holm no la rechaza, el método de Bonferroni tampoco lo hace, y en ambos casos estaremos cometiendo un error de tipo II. Sin embargo, también puede ser que el método de Bonferroni no la rechace mientras que el de Holm sí que lo haga, y en ese caso solo el método de Bonferroni comete un error de tipo II. □

El método de Holm es más potente que el de Bonferroni y ambos se pueden utilizar bajo las mismas hipótesis, por lo que generalmente se recomienda utilizar el método de Holm antes que el de Bonferroni.

De todos modos, otros métodos más potentes deben ser utilizados cuando la estructura de dependencia de los p-valores sea adecuada.

Algoritmo 2 Método de Holm con p-valores ajustados

Input: m, $P_n(j) \quad j = 1,\ldots,m$
Ordenar los p-valores sin ajustar: $P_n(\mathscr{O}_n(1)) \leq \ldots \leq P_n(\mathscr{O}_n(m))$
for $i = 1$ to m **do**
$\widetilde{P}_{n,holm}(\mathscr{O}_n(i)) = \text{mín}\{(m-i+1)P_n(\mathscr{O}_n(i)),1\}$
$\widetilde{P}_{n,holm}(\mathscr{O}_n(i)) = \text{máx}\left\{\widetilde{P}_{n,holm}(\mathscr{O}_n(h))\right\} \quad h = 1,\ldots,i$
end for
Output: $\widetilde{P}_{n,holm}(\mathscr{O}_n(j)) \quad j = 1,\ldots,m$

Finalmente, aplicamos la regla de decisión de rechazar las hipótesis $H_{0\mathscr{O}_n(j)}$ con $\widetilde{P}_{n,holm}(\mathscr{O}_n(j)) \leq \alpha$.

2.1.3. Método de Hommel

El método de Hommel [7] es un procedimiento *step-up*, es decir, que itera desde el p-valor sin ajustar más alto al más bajo. Este método tiene un control fuerte del FWER a nivel α, y se necesita la hipótesis de independencia de los p-valores, aunque funciona incluso con dependencia positiva [5, pág. 1957, 1958].

Este método tiene la ventaja de ser más potente que Bonferroni y Holm, pero el inconveniente es que necesitamos que se cumpla la hipótesis de independencia de los p-valores, o incluso podemos admitir dependencia positiva, por lo que solo debe ser utilizado por encima de estos métodos en el caso de que se cumpla una de estas hipótesis. El algoritmo clásico para calcular los p-valores ajustados según el método de Hommel viene dado en [8, pág. 1010], y es el siguiente:

Algoritmo 3 Método de Hommel con p-valores ajustados

Input: m, $P_n(j) \quad j = 1,\ldots,m$
Ordenar los p-valores sin ajustar: $P_n(\mathscr{O}_n(1)) \leq \ldots \leq P_n(\mathscr{O}_n(m))$
$\widetilde{P}_{n,hommel}(\mathscr{O}_n(j)) = P_n(\mathscr{O}_n(j)) \ \forall j = 1,\ldots,m$
for $i = m$ to 2 **do** (orden decreciente)
for $j = m-i+1$ to m **do**
$k_j = \dfrac{i * P_n(\mathscr{O}_n(j))}{i+j-m}$
end for
$k_{min} = \text{mín}\{k_j\} \quad j = m-i+1,\ldots,m$
for $j = m-i+1$ to m **do**
if $P_n(\mathscr{O}_n(j)) < k_{min}$ **then** $P_n(\mathscr{O}_n(j)) = k_{min}$
end if
end for
for $j = 1$ to $m-i$ **do**
$k_j = \text{mín}\{k_{min}, i * P_n(\mathscr{O}_n(j))\}$
if $\widetilde{P}_{n,hommel}(\mathscr{O}_n(j)) < k_j$ **then** $\widetilde{P}_{n,hommel}(\mathscr{O}_n(j)) = k_j$
end if
end for
end for
Output: $\widetilde{P}_{n,hommel}(\mathscr{O}_n(j)) \quad j = 1,\ldots,m$

Finalmente, aplicamos la regla de decisión de rechazar las hipótesis $H_{0\mathscr{O}_n(j)}$ con $\widetilde{P}_{n,hommel}(\mathscr{O}_n(j)) \leq \alpha$.

Este algoritmo es casi siempre demasiado costoso. Sin embargo, en [9] se proporciona un algoritmo de tiempo lineal para calcular los p-valores ajustados según el método de Hommel, basado en una generalización de la desigualdad de Simes, vista en el Anexo A, Sección A.0.1.

2.1.4. Software necesario para la implementación de estos métodos

La implementación de estos tres métodos se puede hacer con el software estadístico R. Para ello, está disponible la función `p.adjust()`, en el paquete del sistema `stats`. Dado un vector de p-valores sin ajustar, esta función nos devolverá los p-valores ajustados correspondientes al método indicado.

El esquema de uso es

```
p.adjust(p, method = p.adjust.methods, n = length(p))
```

donde los argumentos son:

- `p`: vector de p-valores sin ajustar.
- `method`: sirve para elegir uno de los distintos métodos que podemos utilizar para calcular los p-valores ajustados. Los tres que hemos visto hasta ahora están disponibles, aunque hay más. Basta con escribir escribir el deseado: `"bonferroni"` `"holm"` `"hommel"` .
- `n`: número de hipótesis que estamos contrastando. Por defecto, es la longitud del vector `p`. En nuestro caso, siempre será así.

La función devuelve un vector de tamaño `n` con los p-valores ajustados correspondientes.

2.2. Métodos para controlar el FDR

El concepto de FDR fue introducido por Benjamini y Hochberg en 1995 [10], y su definición ya ha sido dada en la Definición 5. La necesidad de su introducción es debida a que el FWER presentaba algunos problemas que con el FDR se resolvían. Éstos son dos principalmente:

- Los métodos clásicos para controlar el FWER, Bonferroni, Holm y Hommel, tienden a tener poca potencia.
- Dependiendo del contexto, el control del FWER no es realmente necesario. El control del FWER es importante cuando una decisión depende de si al menos una de las hipótesis nulas a contrastar es cierta o no. Por ejemplo, supongamos que tenemos muchos nuevos tratamientos para una enfermedad compitiendo contra un tratamiento estándar conocido. Entonces, lo que nos interesa es que al menos uno de ellos sea significativamente mejor que el tratamiento estándar.

Sin embargo, el FDR es preferible en otras situaciones. Por ejemplo, en la situación experimental descrita en el Capítulo 1, si queremos comparar varios parámetros entre los dos grupos para ver si un tratamiento tiene efectos positivos o no, la conclusión de si ese tratamiento es efectivo no tiene porqué ser errónea aunque alguna de las hipótesis verdaderas sea falsamente rechazada.

Esto no quiere decir que el FDR sea siempre preferible al FWER, pues esto dependerá del experimento que estemos realizando, como ya hemos ilustrado. Recordar que

$$\text{FDR} = E\left[\frac{V_n}{R_n} | R_n > 0\right] \Pr(R_n > 0)$$

por lo que en el caso de que no haya rechazos, $R_n = 0$ y $\text{FDR} = 0$, mientras que si sí que hay rechazos, $R_n > 0$ y $\text{FDR} = E\left[\frac{V_n}{R_n}\right]$.

Veamos dos importantes consecuencias de la definición de FDR:

Proposición 2.4. *Si todas las hipótesis nulas son ciertas ($m_0 = m$), el FDR es equivalente al FWER. Es decir, control débil del FDR implica control débil del FWER y viceversa.*

Demostración. Como todas las hipótesis nulas son ciertas, entonces todas las hipótesis nulas rechazadas serán hipótesis nulas ciertas, es decir, $V_n = R_n$.

Si $V_n = 0$, tenemos $R_n = 0$, y hemos definido que en este caso FDR$= 0$. Además, $\Pr(V_n \geq 1) = 0$, y por la definición de FWER (Definición 3), tenemos también FWER$= 0$.

Si $V_n > 0$, entonces $\frac{V_n}{R_n} = 1$ y FDR$= E\left[\frac{V_n}{R_n}\right] = 1$. Además, $\Pr(V_n \geq 1) = 1$, y de nuevo por la Definición 3 tenemos FWER$= 1$. □

Proposición 2.5. *Si hay al menos una hipótesis nula falsa ($m_0 < m$), entonces FDR$\leq$FWER.*

Demostración. Recordar que R_n y V_n son variables aleatorias. Denotaremos con r y v a los valores que toman en una realización del experimento. Sabemos que en este caso, si $v > 0$, entonces $\frac{v}{r} \leq 1$, por lo tanto $I_{\{V_n \geq 1\}} \geq \frac{V_n}{R_n}$, donde I es la función característica. Tomando esperanzas, se tiene $\Pr(V_n \geq 1) \geq E\left[\frac{V_n}{R_n}\right]$, luego FDR$\leq$FWER, basta aplicar las Definiciones 3 y 5. Para el caso en el que FDR$= 0$ el resultado es trivial. □

Observación: Como consecuencia de las dos proposiciones anteriores, cualquier método que controle el FWER también controla el FDR. Sin embargo, también puede ser que el método no controle el FWER y sí controle el FDR. En este caso, el método es menos exigente que si controlara ambos, por lo que es previsible que ganemos potencia si estamos en esta situación.

2.2.1. Método de Benjamini-Hochberg

El método de Benjamini-Hochberg [10] es un método *step-up* que controla el FDR a nivel α. Para ello, se necesita la hipótesis de que los p-valores sin ajustar sean independientes, aunque puede funcionar también cuando la dependencia es positiva. El método consiste en considerar los p-valores sin ajustar ordenados, $P_n(\mathscr{O}_n(1)) \leq \ldots \leq P_n(\mathscr{O}_n(m))$, con sus respectivas hipótesis nulas asociadas $H_{0\mathscr{O}_n(j)}$. Definimos entonces

$$k = \max\left\{i = 1,\ldots,m \mid P_n(\mathscr{O}_n(i)) \leq \frac{i}{m}\alpha\right\} \tag{2.5}$$

Finalmente, rechazamos todas las hipótesis $H_{0\mathscr{O}_n(j)}$ con $j = 1,\ldots,k$. Si tal k no existe, no rechazamos ninguna hipótesis.

Para comprobar que este método controla el FDR a nivel α para cualquier combinación de hipótesis nulas falsas y verdaderas, y para p-valores independientes, utilizamos el siguiente lema, cuya demostración se puede encontrar en [10].

Lema 2.6. *Para cualquier m_0, número de hipótesis nulas verdaderas, con $0 \leq m_0 \leq m$, tales que sus p-valores sin ajustar correspondientes sean independientes, y para cualquier valor que los $m_1 = m - m_0$ p-valores sin ajustar puedan tomar, el método de Benjamini-Hochberg definido en* (2.5) *cumple la desigualdad*

$$E\left[\frac{V_n}{R_n} | P_n(m_0+1) = p_1,\ldots,P_n(m) = p_{m_1}\right] \leq \frac{m_0}{m}\alpha \tag{2.6}$$

Donde hemos ordenado las hipótesis nulas de modo que primero consideramos las m_0 primeras que son verdaderas y después las m_1 siguientes que son falsas. Notar que $P_n(m_0+1),\ldots,P_n(m)$ son los p-valores sin ajustar correspondientes a las hipótesis nulas falsas, que recordemos que son variables aleatorias, mientras que $p_1,\ldots,p_{m_1}$ son los valores que toman estas variables aleatorias.

