Comparaciones múltiples: corrección de p-value y FDR

Joaquín Amat Rodrigo *j.amatrodrigo@gmail.com*

Diciembre, 2016

Índice

Introducción

En el contexto que trata este capítulo, el termino *comparación* hace referencia a la comparación de dos grupos mediante un test de contraste de hipótesis (t-test, U-test,...). El *p-value* calculado por del test refleja la probabilidad de obtener una diferencia igual o más extrema que la observada, siendo cierta la hipótesis nula. Considerar la diferencia observada como significativa depende de sí el *p-value* obtenido está por debajo de un límite establecido por el investigador al que se conoce como nivel de significancia α.

Las comparaciones múltiples surgen cuando un estudio estadístico conlleva la realización de varias comparaciones con el objetivo de identificar aquellos grupos para los que las diferencias son más significativas. La siguiente tabla recoge los posibles resultados obtenidos al contrastar *m* hipótesis nulas.

	Test considerado significativo	Test considerado no significativo	Total
Hipótesis nula verdadera (Ho)	F	$m_0 - F$	m_0
Hipótesis alternativa verdadera (HA)	T	$m_1 - T$	m_1
Total	S	$m - S$	m

- m: número total de hipótesis contrastadas.
- m_0: número de hipótesis nulas verdaderas, en la práctica este parámetro se desconoce.
- m_1: número de hipótesis alternativas verdaderas (hipótesis nulas falsas), en la práctica este parámetro se desconoce.
- F: número de falsos positivos (error tipo I, *false discoveries*).
- T: número de verdaderos positivos, hipótesis alternativas correctamente detectadas (*true discoveries*).
- $m_0 - F$: número de verdaderos negativos.
- $m_1 - T$: número de falsos negativos (error tipo II).
- S: número de hipótesis nulas rechazadas 8test significativo), independientemente de que sean ciertas o falsas.
- $m - S$: número de hipótesis alternativas rechazadas, independientemente de que sean ciertas o falsas.

La estimación del *false positive rate*, probabilidad de rechazar la hipótesis nula (test significativo) siendo esta cierta, se obtiene como:

$$\frac{F}{m_0}$$

La estimación del *false discovery rate*, probabilidad de que la hipótesis nula sea cierta a pesar de haber sido rechazada (test significativo), se obtiene como:

$$\frac{F}{S}$$

Family-Wise Error Rate (FWER)

Si se considera un nivel de significancia $\alpha = 0.05$, existe un 5% de probabilidad de rechazar la hipótesis nula siendo esta verdadera, a esto se le conoce como error tipo I, *false positive rate* o probabilidad de falso positivo. Si se realiza un único test, esta probabilidad de falso positivo es relativamente baja. Sin embargo, si se realizan 100 test, se esperan en promedio 5 falsos positivos a pesar de que todas hipótesis nulas fuesen ciertas. Si los test son independientes, la probabilidad de que al menos 1 de los 100 test rechace la hipótesis nula de forma incorrecta es del 99.4%.

```
#probabilidad de no obtener ningún evento verdadero en 100 intentos siendo p=0.05
1 - dbinom(x = 0, size = 100, prob = 0.05)
```

```
## [1] 0.9940795
```

```
#o equivalentemente
1 - (1 - 0.05) ^ 100
```

```
## [1] 0.9940795
```

Por lo tanto, el problema de las comparaciones múltiples se debe al incremento en el error de tipo I como consecuencia del uso repetido de test estadísticos. Si se realizan k comparaciones independientes, la probabilidad de que al menos ocurra un falso positivo, también conocida como *family-wise error rate (FWER)* viene dada por la ecuación:

$$FWER = \overline{\alpha} = 1 - (1 - \alpha_{per\ comparison})^k$$

Ejemplo 1 family-wise error rate (FWER)