Teorema 2.7. *El método de Benjamini-Hochberg descrito en* (2.5) *controla el FDR a nivel α para cualquier combinación de hipótesis nulas falsas y verdaderas, y para p-valores independientes.*

Demostración. Sea cual sea la distribución de los p-valores correspondientes a las hipótesis nulas falsas, tomando esperanzas en los p-valores de las hipótesis nulas falsas en la desigualdad (2.6), se tiene $\mathrm{E}\left[\frac{V_n}{R_n}\right] \leq \frac{m_0}{m}\alpha \leq \alpha$ □

Recordar que hemos definido $\pi_0 = \frac{m_0}{m}$, luego podemos decir que el método de Benjamini-Hochberg controla el FDR a nivel $\pi_0\alpha$, siendo π_0 la proporción de hipótesis nulas verdaderas. Si esta proporción es mucho menor que 1, se tratará de un método conservador.

Algoritmo 4 Método de Benjamini-Hochberg con p-valores ajustados

Input: m, $P_n(j) \quad j = 1,\ldots,m$
Ordenar los p-valores sin ajustar: $P_n(\mathscr{O}_n(1)) \leq \ldots \leq P_n(\mathscr{O}_n(m))$
for $i = m$ to 1 **do** (orden decreciente)
$\widetilde{P}_{n,BH}(\mathscr{O}_n(i)) = \text{mín}\left\{\frac{m}{i}P_n(\mathscr{O}_n(i)), 1\right\}$
$\widetilde{P}_{n,BH}(\mathscr{O}_n(i)) = \text{mín}\left\{\widetilde{P}_{n,BH}(\mathscr{O}_n(h))\right\} \quad h = i,\ldots,m$
end for
Output: $\widetilde{P}_{n,BH}(\mathscr{O}_n(j)) \quad j = 1,\ldots,m$

Finalmente, aplicamos la regla de decisión de rechazar las hipótesis $H_{0\mathscr{O}_n(j)}$ con $\widetilde{P}_{n,BH}(\mathscr{O}_n(j)) \leq \alpha$.

2.2.2. Método de Benjamini-Yekutieli

El método de Benjamini-Yekutieli [11] es un método *step-up* que controla el FDR a nivel α. A diferencia del método de Benjamini-Hochberg, este método es válido bajo cualquier estructura de dependencia de los p-valores sin ajustar. El precio a pagar por ello es que este método es en general menos potente que el de Benjamini-Hochberg, por lo que el método de Benjamini-Yekutiely solo debe usarse en su lugar en el caso en que no podamos utilizar el método de Benjamini-Hochberg.

Al igual que en el método de Benjamini-Hochberg, el FDR es controlado realmente a nivel $\pi_0\alpha$.

El método consiste en considerar los p-valores sin ajustar ordenados, $P_n(\mathscr{O}_n(1)) \leq \ldots \leq P_n(\mathscr{O}_n(m))$, con sus respectivas hipótesis nulas asociadas $H_{0\mathscr{O}_n(j)}$. Definimos entonces

$$k = \text{máx}\left\{i = 1,\ldots,m \mid P_n(\mathscr{O}_n(i)) \leq \frac{i}{m \cdot c(m)}\alpha\right\}$$

Donde $c(m) = \sum_{i=1}^{m} \frac{1}{i}$

Finalmente, rechazamos todas las hipótesis $H_{0\mathscr{O}_n(j)}$ con $j = 1,\ldots,k$. Si tal k no existe, no rechazamos ninguna hipótesis.

Notar que $c(m)$ puede aproximarse usando la fórmula de Euler-Maclaurin y la constante de Euler-Mascheroni ($\gamma = 0{,}57721\ldots$) [12]

$$c(m) = \sum_{i=1}^{m} \frac{1}{i} \approx \ln(m) + \gamma + \frac{1}{2m}$$

Algoritmo 5 Método de Benjamini-Yekutieli con p-valores ajustados

Input: m, $P_n(j) \quad j = 1,\ldots,m$
Ordenar los p-valores sin ajustar: $P_n(\mathscr{O}_n(1)) \leq \ldots \leq P_n(\mathscr{O}_n(m))$
for $i = m$ to 1 **do** (orden decreciente)
$\widetilde{P}_{n,BY}(\mathscr{O}_n(i)) = \text{mín}\left\{\frac{m \cdot c(m)}{i} P_n(\mathscr{O}_n(i)), 1\right\}$
$\widetilde{P}_{n,BY}(\mathscr{O}_n(i)) = \text{mín}\left\{\widetilde{P}_{n,BY}(\mathscr{O}_n(h))\right\} \quad h = i,\ldots,m$
end for
Output: $\widetilde{P}_{n,BY}(\mathscr{O}_n(j)) \quad j = 1,\ldots,m$

Finalmente, aplicamos la regla de decisión de rechazar las hipótesis $H_{0\mathscr{O}_n(j)}$ con $\widetilde{P}_{n,BY}(\mathscr{O}_n(j)) \leq \alpha$.

2.2.3. Método BKY

El método BKY, desarrollado por Benjamini, Krieger y Yekutieli [13] está basado en hacer dos veces el método de Benjamini-Hochberg. Es un método que se denomina *adaptativo*, que quiere decir que es un método que estima m_0 en un primer paso con el objetivo de ganar potencia con respecto al método de Benjamini-Hochberg. Pero, ¿qué ventaja tiene conocer m_0 o una estimación suya?

Supongamos que conocemos el valor de m_0. Sabemos que el método de Benjamini-Hochberg controla el FDR a nivel $\frac{m_0}{m}\alpha$., entonces podemos aplicar el método de Benjamini-Hochberg con $\alpha' = \frac{m}{m_0}\alpha$ y de este modo podemos controlar el FDR a nivel $\frac{m_0}{m}\frac{m}{m_0}\alpha = \alpha$, que es lo que queremos. Como $\alpha \leq \frac{m}{m_0}\alpha = \alpha'$, podemos rechazar más hipótesis, por lo que tenemos más potencia, mientras que seguimos controlando el FDR a nivel α.

Para aplicar el método BKY para controlar el FDR a nivel α, se deben seguir los siguientes pasos:

- *Paso 1:* Aplicar el método de Benjamini-Hochberg a nivel $\alpha' = \frac{\alpha}{1+\alpha}$. Sea r el número de hipótesis rechazadas. Si $r = 0$, entonces no rechazamos ninguna hipótesis y hemos terminado. Si $r = m$, entonces rechazamos las m hipótesis y hemos terminado. En otro caso, pasar al siguiente paso.
- *Paso 2:* Sea $\hat{m}_0 = m - r$, la estimación de m_0.
- *Paso 3:* Aplicar el método de Benjamini-Hochberg a nivel $\alpha^* = \alpha'\frac{m}{\hat{m}_0}$. Rechazamos las hipótesis resultantes de aplicar este método con α^*.

El método BKY controla el FDR a nivel α bajo la hipótesis de independencia o dependencia positiva, e incluso se mantiene conservador y con una potencia relativamente buena cuando el grado de dependencia es desconocido [14].

La demostración al hecho de que el método BKY controla el FDR a nivel α se encuentra en [13, pág. 498].

2.2.4. Software necesario para la implementación de estos métodos

Los dos primeros métodos descritos en esta sección, el de Benjamini-Hochberg y el de Benjamini-Yekutieli, están implementados en la función `p.adjust` que hemos definido en 2.1.4. Para obtener los p-valores ajustados correspondientes a estos dos métodos, la implementación es igual, basta con cambiar el segundo argumento a `"BH"` o `"BY"`.

En cambio, el método BKY no está implementado en la función `p.adjust`. Aun así, disponemos de otro paquete de R que nos permite implementar este método, el paquete `cp4p` [23]. En este paquete, la función `adjust.p()` nos permite utilizar el método BKY. Su esquema de uso es

```
adjust.p(p, pi0.method = "bky", alpha = 0.05)
```

donde p es el vector de p-valores sin ajustar.

La función devuelve una lista compuesta por `pi0`, una estimación de π_0, y `adjp`, una matriz de dos columnas con los p-valores sin ajustar y ajustados, donde la primera columna corresponde a los p-valores sin ajustar, la segunda a los ajustados y cada fila corresponde a un test. Para seleccionar los p-valores ajustados por el método BKY, que es lo que nos interesa, basta con poner

```
adjust.p(p, pi0.method = "bky", alpha = 0.05)$adjp[,2]
```

2.3. Control del FDP

Primeramente, notar que, aplicando las Definiciones 4 y 5, la cantidad FRD es la esperanza de la variable FDP, es decir, FDR $= E[\text{FDP}]$.

Se podría pensar que controlar la variable FDP$= \dfrac{V_n}{R_n}$ es más relevante que controlar la cantidad FDR, puesto que FDP está directamente relacionado con el experimento. Sin embargo, notar que en el caso de que todas las hipótesis nulas sean verdaderas $(m = m_0)$, se tiene que con que tengamos un solo rechazo, se cumple $\dfrac{V_n}{R_n} = 1$, ya que todos los rechazos son falsos rechazos, luego FDP$= 1$, por lo que el FDP no se puede controlar. Lo mismo ocurre con $\left(\dfrac{V_n}{R_n}|R_n > 0\right)$, por lo que $E\left[\dfrac{V_n}{R_n}|R_n > 0\right]$ tampoco puede ser controlado. Sin embargo, como hemos visto en los diversos métodos que hemos presentado, FDR$= E\left[\dfrac{V_n}{R_n}|R_n > 0\right]\Pr(R_n > 0)$ sí que puede ser controlado a nivel α.

En el caso $m_0 = m$, la variable FDP es de tipo binario, pues solo puede tomar los valores FDP$= 0$ (si $R_n = 0$, por la Definición 4) o FDP$= 1$ (si $R_n > 0$, como hemos dicho en el párrafo anterior). Pese a que el FDP solo pueda tomar los valores 0 y 1, su esperanza, que es el FDR, sí que puede ser controlado a nivel α. Basta con aplicar un método en el que la probabilidad de que rechacemos al menos una hipótesis sea menor o igual que α. De este modo, el FDR vale a lo sumo FDR $= E[\text{FDP}] = 0 * \Pr(\text{FDP} = 0) + 1 * \Pr(\text{FDP} = 1) = 0 * (1-\alpha) + 1 * \alpha = \alpha$. Cualquiera de los métodos descritos anteriormente cumple esto, pues ya hemos visto que controlaban el FDR a nivel α si $\pi_0 = 1$.

2.3.1. Estimación del FDP y q-valores

Definimos un nuevo concepto directamente relacionado con el FDP.

Definición 13. *Positive false discovery rate* (pFDR), es la esperanza de la proporción entre los falsos rechazos, V_n y el total de rechazos, R_n, condicionado a que ha ocurrido al menos un rechazo,

$$\text{pFDR} = E\left[\frac{V_n}{R_n}|R_n > 0\right]$$

Hemos introducido esta definición debido a que el pFDR puede ser considerado como un estimador del FDP [5, pág. 1965].

Notar que si m es grande, como suele ocurrir en este contexto, $\Pr(R_n > 0) \approx 1$, luego pFDR$\approx$FDR en este caso. Sin embargo, si m no es demasiado grande, puede ser que $R_n = 0$. En este caso, pFDR no está definido, pero si ocurre esto sustituiremos $R_n = 0$ por $R_n = 1$.

Como hemos comprobado en la Sección 2.3, el pFDR tampoco puede ser controlado a nivel α, pues si todas las hipótesis nulas son verdaderas vale siempre 1. Como no puede ser controlado, los p-valores ajustados tampoco pueden ser definidos formalmente. Es por ello que se definen los q-valores de la siguiente forma [15, pág. 9443].

Definición 14. El q-valor para la hipótesis H_{0j} es

$$Q_j = \min_{t \geq P_n(j)} \text{pFDR}(t)$$

donde $\text{pFDR}(t) = E\left[\frac{V_n(t)}{R_n(t)} \mid R_n(t) > 0\right]$, siendo $V_n(t)$ el número de hipótesis nulas verdaderas rechazadas con nivel de significación t y $R_n(t)$ el número total de hipótesis nulas rechazadas con nivel de significación t, utilizando los p-valores sin ajustar.

Recordar que, por (A.2), el p-valor sin ajustar es el nivel de significación más pequeño para el cual una hipótesis es rechazada. Una hipótesis es rechazada si los valores observados pertenecen a la región de rechazo. El uso de los p-valores sin ajustar nos permite controlar el error de tipo I en test de hipótesis simples.

Ahora, en lugar de querer controlar el error de tipo I queremos controlar el pFDR. Además, podemos considerar que las regiones de rechazo están dadas en función de los p-valores sin ajustar, por lo que son de la forma $[0,t]$, de modo que si $P_n(j) \in [0,t]$ se rechaza la hipótesis H_{0j}. Por tanto, se pueden definir análogamente los q-valores Q_j como el mínimo valor de $\text{pFDR}(t)$, con $P_n(j) \leq t$. Los q-valores nos permitirán controlar el pFDR.

2.3.2. Método de Storey

El método de Storey [15] controla el pFDR *asintóticamente* a nivel α. Con asintóticamente nos referimos a que $\text{pFDR} \leq \alpha$ cuando $m \to \infty$. En los experimentos reales, no podemos asegurar $\text{pFDR} \leq \alpha$, pero como m suele ser grande, el valor se acerca mucho a α.

Este método asume que los p-valores sin ajustar son independientes, aunque funciona bien si la dependencia es débil. Si la dependencia es más fuerte, como suele ocurrir en los experimentos, el pFDR puede tener demasiada varianza, asimetría y sesgo [16] (recordar que pFDR es un estimador de FDP).

Es un método que es bastante potente, aunque el precio a pagar por ello es que no podemos asegurar $\text{pFDR} \leq \alpha$, como hemos dicho.

El método consiste en calcular los q-valores Q_j correspondientes a cada hipótesis H_{0j} y rechazar las que cumplen $Q_j \leq \alpha$. La demostración de que este método controla el pFDR asintóticamente a nivel α se puede encontrar en [17, pág. 194, 195].

Sin embargo, calcular el valor exacto de los q-valores es muy complicado. Por ello, utilizaremos un valor aproximado de éstos al que llamamos $\widehat{Q}_j$ y aplicamos el método con este valor. Esta aproximación de lo q-valores converge a éstos, y por tanto el método sigue siendo válido (ver [17, pág. 196]).

Para calcular los q-valores aproximados, aplicando su definición, necesitamos una aproximación de $\text{pFDR}(t) = E\left[\frac{V_n(t)}{R_n(t)} | R_n(t) > 0\right]$. Como m es muy grande, se puede demostrar que $\text{pFDR}(t) \approx \frac{E[V_n(t)]}{E[R_n(t)]}$ [18, pág. 2019]. Una buena estimación de $E[R_n(t)]$ es $R_n(t) = \#\{P_n(j) \leq t\}$. Por otro lado, recordar que $V_n(t) = \#\{P_n(j) \leq t \mid H_{0j} \text{ verdadera}\}$. Sabemos que los p-valores correspondientes a una hipótesis nula verdadera siguen una distribución uniforme, es decir, se cumple (2.2) (hemos dado las condiciones para esto en el Teorema A.1), por lo que tenemos $E[V_n(t)] = m_0 t = m\pi_0 t$. Una buena estimación de π_0 es $\hat{\pi}_0(\lambda) = \frac{\#\{P_n(j) > \lambda\}}{m(1-\lambda)}$, donde hemos introducido el parámetro $\lambda \in (0,1)$. La justificación de esta estimación de π_0 es la siguiente:

Suponer en primer lugar que tenemos 10000 hipótesis independientes a contrastar, y todas ellas son ciertas ($m = m_0$). Como sabemos, los p-valores sin ajustar son uniformes bajo la hipótesis nula, como se demuestra en el Teorema A.1.

Si, por otro lado, tenemos 10000 hipótesis independientes a contrastar, todas ellas falsas ($m = m_1$), los p-valores están concentrados en valores pequeños, puesto que rechazamos la mayoría de las hipótesis. Notar que en este caso prácticamente no hay valores altos.

Veamos las dos situaciones:

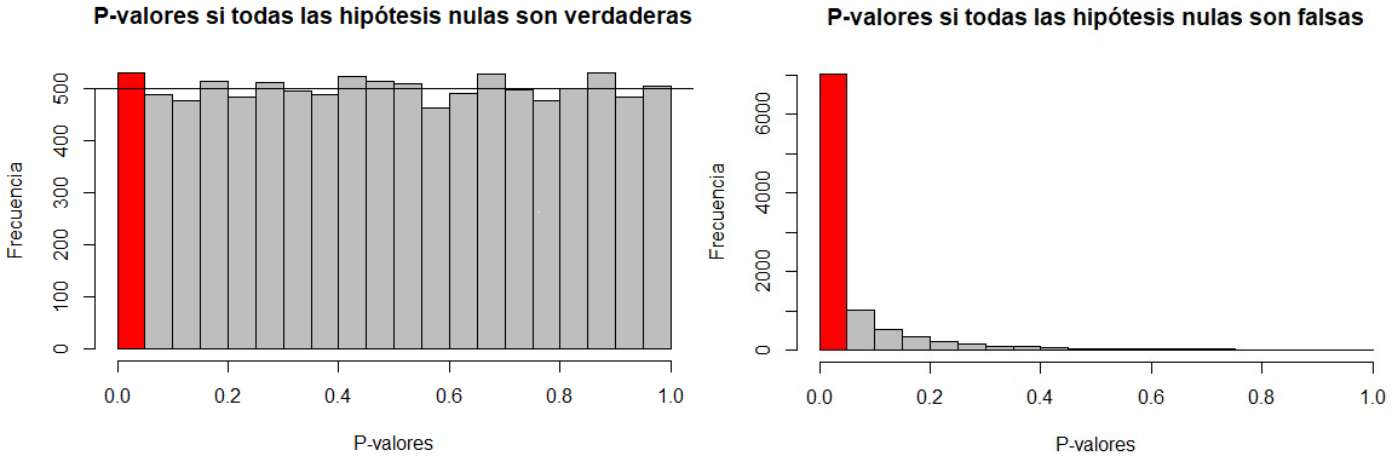

Figura 2.1: Ejemplo de histograma de P-valores si todas las hipótesis nulas son verdaderas o falsas.

Sin embargo, cuando estamos realizando un experimento, no sabemos qué hipótesis nulas son verdaderas o falsas, pues es lo que tratamos de estimar. Por lo tanto, tenemos una mixtura de los dos histogramas anteriores, y obtenemos algo parecido a lo siguiente:

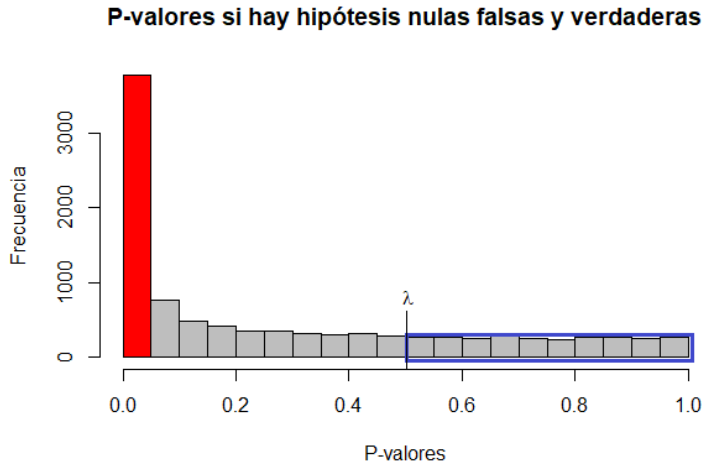

Figura 2.2: Ejemplo de histograma de P-valores si tenemos hipótesis nulas verdaderas y falsas.

Estos histogramas son de elaboración propia, y están basados en [15, fig. 1].

Notar que la última parte del histograma es plana, pues esto se debe a que los p-valores altos corresponden prácticamente todos ellos a hipótesis nulas verdaderas, que son uniformes, ya que no hay p-valores altos correspondientes a hipótesis nulas falsas que modifiquen la uniformidad.

Por tanto, si queremos estimar el número de hipótesis nulas verdaderas, hay que estimar la altura del rectángulo azul de la figura. Para ello, hay que contar el número de p-valores que corresponden a la parte horizontal del histograma. Suponer que a partir de λ, el histograma es plano, entonces la estimación del número de hipótesis nulas verdaderas es $\dfrac{\#\{P_n(j) > \lambda\}}{1-\lambda}$. Para estimar π_0 basta dividir entre m, y obtenemos finalmente la estimación de la proporción de hipótesis nulas verdaderas, $\hat{\pi}_0(\lambda) = \dfrac{\#\{P_n(j) > \lambda\}}{m(1-\lambda)}$.

La elección de $\lambda = 0{,}5$ da resultados razonablemente buenos en la mayoría de los casos, pero lo que se propone en [15] es utilizar un spline cúbico natural con tres grados de libertad para aproximar $\hat{\pi}_0(\lambda)$

y evaluarlo en 1.