Un experimento pretende determinar si una moneda está trucada realizando 10 lanzamientos. Si se cumple la hipótesis nula de que la moneda no está trucada, la probabilidad de obtener 9 veces cara en 10 lanzamientos es de 0.0107. Si tal escenario ocurriera y se emplease un límite de significancia de $\alpha = 0.05$, se rechazaría la hipótesis nula y se consideraría la moneda como trucada. El problema surge si se quiere emplear este mismo test (que es apropiado para testar una moneda) para evaluar muchas monedas, por ejemplo 100. A pesar de que ninguna de las 100 monedas esté trucada, siendo 0.0107 la probabilidad de obtener 9 caras en 10 lanzamientos de una moneda normal, la probabilidad de que el método empleado considere como buenas todas ellas es de $(1 - 0.0107)^{100} \approx 0.34$. Esto equivale a decir que el *family-wise error rate (FWER)* o probabilidad de al menos un falso positivo es de $1 - (1 - 0.0107)^{100} \approx 0.66$.

```
#simulación de 10 lanzamientos de una moneda normal (50%) y evaluación de la moneda
set.seed(5351)
lanzamientos <- function() {
  if (sum(sample(x = c(TRUE, FALSE), size = 10, replace = TRUE)) >= 9) {
    return("trucada")
  }else{
    return("normal")
  }
}

table(replicate(n = 100, expr = lanzamientos()))
```

```
##
##  normal trucada
##      98       2
```

Ejemplo 2 family-wise error rate (FWER)

Supóngase una máquina empleada para la fabricación de varillas de metal. Se sabe que el diámetro de las varillas se distribuye de forma normal con media 3mm y desviación estándar de 1mm. Se produce un lote de 1000 varillas del que se extraen dos muestras de 15 varillas cada una. Se realiza un *t-test* para contrastar la hipótesis nula de que ambas muestras proceden de la misma población, empleando un nivel de significancia $\alpha = 0.05$.

```r
set.seed(863)
poblacion <- rnorm(n = 1000, mean = 3, sd = 1)
muestra_a <- sample(poblacion, size = 15)
muestra_b <- sample(poblacion, size = 15)
t.test(muestra_a, muestra_b, alternative = "two.sided", var.equal = TRUE)$p.value
```

```
## [1] 0.8531612
```

Como es de esperar, dado que las varillas proceden de la misma población, el *p-value* obtenido no muestra evidencias en contra de la hipótesis nula.

Supóngase que se repite el proceso 100 veces, extrayendo cada vez dos nuevas muestras de esa misma población.

```r
comparacion_muestras <- function() {
  muestra_a <- sample(poblacion, size = 15)
  muestra_b <- sample(poblacion, size = 15)
  t.test(muestra_a, muestra_b, alternative = "two.sided", var.equal = TRUE)$p.value
}

set.seed(9651)
p_values <- replicate(n = 100, expr = comparacion_muestras())
sum(p_values < 0.05)
```

```
## [1] 6
```

A pesar de que todas las muestras se han obtenido siempre de la misma población y por lo tanto todas las hipótesis nulas contrastadas eran ciertas, al realizar 100 comparaciones, 6 test han resultado significativos.

Comparaciones múltiples tras un ANOVA

Si un Análisis de Varianza resulta significativo, implica que al menos dos de las medias comparadas son significativamente distintas entre sí, pero no se indica cuáles. Para identificarlas hay que comparar dos a dos las medias de todos los grupos introducidos en el análisis mediante un *t-test* u otro test que compare 2 grupos, ha esto se le conoce como análisis *post-hoc*. Debido a la inflación del error de tipo I, los niveles de significancia pueden ser ajustados en función del número de comparaciones (corrección de significancia). Si no se hace ningún tipo de corrección se aumenta la probabilidad de falsos positivos (error tipo I) pero, si se es muy estricto con las correcciones, se pueden considerar como no significativas diferencias

que realmente sí lo son (error tipo II). La necesidad de corrección o no, y de qué tipo, se ha de estudiar con detenimiento en cada caso.

Intervalos *LSD* de Fisher (Least Significance Method)

Siendo $\overline{x}_i$ la media muestral de un grupo, la desviación típica estimada de dicha media (asumiendo la homocedasticidad de los grupos) es igual a la raíz cuadrada de los Cuadrados Medios del Error (que como se ha visto es la estimación de la intravarianza o varianza del error) dividida por el número de observaciones de dicho grupo. Asumiendo también la normalidad de los grupos, se puede obtener el intervalo LSD como:

$$\overline{x}_i \pm \frac{\sqrt{2}}{2} t^{\alpha}_{gl(error)} \sqrt{\frac{SSE}{n}}$$

Los intervalos *LSD* son básicamente un conjunto de *t-test* individuales con la única diferencia de que en lugar de calcular una *pooled SD* empleando solo los dos grupos comparados, calcula la *pooled SD* a partir de todos los grupos.