Introduciendo estas aproximaciones en el cálculo del pFDR nos queda $\widehat{\text{pFDR}}(t) = \dfrac{\hat{\pi}_0 mt}{\#\{P_n(j) \leq t\}}$. Por lo tanto, los q-valores aproximados son $\widehat{Q}_j = \underset{t \geq P_n(j)}{\text{mín}} \widehat{\text{pFDR}}(t) = \underset{t \geq P_n(j)}{\text{mín}} \dfrac{\hat{\pi}_0 mt}{\#\{P_n(j) \leq t\}}$

El algoritmo del método de Storey con q-valores aproximados queda por lo tanto de la siguiente manera:

Algoritmo 6 Método de Storey con q-valores aproximados

Input: m, $P_n(j) \quad j = 1, \ldots, m$

Ordenar los p-valores sin ajustar: $P_n(\mathscr{O}_n(1)) \leq \ldots \leq P_n(\mathscr{O}_n(m))$

Para un rango de λ, $\lambda = 0;\ 0{,}01;\ 0{,}02; \ldots; 0{,}95$ calcular $\hat{\pi}_0(\lambda) = \dfrac{\#\{P_n(j) > \lambda\}}{m(1-\lambda)}$

Sea $\hat{f}$ el spline cubico natural con 3 grados de libertad de $\hat{\pi}_0(\lambda)$

Calculamos el estimador de π_0 como $\hat{\pi}_0 = \hat{f}(1)$

Sea $\widehat{Q}_{\mathscr{O}_n(m)} = \underset{t \geq P_n(\mathscr{O}_n(m))}{\text{mín}} \left\{ \dfrac{\hat{\pi}_0 mt}{\#\{P_n(j) \leq t\}} \right\} = \hat{\pi}_0 P_n(\mathscr{O}_n(m))$

for $i = m-1$ to 1 **do** (orden decreciente)

$\widehat{Q}_{\mathscr{O}_n(i)} = \underset{t \geq P_n(\mathscr{O}_n(i))}{\text{mín}} \left\{ \dfrac{\hat{\pi}_0 mt}{\#\{P_n(j) \leq t\}} \right\} = \text{mín} \left\{ \dfrac{\hat{\pi}_0 m P_n(\mathscr{O}_n(i))}{i}, \widehat{Q}_{\mathscr{O}_n(i+1)} \right\}$

end for

Output: $\widehat{Q}_{\mathscr{O}_n(j)} \quad j = 1, \ldots, m$

Finalmente, aplicamos la regla de decisión de rechazar las hipótesis $H_{0\mathscr{O}_n(j)}$ con $\widehat{Q}_{\mathscr{O}_n(j)} \leq \alpha$.

Para implementar el método de Storey para calcular los q-valores aproximados en el software estadístico R, utilizamos la función `qvalue()` del paquete con el mismo nombre `qvalue` [24]. Su esquema de uso es

```
qvalue(p)
```

donde `p` es la lista de p-valores sin ajustar. La función devuelve una lista con varios elementos, de los cuales el que nos interesa es el llamado `qvalues`. Por tanto, para seleccionar los q-valores aproximados hacemos

```
qvalue(p)$qvalues
```

Capítulo 3

Estudio de simulación e implementación de los distintos métodos

En este capítulo simulamos datos correspondientes a la situación experimental que hemos puesto de ejemplo en el Capítulo 1, en la que había dos grupos, uno de control y otro de casos. Simulamos la expresión de cada gen para cada individuo, y comparamos su media entre los dos grupos, para contrastar si son iguales o no. Utilizamos los distintos métodos descritos hasta ahora para ajustar los p-valores y contrastar todas las hipótesis, y comprobamos que se cumplen las características de cada uno, y los comparamos entre ellos.

Suponemos en todas la simulaciones que las varianzas entre cada par de muestras son iguales, y fijamos $n_1 = n_2 = 20$, por lo que $n = 40$, es decir, hay 20 individuos en el grupo de casos y otros 20 en el grupo de control. Estudiamos las expresiones de 1000 genes, por lo que contrastamos 1000 hipótesis ($m = 1000$). El nivel de significación será de ahora en adelante $\alpha = 0{,}05$. Además, las diferencias de medias en las hipótesis nulas falsas están fijadas para que la potencia individual de cada test sea de 0,8.

Además, podemos cambiar la estructura de dependencia de los p-valores y la proporción de hipótesis verdaderas π_0, y ver cómo se comporta cada método según cómo varían estos parámetros. Sin embargo, solo estudiaremos el caso $\pi_0 = 0{,}5$ por falta de espacio.

Para la simulación de datos y el control de los parámetros utilizamos el software *Myriads* [19]. Este software nos devuelve los datos en un formato adecuado para luego poder importarlos desde R y poder calcular los p-valores ajustados con los distintos métodos ya descritos.

Para ejemplificar nuestro método de trabajo, simulamos la situación en la que los p-valores son independientes y $\pi_0 = 0{,}5$. Los datos simulados por Myriads tienen el siguiente formato:

	X_1	$\cdots$	X_{20}	X_{21}	$\cdots$	X_{40}
Gen 1	-0.386832	$\cdots$	-0.266555	-0.745515	$\cdots$	-0.123823
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Gen 1000	1.498470	$\cdots$	-0.586187	-1.514290	$\cdots$	0.040948

Tabla 3.1: Expresión de cada gen para cada individuo

El código utilizado a lo largo de todo el procedimiento descrito a continuación se encuentra en el Anexo B. Importamos los datos de la tabla en R, y a partir de estos datos obtenemos un vector de 1000 p-valores sin ajustar. Para ello, tenemos que aplicar un *t-test* para cada fila de la tabla, en el que comparamos las igualdad de medias entre los dos grupos. A partir de ese vector de p-valores sin ajustar, aplicamos las funciones correspondientes a los distintos métodos, que hemos descrito en las Secciones 2.1.4, 2.2.4 y al final de 2.3.2. Tenemos ahora un vector con 1000 p-valores ajustados (o q-valores) para cada método.

Gracias al modo en el que funciona Myriads, sabemos que las hipótesis nulas falsas están colocadas al principio, y las verdaderas al final. Como en este ejemplo $\pi_0 = 0{,}5$, tenemos que las primeras 500 hipótesis nulas son falsas y las 500 siguientes verdaderas. Por lo tanto, podemos conocer el número de

hipótesis falsas que son rechazadas ($m_1 - U_n$), el número de hipótesis falsas no rechazadas (U_n: número de errores de tipo II), el número de hipótesis verdaderas que son rechazadas (V_n: número de errores de tipo I) y el número de hipótesis verdaderas no rechazadas ($m_0 - V_n$). Recordar la Tabla 1.1 en la que describimos esta situación. La siguiente tabla recoge las cantidades que acabamos de describir para los p-valores sin ajustar y para los métodos estudiados en el Capítulo 2:

	Hipótesis nulas falsas		Hipótesis nulas verdaderas	
	Rechazadas	No rechazadas	Rechazadas	No rechazadas
P-valores sin ajustar	398	102	26	474
Bonferroni	28	472	0	500
Holm	29	471	0	500
Hommel	30	470	0	500
BH	309	191	6	494
BY	125	375	1	499
BKY	354	146	9	491
Q-valores	375	125	16	484

Tabla 3.2: Número de rechazos para cada método

Notar que esta tabla se tiene para una sola simulación en la que $\pi_0 = 0{,}5$ y los p-valores son independientes. Nuestro objetivo es hacer 100 simulaciones como la descrita anteriormente y con las 100 tablas obtenidas calcular estimadores del FWER, del FDR y de la potencia para cada método.

Para la estimación del FWER lo que hacemos es contar el número de veces que $V_n \geq 1$ para cada método y dividir el resultado entre el total de simulaciones, 100. Lo denotamos por $\widehat{\text{FWER}}$.

Para la estimación del FDR lo que hacemos es obtener para cada método el valor $\frac{V_n}{R_n}$ y hacemos la media entre los valores de todas las tablas. Lo denotamos por $\widehat{\text{FDR}}$.

Para la estimación de la potencia lo que hacemos es obtener para cada método el valor $\frac{m_1 - U_n}{m_1}$ (Definición 2) y hacemos la media entre los valores de todas las tablas. Lo denotamos por $\widehat{\text{Potencia}}$.

Por tanto, tenemos ahora un estimador del FWER, del FDR y de la potencia para cada método en el caso en el que $\pi_0 = 0{,}5$ y los p-valores son independientes. Haciendo 100 simulaciones con estos parámetros y calculando los estimadores del FWER, del FDR y de la potencia como hemos explicado, quedan así:

	$\widehat{\text{FWER}}$	$\widehat{\text{FDR}}$	$\widehat{\text{Potencia}}$
P-valores sin ajustar	1	0.0610403211	0.78160
Bonferroni	0.03	0.0008629501	0.06128
Holm	0.03	0.0008442609	0.06238
Hommel	0.03	0.0008228359	0.06632
BH	1	0.0254878603	0.61142
BY	0.4	0.0045104643	0.22010
BKY	1	0.0366233288	0.68570
Q-valores	1	0.0524896303	0.75122

Tabla 3.3: Estimaciones del FWER, FDR y Potencia para cada método

Repitiendo el proceso desde el principio, podemos hacer lo mismo variando el grado de dependencia de los p-valores, y así ver cómo varían los estimadores según cambia esta variable. Myriads nos permite cambiar el grado de dependencia entre los individuos del grupo de control, ρ_1, y entre los individuos del grupo de casos, ρ_2. Lo que hacemos es fijar $\rho_1 = 0$ y dar distintos valores a ρ_2. Los valores de ρ_2 que estudiamos son 0, 0.1,...,0.9 (notar que el caso $\rho_1 = \rho_2 = 0$ es el que ya hemos dado de ejemplo). Repitiendo el proceso para estos valores de ρ_2 y dividiendo los datos en tres tablas distintas, una pa-

ra cada estimador que estamos estudiando, quedan los siguientes resultados. Para visualizarlos mejor, utilizaremos mapas de calor:

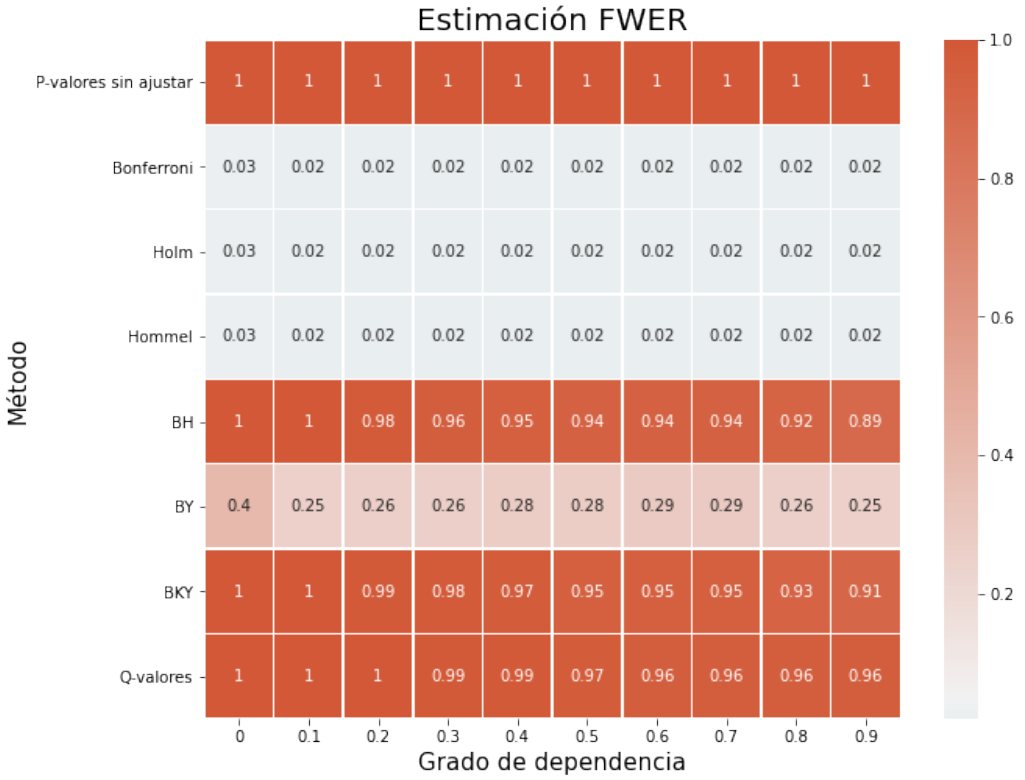

Figura 3.1: Estimación FWER

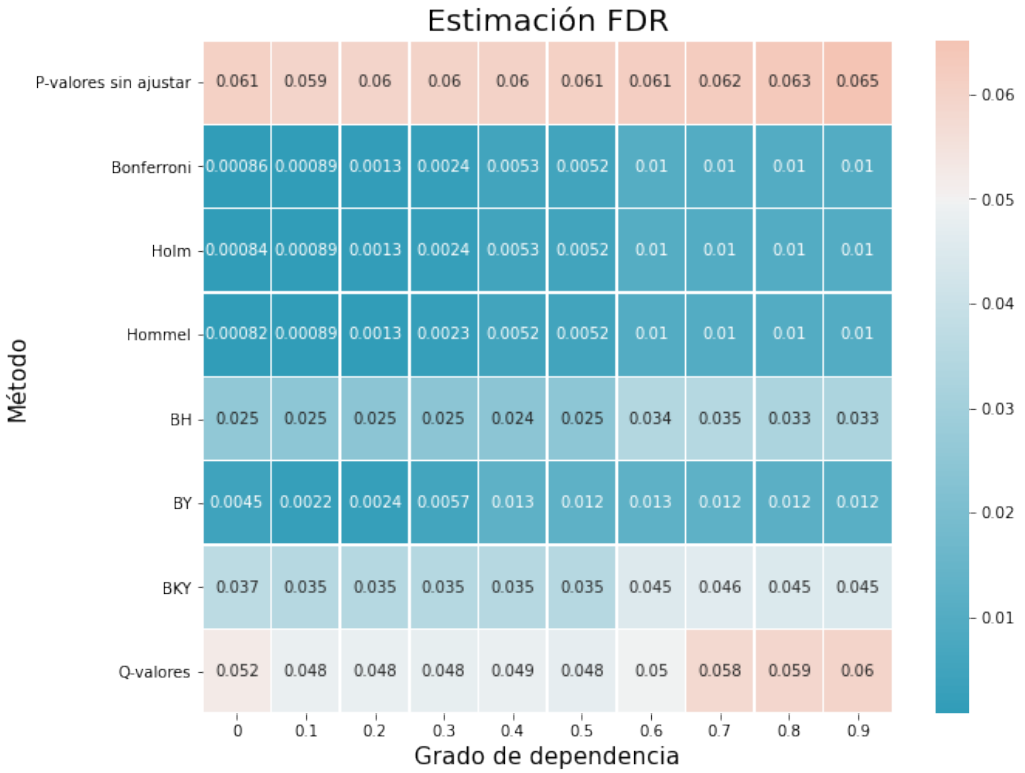

Figura 3.2: Estimación FDR

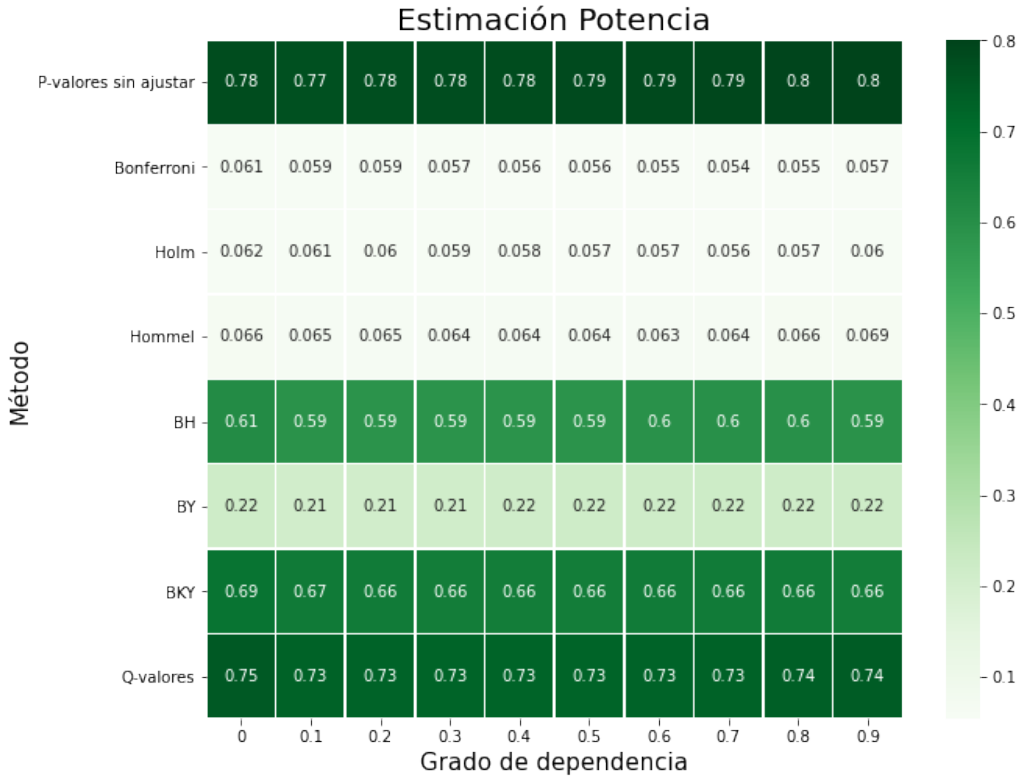

Figura 3.3: Estimación Potencia

Notar que hemos fijado desde el principio $\pi_0 = 0{,}5$, pero podríamos haber hecho el mismo análisis con cualquier otra proporción de hipótesis nulas verdaderas. No lo hacemos por falta de espacio.

Comentamos ahora los tres mapas de calor obtenidos y vemos si los resultados concuerdan con lo expuesto en la teoría del Capítulo 2.

Estimación del FWER y del FDR

En las Figuras 3.1 y 3.2, en las que vemos las distintas estimaciones del FWER y el FDR respectivamente, hemos elegido la escala de colores de modo que los valores mayores que 0,05 sean cada vez más rojos y los valores menores que 0,05 sean cada vez más azules. Gracias a esto, podemos ver fácilmente los métodos con los que controlamos el FWER o el FDR con nivel de significación 0,05.

El FWER solo es controlado por los métodos de Bonferroni, Holm y Hommel, independientemente del grado de dependencia. Notar que el método de Hommel teóricamente no controla el FWER si el grado de dependencia es muy alto, pero podemos ver que el método es lo suficientemente robusto y que lo controla igualmente. El resto de métodos, como era de esperar, no controlan el FWER a nivel 0,05.

El FDR es controlado por los métodos BH, BY y BKY, que lo controlan a nivel 0,05. Se podría pensar que los métodos BH y BKY no iban a controlar el FDR cuando el grado de dependencia fuera alto, como se dice en el Capítulo 2. Si bien el valor de la estimación del FDR crece en estos casos, no llega a superar el valor de 0,05, por lo que lo siguen controlando. Sin embargo, hay que tener mucha precaución a la hora de aplicar los métodos BH y BKY, sobre todo este último, si no conocemos el grado de dependencia, pues en el caso de que sea alto, el valor del FDR se acerca mucho a 0,05 y puede que en otro tipo de experimentos no sea controlado. Si se da el caso, habrá que usar el método BY, que controla el FDR bajo cualquier estructura de dependencia, como podemos ver.

En cuanto a los q-valores, como hemos visto en la Sección 2.3.2, el control del FDR a nivel 0,05 no está garantizado, pues esta es una propiedad que se cumple asintóticamente. Como $m = 1000$ es grande, los resultados obtenidos se aproximan mucho a este valor, excepto en el caso de que el grado de dependencia sea alto, como hemos dicho en la teoría. Como podemos ver, la estimación del FDR supera

claramente el valor 0,05 si el grado de dependencia es alto, por lo que no debemos usar este método si se da este caso.

Además, como hemos visto en la Proposición 2.5, se tiene FDR$\leq$FWER, por lo que los métodos que controlan el FWER (Bonferroni, Holm y Hommel) también controlan el FDR, pero no al revés.

Como era de esperar, los p-valores sin ajustar no controlan ni el FWER ni el FDR.

Estimación de la potencia

Para la estimación de la potencia, que podemos ver representada en la Figura 3.3, hemos elegido una escala de colores que va de verde menos intenso, para valores bajos de la potencia, a un verde más intenso, para valores altos de la potencia. Como se dice al principio de la Sección 2.2, los métodos que controlan el FDR son más potentes que los que controlan el FWER, como se puede ver en la figura. Además, como hemos dicho en la teoría, también podemos ver que el método de Hommel es más potente que el de Holm, y el de Holm es más potente que el de Bonferroni. Por otro lado, el método BKY es más potente que el BH, y el BH es más potente que el BY. El método de los q-valores es muy potente, pero recordar que el control del FDR no está garantizado si lo usamos, sobre todo en el caso de que el grado de dependencia sea alto.

Los p-valores sin ajustar son el método más potente, pero no controlan el FWER ni el FDR, por lo que no se usarán en test de hipótesis múltiples.

3.1. Conclusiones

Hemos visto las ventajas e inconvenientes de los distintos métodos que hemos dado para controlar el FWER y el FDR. Dependiendo de las características del experimento, es más adecuado el control de una cantidad o de otra.

Si estamos interesados en controlar el FWER, el método más potente es Hommel, por lo que es la opción más conveniente. Si el grado de dependencia es alto, no está asegurado el control del FWER, pero en nuestra simulación hemos visto que el método es lo suficientemente robusto y lo controlaba. Ahora bien, si queremos asegurarnos, es más conveniente usar el método de Holm, que, aunque sea algo menos potente, nos asegura el control del FWER independientemente de la estructura de dependencia de los p-valores.

Si en cambio estamos interesados en el control del FDR, el método que nos lo asegura, sea cual sea la estructura de dependencia de los p-valores, es el método BY. La desventaja de BY es que no es muy potente, por lo que otros métodos son preferibles en el caso de que se puedan utilizar. BH y BKY controlan el FDR cuando el grado de dependendencia no es excesivamente alto, como hemos comprobado, por lo que son preferibles. Hay que tener precaución sobre todo con BKY, que es el más potente de estos tres, ya que cuando el grado de dependencia es muy alto, puede que el valor del FDR supere 0,05. En cuanto a los q-valores, que son muy potentes, solo han de ser utilizados en el caso de que m sea un valor grande, del orden de miles, ya que el control del FDR es asintótico en este caso. Si el grado de dependencia de los p-valores es alto, no debemos utilizar los q-valores, ya que en este caso no tenemos control del FDR.

Para finalizar, señalar que en este trabajo hemos descrito varios de los principales métodos que hay para controlar el FWER y el FDR, aunque hay muchos más. El desarrollo de nuevos contrastes de hipótesis múltiples es un área de estudio relativamente reciente y activa, por lo que están apareciendo nuevos métodos de ajuste del p-valor continuamente. Con el creciente volumen de datos que se está manejando actualmente, los test de hipótesis múltiples y métodos de ajuste del p-valor se están utilizando y perfeccionando cada vez más.

Anexo A

Test de hipótesis simples

Las definiciones y resultados de este anexo están basados en [1] y [21]. Definiremos en este anexo los test de hipótesis simples, dando sus principales características y resultados. Comprender bien los test de hipótesis simples es necesario para luego poder definir y entender los test de hipótesis múltiples, que básicamente consisten en decidir sobre qué hipótesis rechazamos y cuáles no sobre un conjunto de hipótesis, a las que podemos aplicar la teoría de test de hipótesis simples que veremos a continuación.