Cuanto más se alejen los intervalos de dos grupos más diferentes son sus medias, siendo significativa dicha diferencia si los intervalos no se solapan. Es importante comprender que los intervalos *LSD* se emplean para comparar las medias pero no se pueden interpretar como el intervalo de confianza para cada una de las medias. El método *LSD* no conlleva ningún tipo de corrección de significancia, es por esto que su uso parece estar desaconsejado para determinar significancia aunque sí para identificar que grupos tienen las medias más distantes.

En R se pueden obtener los intervalos *LSD* y su representación gráfica mediante la función `LSD.test {agricolae}`.

Bonferroni adjustment

Este es posiblemente el ajuste de significancia más extendido a pesar de que no está recomendado para la mayoría de las situaciones que se dan en el ámbito de la biomedicina. La corrección de *Bonferroni* consiste en dividir el nivel de significancia α entre el número de comparaciones dos a dos realizadas.

$$\alpha_{corregido} = \frac{\alpha}{número\ de\ grupos}$$

Con esta corrección se asegura que la probabilidad de obtener al menos un falso positivo entre todas las comparaciones (*family-wise error rate*) es $\leq \alpha$. Permite por lo tanto contrastar una hipótesis nula general (la de que toda las hipótesis nulas testadas son verdaderas) de

forma simultanea, cosa que raramente es de interés en las investigaciones. Se considera un método excesivamente conservativo sobre todo a medida que se incrementa el número de comparaciones. Se desaconseja su utilización excepto en situaciones muy concretas. *False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies, Mark E. Glickman*.

En R se puede calcular mediante la función `pairwise.t.test()` indicando en los argumentos `p.adj = "bonferroni"`.

Holm–Bonferroni Adjustment

Con este método, el valor de significancia α se corrige secuencialmente haciéndolo menos conservativo que el de *Bonferroni*. Aun así, parece que tampoco es indicado si se realizan más de 6 comparaciones.

El proceso consiste en realizar un *t-test* para todas las comparaciones y ordenarlas de menor a mayor *p-value*. El nivel de significancia para la primera comparación (la que tiene menor *p-value*) se corrige dividiendo α entre el número total de comparaciones, si no resulta significativo se detiene el proceso, si sí que lo es, se corrige el nivel de significancia de la siguiente comparación (segundo menor *p-value*) dividiendo entre el número de comparaciones menos uno. El proceso se repite hasta detenerse cuando la comparación ya no sea significativa.

Tukey-Kramer Honest Significant Difference (HSD)

Es el ajuste recomendado cuando el número de grupos a comparar es mayor de 6 y el diseño es equilibrado (mismo número de observaciones por grupo). En el caso de modelos no equilibrados el método *HSD* es conservativo, requiere diferencias grandes para que resulte significativo. Solo aplicable si se trata de datos no pareados.

El *Tukey's test* es muy similar a un *t-test*, excepto que corrige el *experiment wise error rate*. Esto lo consigue empleando un estadístico que sigue una distribución llamada *studentized range distribution* en lugar de una *t-distribution*. El estadístico empleado se define como:

$$q_{calculado} = \frac{\overline{x}_{max} - \overline{x}_{min}}{S\sqrt{2/n}}$$

Donde: $\overline{x}_{max}$ es la mayor de las medias de los dos grupos comparados, $\overline{x}_{min}$ la menor, S la *pooled SD* de estos dos grupos y n el número total de observaciones en los dos grupos.

Para cada par de grupos, se obtiene el valor $q_{calculado}$ y se compara con el esperado acorde a una *studentized range distribution* con los correspondientes grados de libertad. Si la probabilidad es menor al nivel de significancia α establecido, se considera significativa la diferencia de medias. Al igual que con los intervalos *LSD*, es posible calcular intervalos *HSD* para estudiar su solapamiento.

En R, las funciones `TukeyHSD()` y `plot(TukeyHSD)` permiten calcular los *p-value* corregidos por Tukey y representar los intervalos.

Dunnett's correction (Dunnett's test)

Es el equivalente al test Tukey-Kramer (HSD) recomendado cuando en lugar de comparar todos los grupos entre sí ($\frac{(k-1)k}{2}$ comparaciones) solo se quieren comparar frente a un grupo control ($k-1$ comparaciones). Se emplea con frecuencia en experimentos médicos.