Primeras nociones

Sea $\mathbf{x} = (X_1, \ldots, X_n)$ una muestra aleatoria simple de una población $X \sim P_\theta$ con $\theta \in \mathbb{R}$ desconocido.

En los **contrastes paramétricos**, la distribución poblacional es totalmente conocida salvo por el parámetro θ.

El **espacio paramétrico** $\Theta \subseteq \mathbb{R}$ (que contiene al menos dos puntos) se puede particionar como $\Theta = \Theta_1 \cup \Theta_2$, con $\Theta_1 \cap \Theta_2 = \emptyset$ $(\Theta_i \neq \emptyset, i = 0, 1)$.

El **espacio muestral**, $\mathscr{X} \subset \mathbb{R}^n$, puede ser particionado en dos regiones disjuntas, $\mathscr{X} = \mathscr{X}_1 \cup \mathscr{X}_2$.

Definición 15. La afirmación "$\theta \in \Theta_0$" se denomina **hipótesis nula** (paramétrica) y la denotamos por H_0. Al haber asumido una partición del espacio paramétrico, disponemos de una **hipótesis alternativa** (i.e. "$\theta \in \Theta_1$") y la denotamos por H_1.

Enfrentamos las dos hipótesis:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

El *objetivo* es confirmar o refutar la hipótesis nula una vez observados los valores de $\mathbf{x}$.

Creemos de entrada que $\theta \in \Theta_0$ (en ausencia de información o nueva evidencia, es la hipótesis aceptada).

Definición 16. Un **contraste estadístico** (puro) para contrastar H_0 frente a H_1 es una partición del espacio muestral de forma que si $\mathbf{x} \in \mathscr{X}_1$ se rechaza H_0 en favor de H_1, y si $\mathbf{x} \in \mathscr{X}_0$ se acepta (mejor, 'no se rechaza') H_0. A la región $\mathscr{X}_1$ se la denomina región crítica del test y a la función indicadora de $\mathscr{X}_1$ se la denomina **función test**.

Solemos denotar la función test como $\phi : \mathscr{X} \to \{0, 1\}$,

$$\phi(\mathbf{x}) = \begin{cases} 0 & si \quad \mathbf{x} \notin \mathscr{X}_1 \\ 1 & si \quad \mathbf{x} \in \mathscr{X}_1 \end{cases}$$

Es muy frecuente denotar la región crítica o región de rechazo por la letra C. Su complementaria se suele denominar región de aceptación y se denota en ocasiones por la letra A.

Definición 17. Si $\Theta_0 = \{\theta_0\}$, (i.e. consta únicamente de un único punto), la hipótesis $H_0 : \theta \in \Theta_0$ se denomina simple. En caso contrario, se denomina hipótesis compuesta. La misma terminología se aplica a H_1.

En caso que enfrentemos hipótesis simple contra hipótesis simple, los contrastes son de la forma

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1$$

Si enfrentamos una hipótesis simple (o compuesta) contra una hipótesis compuesta, hay dos posibiliadades:

- Contrastes unilaterales: $H_0 : \theta = \theta_0$ (o $\theta \leq \theta_0$) vs. $H_1 : \theta > \theta_0$ o
  $H_0 : \theta = \theta_0$ (o $\theta \geq \theta_0$) vs. $H_1 : \theta < \theta_0$
- Contrastes bilaterales: $H_0 : \theta \in [\theta_1, \theta_2]$ vs. $H_1 : \theta \notin [\theta_1, \theta_2]$ y especialmente
  $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

Decisión y error asociado

De forma general, siempre que se toma la decisión de aceptar o rechazar la hipótesis nula, estamos sujetos a error. El papel asimétrico que juegan las hipótesis entre las que elegimos imponen distinguir también dos tipos de posibles errores.

Definición 18. Rechazar H_0 cuando es cierta, se denomina **error de tipo I**, o falso positivo, y aceptar H_0 cuando es falsa, se denomina **error de tipo II**, o falso negativo.

Esta situación se puede resumir en la siguiente tabla:

	No rechazar H_0	Rechazar H_0
H_0 cierta	✓	Error de tipo I
H_0 falsa	Error de tipo II	✓

Tabla A.1: Errores de tipo I y II

Si H_0 es simple (i.e. $\Theta_0 = \theta_0$), la probabilidad de error de tipo I es un valor, $\Pr_{\theta_0}(\mathbf{x} \in C) = \alpha$. Si H_0 es compuesta, la probabilidad de error de tipo I es función de $\theta \in \Theta_0$.

Análogamente, Si H_1 es simple (i.e. $\Theta_1 = \theta_1$), la probabilidad de error de tipo II es un valor, $\Pr_{\theta_1}(\mathbf{x} \notin C) = \beta$. Si H_1 es compuesta, la probabilidad de error de tipo II es función de $\theta \in \Theta_1$.

Un test de H_0 frente a H_1 se reduce a una partición del espacio muestral, $\mathscr{X} = \mathscr{X}_0 \cup \mathscr{X}_1$, con $\mathscr{X}_1 = C$ (región de rechazo).

Nivel y potencia de un test

Hay que disponer de algún criterio que mida o compare el rendimiento/potencia de un test.

Definición 19. La **función potencia** $\beta(\cdot)$ mide la probabilidad de rechazar H_0 como función de $\theta \in \Theta$:

$$\beta(\theta) = \Pr_\theta(\mathbf{x} \in C)$$

Definición 20. Se define el **tamaño de un test**,

$$\widetilde{\alpha} = \sup_{\theta \in \Theta_0} \beta(\theta)$$

y se dice que un test tiene un **nivel de significación** $\alpha \in (0,1)$ si

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$$

Se sigue que si un test tiene tamaño $\widetilde{\alpha}$ tendrá nivel de significación α, $\forall \alpha \geq \widetilde{\alpha}$.

Idealmente, querríamos que $\beta(\theta) = 0$ para $\theta \in \Theta_0$ y que $\beta(\theta) = 1$ para $\theta \in \Theta_1$ pero esto es generalmente imposible.

Hipótesis preferente

En un planteamiento puramente formal, se podrían plantear contrastes donde ambas hipótesis juegan un papel simétrico; sin embargo, en la práctica H_0 juega el papel de *"status quo"* (lo aceptado o aceptable en ausencia de evidencia).

Esta situación hace que se adopte como punto de partida la obtención de contrastes de hipótesis que tengan una pequeña probabilidad de rechazar H_0 cuando ésta es cierta, es decir, queremos una pequeña probabilidad de error de tipo I. Buscamos pues un valor α que acote la probabilidad de error de tipo I:

$$\sup_{\theta\in\Theta_0} \beta(\theta) \leq \alpha, \text{ es decir, } \sup_{\theta\in\Theta_0} \Pr_\theta(\mathbf{x}\in C) \leq \alpha$$

y tomaremos usualmente el valor de $\alpha = 0{,}05$, aunque es solo un convenio.

Contrastes MP/UMP

En ocasiones podemos disponer de clases de funciones test y queremos determinar el contraste óptimo o el mejor en algún sentido entre todas las funciones test disponibles. Como hemos dicho, debemos controlar el nivel de significación (cuanto menor sea, mejor) y la potencia del test (cuanto mayor sea, mejor).

Una vez fijadas H_0, H_1 y el nivel de significación α sigue habiendo varios test alternativos (basados en diferentes regiones de rechazo, por ejemplo), y deberemos elegir el que tenga más potencia en la alternativa, es decir, el que tenga más probabilidad de rechazar una hipótesis falsa, $1 - \Pr(\text{error tipo II})$.

Denotamos por $\mathscr{C}$ una familia de test de hipótesis para H_0 frente a H_1 con muestras de tamaño n. Será especialmente importante la clase $\mathscr{C}_\alpha$ de test de nivel α para H_0 frente a H_1.

Definición 21. El test $\phi_0 \in \mathscr{C}_\alpha$ (un test de nivel α), con función de potencia $\beta_0(\theta)$, se dice que es el **test más potente (MP)** para $\theta_1 \in \Theta_1$ fija, si

$$\beta_0(\theta_1) \geq \beta(\theta_1), \quad \forall \phi \in \mathscr{C}_\alpha$$

El test $\phi_0 \in \mathscr{C}_\alpha$ se dice que es el test **uniformemente más potente (UMP)** en $\mathscr{C}_\alpha$, si

$$\beta_0(\theta) \geq \beta(\theta), \quad \forall \phi \in \mathscr{C}_\alpha \quad \forall \theta \in \Theta_1$$

En resumen, entre todos los test que controlen el nivel de significación a nivel α, es decir, que la probabilidad de cometer un error de tipo I sea $\leq \alpha$, buscamos el que aumente todo lo posible la potencia, es decir, el que reduzca lo más posible la probabilidad de cometer un error de tipo II.

Obtener o diseñar contrastes óptimos para un nivel de significación α dado puede resultar muy atractivo, pero en general es un procedimiento muy complicado. Para algunos casos concretos, hay teoremas que nos dan criterios de construcción sencillos y fáciles de implementar. Por ejemplo, si contrastamos hipótesis simple frente a hipótesis simple, es decir, $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ el lema de Neyman-Pearson nos da el test MP de tamaño α [1, pág. 60]. El teorema de Karlin-Rubin [1, pág. 65] nos da el test UMP para contrastar $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$.

Estadísticos de contraste

Hasta ahora, nuestra decisión de rechazar H_0 se basaba en elegir una región de rechazo C y si $\mathbf{x} \in C$, rechazamos H_0. Sin embargo, es más práctico elegir la región de rechazo en función de un estadístico $T(\mathbf{x})$, denominado **estadístico de contraste**. Entonces, elgiendo una región de rechazo adecuada, rechazaremos H_0 si $T(\mathbf{x}) \in C$. Asumiremos de ahora en adelante que valores crecientes del estadístico de contraste apoyan la hipótesis alternativa, por lo que es habitual dar esta región de rechazo en función de un valor límite k, de modo que rechazaremos H_0 si se cumple $T(\mathbf{x}) \geq k$. Para valores decrecientes, basta con cambiar la desigualdad, y para valores extremos, basta tomar valor absoluto.

P-valores

Otro modo habitual de decidir si rechazamos o no H_0 es utilizar el concepto de p-valor. Sea C_α la región de rechazo para un contraste a nivel α. Si se cumple

$$C_\alpha \subset C_{\alpha'} \quad \forall \alpha < \alpha' \tag{A.1}$$

que es lo más habitual, entonces definimos:

Definición 22. El **p-valor** $P_n = P_n(\mathbf{x})$ es el nivel de significación más pequeño para el cual la hipótesis H_0 es rechazada para una observación $\mathbf{x}$. Es decir,

$$P_n(\mathbf{x}) = \text{ínf}\{\alpha \mid \mathbf{x} \in C_\alpha\} \tag{A.2}$$

Igual que hemos dicho que en lugar de rechazar una hipótesis si $\mathbf{x} \in C$, podemos utilizar un estadístico de contraste y rechazarla si $T(\mathbf{x}) \geq k$, por lo que el p-valor se puede redefinir como

$$P_n(\mathbf{x}) = \text{ínf}\{\alpha \mid T(\mathbf{x}) \geq k_\alpha\} \tag{A.3}$$

donde k_α quiere decir que el contraste dado por la regla de decisión $T(\mathbf{x}) \geq k_\alpha$ es de tamaño α.

En este caso, la regla de decisión será que si $P_n(\mathbf{x}) \leq \alpha$ entonces rechazaremos H_0.

Los p-valores tienen dos importantes propiedades:

Teorema A.1. *Sea $X \sim P_\theta$ con $\theta \in \Theta$. Suponer que estamos contrastando $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$. Asumamos que las regiones de rechazo satisfacen* (A.1)*.*

1. *Si*

$$\sup_{\theta \in \Theta_0} Pr_\theta(\boldsymbol{x} \in C_\alpha) \leq \alpha \quad \forall \alpha \in (0,1)$$

entonces la distribución de los p-valores si $\theta \in \Theta_0$ cumple

$$Pr_\theta(P_n \leq u) \leq u \quad \forall u \in [0,1]$$

2. *Si, para $\theta \in \Theta_0$*

$$Pr_\theta(\boldsymbol{x} \in C_\alpha) = \alpha \quad \forall \alpha \in (0,1)$$

entonces para $\theta \in \Theta_0$ los p-valores cumplen

$$Pr_\theta(P_n \leq u) = u \quad \forall u \in [0,1]$$

es decir, los p-valores están uniformemente distribuidos en $(0,1)$

Demostración.

1. Si $\theta \in \Theta_0$ el evento $\{P_n \leq u\}$ implica $\{\mathbf{x} \in C_v\}$ para todo $u < v$. Por tanto, $\text{Pr}_\theta(P_n \leq u) \leq \text{Pr}_\theta(\mathbf{x} \in C_v)$ para todo $u < v$. El resultado se sigue de hacer $v \to u$ y de la hipótesis del enunciado.

2. Como el evento $\{\mathbf{x} \in C_u\}$ implica $\{P_n \leq u\}$, se sigue que $\text{Pr}_\theta(P_n \leq u) \geq \text{Pr}_\theta(\mathbf{x} \in C_u)$. Por la hipótesis del enunciado, $\text{Pr}_\theta(\mathbf{x} \in C_u) = u$, luego $\text{Pr}_\theta(P_n \leq u) \geq u$. Combinando esta desigualdad con la otra desigualdad que ya hemos probado, se tiene el resultado.

□

Tenemos por lo tanto dos reglas de decisión para rechazar una hipótesis H_0 y que controlan el nivel de signifiación a nivel α. La primera, basada en los valores límite para el estadístico de contraste T, en la que rechazamos la hipótesis si

$$T(\mathbf{x}) \geq k_\alpha$$

y la segunda basada en los p-valores, en la que rechazamos la hipótesis si

$$P_n(\mathbf{x}) \leq \alpha$$

La ventaja de utilizar los p-valores como regla de decisión es que el nivel de significación no tiene que ser fijado previamente, si no que podemos calcular $P_n(\mathbf{x})$ y luego elegir si rechazamos o no cada hipótesis según el nivel de significación que estemos dispuestos a tolerar. Además, los p-valores nos sirven para ver cuánta evidencia tenemos para rechazar una hipótesis. Cuanto más pequeño sea el p-valor, más evidencia.

A.0.1. Desigualdad de Simes

El método de Hommel descrito en la Sección 2.1.3 está basado en una generalización de la desigualdad de Simes, probada por Simes en [22].

Sean $P_n(\mathscr{O}_n(1)),\ldots,P_n(\mathscr{O}_n(m))$ los m p-valores sin ajustar ordenados para contrastar las m hipótesis $H_{0\mathscr{O}_n(1)},\ldots,H_{0\mathscr{O}_n(m)}$. Simes consideró el siguiente procedimiento: rechazar las m hipótesis si $P_n(\mathscr{O}_n(j)) \leq \frac{j\alpha}{m}$ para algún $j = 1,\ldots,m$, y en caso contrario no rechazar ninguna.

La desigualdad probada por Simes es la siguiente:

Teorema A.2. *Sean $P_n(\mathscr{O}_n(1)),\ldots,P_n(\mathscr{O}_n(m))$ los estadísticos ordenados de m variables aleatorias independientes y uniformes en $(0,1)$ y sea $A_m(\alpha) = Pr\left\{P_n(\mathscr{O}_n(j)) > \frac{j\alpha}{m} \ \forall j = 1,\ldots,m\right\}$ $(0 \leq \alpha \leq 1)$. Entonces, $A_m(\alpha) = 1-\alpha$.*

Demostración. El resultado es claramente cierto para $m = 1$. Para $m > 1$, $\left\{\frac{P_n(\mathscr{O}_n(1))}{P_n(\mathscr{O}_n(m))},\ldots,\frac{P_n(\mathscr{O}_n(m-1))}{P_n(\mathscr{O}_n(m))}\right\}$ son los estadísticos ordenados de $m-1$ variables independientes uniformes en $(0,1)$, independientes de $P_n(\mathscr{O}_n(m))$, y $P_n(\mathscr{O}_n(m))$ tiene una distribución de p^m $(0 < p < 1)$. Por tanto,

$$A_m(\alpha) = \int_\alpha^1 A_{m-1}\left(\frac{\alpha(m-1)}{pm}\right) m p^{m-1} dp$$

Si $A_{m-1}(\alpha) = 1-\alpha$ entonces se sigue $A_m(\alpha) = 1-\alpha$ y el resultado queda probado por inducción. □

Si todas las hipótesis nulas son ciertas, se cumple que todos sus p-valores están uniformemente distribuidos en $(0,1)$, como hemos visto en el Teorema A.1. Por tanto, si las hipótesis son independientes, podemos aplicar la desigualdad de Simes, y obtenemos que la probabilidad de que $P_n(\mathscr{O}_n(j)) \leq \frac{j\alpha}{m}$ para algún $j = 1,\ldots,m$ vale α, por lo que la probabilidad de rechazar todas las hipótesis es α.

Es decir, en el caso de que todas las hipótesis nulas sean ciertas, si utilizamos el procedimiento descrito por Simes, tenemos una probabilidad de rechazar todas de α, y una probabilidad de no rechazar ninguna de $1-\alpha$. Por tanto, como todas las hipótesis son ciertas, el FWER, que es la probabilidad de rechazar al menos una hipótesis cierta, como vemos en la Definición 3, vale α. Recordar que nuestro objetivo es controlar el FWER a nivel α, que es lo que hemos conseguido en este caso.

Anexo B

Código de simulación

En este anexo explicamos el código de R que hemos utilizado para hacer la simulación del Capítulo 3.

Para obtener los p-valores sin ajustar, lo primero que tenemos que hacer es importar los datos que ha simulado Myriads. Una vez importados, hacemos un *t-test* para cada gen que estamos estudiando, y seleccionamos el p-valor que nos devuelve.

```
DataMatrix<-read.delim("C:/Users/jprub/Myriads_Data/DataMatrix_n1_20_n2_20_1.txt")

pvalues<-rep(0,dim(DataMatrix)[1])

for(i in 1:dim(DataMatrix)[1]){
  pvalues[i]<-t.test(DataMatrix[i,2:21],
                     DataMatrix[i,22:41],
                     var.equal = TRUE)$p.value
}
```

Este es el método que hay que seguir para calcular los p-valores sin ajustar. Sin embargo, Myriads nos ofrece la posibilidad de, en lugar de dar una tabla con los datos, dar un vector con los p-valores correspondientes a cada gen. Esto es preferible porque cuando posteriormente hagamos muchas simulaciones nos ahorraremos un tiempo de computación considerable. De este modo, podemos ahorrarnos el trozo de código anterior, e importar directamente el vector de p-valores.

```
MyriadsPval <-read.csv("C:/Users/jprub/Myriads_Sims_t/MyriadsPval_1.txt",sep="")

pval_sin_ajustar <- MyriadsPval[[1]] #para pasar de lista a vector
```

Aplicamos ahora los métodos de ajuste del p-valor descritos, y guardamos los p-valores de cada método en las filas de una matriz.

```
matriz_pval <- matrix(NA, nrow=8, ncol=1000)

matriz_pval[1,] <- pval_sin_ajustar
matriz_pval[2,] <- p.adjust(pval_sin_ajustar,
                   method = "bonferroni", n = length(pval_sin_ajustar))
matriz_pval[3,] <- p.adjust(pval_sin_ajustar,
                   method = "holm", n = length(pval_sin_ajustar))
matriz_pval[4,] <- p.adjust(pval_sin_ajustar,
                   method = "hommel", n = length(pval_sin_ajustar))
```

```
matriz_pval[5,] <- p.adjust(pval_sin_ajustar,
                   method = "BH", n = length(pval_sin_ajustar))
matriz_pval[6,] <- p.adjust(pval_sin_ajustar,
                   method = "BY", n = length(pval_sin_ajustar))
matriz_pval[7,] <- adjust.p(pval_sin_ajustar,
                   pi0.method = "bky", alpha = 0.05)$adjp[,2]
matriz_pval[8,] <- qvalue(pval_sin_ajustar)$qvalues
```

Para calcular la tabla en la que vemos las hipótesis falsas que han sido rechazadas y las que no, y las hipótesis verdaderas que han sido rechazadas y las que no, contamos los p-valores que son $\leq 0{,}05$ y $> 0{,}05$ entre las 500 primeras hipótesis y entre las 500 últimas para cada método. Recordar que en nuestro ejemplo $\pi_0 = 0{,}5$, y que Myriads coloca las hipótesis falsas primero y luego las verdaderas.

```
tabla_rechazos <- matrix(NA, nrow=8, ncol=4)

for(i in 1:8){
    tabla_rechazos[i,1] <- sum(matriz_pval[i,1:500]<=0.05)
    tabla_rechazos[i,2] <- sum(matriz_pval[i,1:500]>0.05)
    tabla_rechazos[i,3] <- sum(matriz_pval[i,501:1000]<=0.05)
    tabla_rechazos[i,4] <- sum(matriz_pval[i,501:1000]>0.05)
}
```

Ahora queremos estimar el FWER, el FDR y la potencia. Lo que hacemos es, para cada método, guardar nuestras estimaciones de estas tres cantidades en una matriz de 8 filas, una para cada método, y 3 columnas, una para cada cantidad que queremos estimar.

Para estimar el FWER, comprobamos si hay una o más hipótesis nulas verdaderas rechazadas. Si las hay, el valor que toma es 1, y si no, 0.

Para estimar el FDR, calculamos $\frac{V_n}{R_n}$, el cociente entre las hipótesis nulas verdaderas rechazadas y el total de rechazos, distinguiendo el caso en el que $R_n = 0$, en cuyo caso le asignamos el valor 0.

Para estimar la potencia, calculamos $\frac{m_1 - U_n}{m_1}$, es decir, el cocientre entre las hipótesis nulas falsas rechazadas y el total de hipótesis nulas falsas.

```
matriz_estimaciones <- matrix(NA, nrow=8, ncol=3)

for (i in 1:8){
if (sum(matriz_pval[i,501:1000]<=0.05) >= 1){
  matriz_estimaciones[i,1] <- 1
} else {
  matriz_estimaciones[i,1] <- 0
}

if (sum(matriz_pval[i,]<=0.05)==0){
  matriz_estimaciones[i,2] <- 0
} else {
matriz_estimaciones[i,2] <- sum(matriz_pval[i,501:1000]<=0.05) /
                            sum(matriz_pval[i,]<=0.05)
}

matriz_estimaciones[i,3] <- sum(matriz_pval[i,1:500]<=0.05) / 500
}
```

Como queremos estimar el FWER, el FDR y la potencia de un modo más preciso, lo que hacemos es repetir el proceso descrito hasta ahora 100 veces, obteniendo de este modo 100 matrices con el cálculo de las estimaciones para cada simulación, y calculamos la media entre los valores obtenidos.

Para hacer esto, le pedimos a Myriads que nos simule 100 vectores de 1000 p-valores cada uno. Nos los devuelve con el formato *MyriadsPval_i.txt*, siendo $i = 1, \ldots, 100$ el número de la simulación. Repetimos entonces el proceso anterior 100 veces y vamos guardando las matrices resultantes en una lista.

```
lista_matrices <- list()

archivos <- paste0("C:/Users/jprub/Myriads_Sims_t/MyriadsPval_", 1:100, ".txt")

for(archivo in archivos){
########## Código ya explicado
MyriadsPval <- read.csv(archivo, sep="")

pval_sin_ajustar <- MyriadsPval[[1]] #para pasar de lista a vector

matriz_pval <- matrix(NA, nrow=8, ncol=1000)

matriz_pval[1,] <- pval_sin_ajustar
matriz_pval[2,] <- p.adjust(pval_sin_ajustar,
                    method = "bonferroni", n = length(pval_sin_ajustar))
matriz_pval[3,] <- p.adjust(pval_sin_ajustar,
                    method = "holm", n = length(pval_sin_ajustar))
matriz_pval[4,] <- p.adjust(pval_sin_ajustar,
                    method = "hommel", n = length(pval_sin_ajustar))
matriz_pval[5,] <- p.adjust(pval_sin_ajustar,
                    method = "BH", n = length(pval_sin_ajustar))
matriz_pval[6,] <- p.adjust(pval_sin_ajustar,
                    method = "BY", n = length(pval_sin_ajustar))
matriz_pval[7,] <- adjust.p(pval_sin_ajustar,
                    pi0.method = "bky", alpha = 0.05)$adjp[,2]
matriz_pval[8,] <- qvalue(pval_sin_ajustar)$qvalues

matriz_estimaciones <- matrix(NA, nrow=8, ncol=3)

for (i in 1:8){
  if (sum(matriz_pval[i,501:1000]<=0.05) >= 1){
    matriz_estimaciones[i,1] <- 1
  } else {
    matriz_estimaciones[i,1] <- 0
  }

  if (sum(matriz_pval[i,]<=0.05)==0){
    matriz_estimaciones[i,2] <- 0
  } else {
    matriz_estimaciones[i,2] <- sum(matriz_pval[i,501:1000]<=0.05) /
                                sum(matriz_pval[i,]<=0.05)
  }
```

```
  matriz_estimaciones[i,3] <- sum(matriz_pval[i,1:500]<=0.05) / 500
}
##########
lista_matrices[[length(lista_matrices) + 1]] <- matriz_estimaciones

}
```

Ahora, con las matrices de la lista, hacemos la media de los valores obtenidos para cada celda.

```
matriz_medias_estimaciones<- matrix(NA, nrow=8, ncol=3)

v <- c()
for(i in 1:8){
  for(j in 1:3){
    for(k in 1:100){
      v[k] <- lista_matrices[[k]][i,j]
    }
    matriz_medias_estimaciones[i,j] <- mean(v)
  }
}
```

Y de este modo hemos obtenido una matriz con las estimaciones del FWER, del FDR y de la potencia para cada método estudiado. Esto es solo para el caso $\pi_0 = 0{,}5$ y en el que los p-valores son independientes. Podemos ir cambiando el grado de dependencia en Myriads, y de este modo obtener los estimadores correspondientes para cada grado de dependencia. Hacemos tres matrices de 8 filas, una para cada método, y 10 columnas, una para cada grado de dependencia a estudiar. En nuestro caso, los valores del grado de dependencia que simularemos serán 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 y 0.9.

```
matriz_est_fwer <- matrix(NA, nrow = 8, ncol = 10)
matriz_est_fdr <- matrix(NA, nrow = 8, ncol = 10)
matriz_est_power <- matrix(NA, nrow = 8, ncol = 10)

matriz_est_fwer[,1] <- matriz_medias_estimaciones[,1]
matriz_est_fdr[,1] <- matriz_medias_estimaciones[,2]
matriz_est_power[,1] <- matriz_medias_estimaciones[,3]
```

Notar que, con la simulación que hemos dado de ejemplo en la que los p-valores son independientes, hemos rellenado una columna de cada matriz. Cambiando el grado de dependencia en Myriads, podemos volver a ejecutar el código y rellenar el resto de columnas de estas tres matrices. Al acabar de hacer este proceso, tenemos las tres matrices que hemos representado como mapas de calor.

Con R podríamos haber utilizado la función `heatmap()` para hacer los mapas de calor, pero los resultado obtenidos no eran satisfactorios. Por ello, utilizamos Python para dibujarlos. Para ello, cambiamos los nombres de las filas y las columnas de las matrices en R y las exportamos en archivos *.csv*.

```
colnames(matriz_est_fwer) <- seq(0, 0.9, by=0.1)
colnames(matriz_est_fdr) <- seq(0, 0.9, by=0.1)
colnames(matriz_est_power) <- seq(0, 0.9, by=0.1)
rownames(matriz_est_fwer) <- c("P-valores sin ajustar", "Bonferroni", "Holm",
                               "Hommel", "BH", "BY", "BKY", "Q-valores")
rownames(matriz_est_fdr) <- c("P-valores sin ajustar", "Bonferroni", "Holm",
                              "Hommel", "BH", "BY", "BKY", "Q-valores")
```

```
rownames(matriz_est_power) <- c("P-valores sin ajustar", "Bonferroni", "Holm",
                                "Hommel", "BH", "BY", "BKY", "Q-valores")

write.csv(matriz_est_fwer,"matriz_est_fwer.csv")
write.csv(matriz_est_fdr,"matriz_est_fdr.csv")
write.csv(matriz_est_power,"matriz_est_power.csv")
```

Ahora importamos los archivos *.csv* desde Python y dibujamos los mapas de calor.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df_fwer = pd.read_csv('matriz_est_fwer.csv')
df_fwer=df_fwer.rename(columns = {'Unnamed: 0':'Métodos'})
df_fwer.set_index('Métodos', inplace=True)
df_fwer.columns.name='Grado de dependencia'

f, ax = plt.subplots(figsize=(10, 8))
s=sns.heatmap(df_fwer, annot=True, linewidths=.5,
              cmap=sns.diverging_palette(220, 20, s=80, l=53, as_cmap=True),
              center=0.05, ax=ax);
plt.title('Estimación FWER', size=20)
plt.xlabel('Grado de dependencia', fontsize=15)
plt.ylabel('Método', fontsize=15)
```

Lo mismo habría que hacer para obtener los mapas de calor correspondientes al FDR y a la potencia.

Bibliografía

[1] E. L. LEHMAN y J. P. ROMANO, *Testing Statistical Hypothesis*, Springer, Nueva York, 2005

[2] S. DUDOIT, M. J. VAN DER LAAN y K. S. POLLARD, *Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates*, U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 138, 2003, 1-15

[3] A. CARVAJAL-RODRÍGUEZ, *Bioinformatics of Correction of Multiple Testing: An Introduction for Life Scientists*, CRC Press, Chapter 17, 2021

[4] `https://en.wikipedia.org/wiki/Bonferroni_correction`

[5] J. J. GOEMANA y A. SOLARI, *Multiple hypothesis testing in genomics*, Statistics in Medicine, **33** (2014) 1946–1978.

[6] S. HOLM, *A Simple Sequentially Rejective Multiple Test Procedure*, Scandinavian Journal of Statistics, **6**, (2) (1979), 65-70. `https://www.jstor.org/stable/4615733`

[7] G. HOMMEL, *A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test*, Biometrika, **75**, (2) (1988), 383-386. `http://www.jstor.org/stable/2336190`

[8] S. P. WRIGHT, *Adjusted P-Values for Simultaneous Inference*, Biometrics, **48**, (1992) 1005–1013.

[9] J. GOEMAN, R. MEIJER y T. KREBS, *hommel: Methods for Closed Testing with Simes Inequality, in Particular Hommel's Method*, R package, version 1.

[10] Y. BENJAMINI y Y. HOCHBERG, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society. Series B (Methodological), **57**, (1) (1995), 289-300. `https://www.jstor.org/stable/2346101`

[11] Y. BENJAMINI y D. YEKUTIELI, *The Control of the False Discovery Rate in Multiple Testing under Dependency*, The Annals of Statistics, **29**, (4) (2001), 1165–1188.

[12] `https://en.wikipedia.org/wiki/Euler%E2%80%93Maclaurin_formula`

[13] Y. BENJAMINI, A. M. KRIEGER y D. YEKUTIELI, *Adaptive linear step-up procedures that control the false discovery rate*, Biometrika, **93**, (3) (2006), 491–507

[14] G. BLANCHARD y É. ROQUAIN, *Adaptive False Discovery Rate Control under Independence and Dependence*, Journal of Machine Learning Research, **10**, (2009), 2837-2871

[15] J. D. STOREY y R. TIBSHIRANI, *Statistical significance for genomewide studies*, PNAS, **100**, (16) (2003), 9440–9445

[16] A. B. OWEN, *Variance of the Number of False Discoveries*, Journal of the Royal Statistical Society. Series B (Statistical Methodology), **67**, (3) (2005), 411-426. `https://www.jstor.org/stable/3647668`

[17] J. D. STOREY, J. E. TAYLOR Y D. SIEGMUND, *Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach*, Journal of the Royal Statistical Society. Series B (Statistical Methodology), **66**, (1) (2004), 187-205. https://www.jstor.org/stable/3647634

[18] J. D. STOREY, *The Positive False Discovery Rate: a Bayesian Interpretation and the Q-value*, The Annals of Statistics, **31**, (6) (2003), 2013–2035

[19] A. CARVAJAL-RODRÍGUEZ, *Myriads*, http://myriads.webs.uvigo.es/

[20] F. J. TEJEDOR TEJEDOR, *Análisis de varianza*, La Muralla S.A., Madrid, 1999

[21] J. T. ALCALÁ NALVÁIZ, *Apuntes del curso Estadística Matemática. Test de hipótesis*, Curso 2021-2022

[22] R. J. SIMES, *An improved Bonferroni procedure for multiple tests of significance*, Biometrika, **73**, (3) (1986), 751-754

[23] CP4P: CALIBRATION PLOT FOR PROTEOMICS, *Q. G. Gianetto, F. Combes, C. Ramus, C. Bruley, Y. Couté y T. Burger*, R package version 0.3.6, (2019), https://CRAN.R-project.org/package=cp4p

[24] QVALUE: Q-VALUE ESTIMATION FOR FALSE DISCOVERY RATE CONTROL, *J. D. Storey, A. J. Bass, A. Dabney, D. Robinson*, R package version 2.32.0, (2023), https://bioconductor.org/packages/qvalue