Inconvenientes de controlar el false positive rate mediante Bonferroni

Tal como se ha visto en los ejemplos, al realizar múltiples comparaciones es importante controlar la inflación del error de tipo I. Sin embargo, correcciones como las de *Bonferroni* o similares conllevan una serie de problemas. La primera es que el método se desarrolló para contrastar la hipótesis nula universal de que los dos grupos son iguales para todas las variables testadas, no para aplicarlo de forma individual a cada test. A modo de ejemplo, supóngase que un investigador quiere determinar si un nuevo método de enseñanza es efectivo empleando para ello estudiantes de 20 colegios. En cada colegio se selecciona de forma aleatoria un grupo control, un grupo que se somete al nuevo método y se realiza un test estadístico entre ambos considerando un nivel de significancia $\alpha = 0.05$. La corrección de *Bonferroni* implica comparar el *p-value* obtenido en los 20 test frente a $0.05/20 = 0.0025$. Si alguno de los *p-values* es significativo, la conclusión de *Bonferroni* es que la hipótesis nula de que el nuevo sistema de enseñanza no es efectivo en todos los grupos (colegios) queda rechazada, por lo que se puede afirmar que el método de enseñanza es efectivo para alguno de los 20 grupos, pero no cuáles ni cuántos. Este tipo de información no es de interés en la gran mayoría de estudios, ya que lo que se desea conocer es qué grupos difieren.

El segundo problema de la corrección de *Bonferroni* es que una misma comparación será interpretada de forma distinta dependiendo del número de test que se hagan. Supóngase

que un investigador realiza 20 contrastes de hipótesis y que todos ellos resultan en un *p-value* de 0.001. Aplicando la corrección de *Bonferroni* si el límite de significancia para un test individual es de $\alpha = 0.05$, el nivel de significancia corregido resulta ser *0.05/20 = 0.0025*, por lo que el investigador concluye que todos los test son significativos. Un segundo investigador reproduce los mismos análisis en otro laboratorio y llega a los mismos resultados, pero para confirmarlos todavía más, realiza 80 test estadísticos adicionales con lo que su nivel de significancia corregido pasa a ser de 0.05/100 = 0.0005. Ahora, ninguno de los test se puede considerar significativo, por lo que debido a aumentar el número de contrastes las conclusiones son totalmente contrarias.

Viendo los problemas que implica ¿Para qué sirve entonces la corrección de Bonferroni?

Que su aplicación en las disciplinas biomédicas no sea adecuada no quita que pueda serlo en otras áreas. Imagínese una factoría que genera bombillas en lotes de 1000 unidades y que testar cada una de ellas antes de repartirlas no es práctico. Una alternativa consiste en comprobar únicamente una muestra de cada lote, rechazando cualquier lote que tenga más de *x* bombillas defectuosas en la muestra. Por supuesto, la decisión puede ser errónea para un determinado lote, pero según la teoría de Neyman-Pearson, se puede encontrar el valor *x* para el que se minimiza el ratio de error. Ahora bien, la probabilidad de encontrar *x* bombillas defectuosas en la muestra depende del tamaño que tenga la muestra, o en otras palabras, del número de test que se hagan por lote. Si se incrementa el tamaño también lo hace la probabilidad de rechazar el lote, es aquí donde la corrección de *Bonferroni* recalcula el valor de *x* que mantiene minimizado el ratio de errores.

False Discovery Rate (FDR)

Los métodos descritos anteriormente se centran en corregir la inflación del el error de tipo I (*false positive rate*), es decir, la probabilidad de rechazar la hipótesis nula siendo esta cierta. Esta aproximación es útil cuando se emplea un número limitado de comparaciones. Para escenarios de *large-scale multiple testing* como los estudios genómicos en los que se realizan miles de test de forma simultánea, el resultado de estos métodos es demasiado conservativo e impide que se detecten diferencias reales. Una alternativa es controlar el *false discovery rate*.

El *false discovery rate (FDR)* se define como: (*todas las definiciones son equivalentes*)

- La proporción esperada de test en los que la hipótesis nula es cierta, de entre todos los test que se han considerado significativos.
- *FDR* es la probabilidad de que una hipótesis nula sea cierta habiendo sido rechazada por el test estadístico.
- De entre todos los test considerados significativos, el *FDR* es la proporción esperada de esos test para los que la hipótesis nula es verdadera.
- Es la proporción de test significativos que realmente no lo son.
- La proporción esperada de falsos positivos de entre todos los test considerados como significativos.

El objetivo de controlar el *false discovery rate* es establecer un límite de significancia para un conjunto de test tal que, de entre todos los test considerados como significativos, la proporción de hipótesis nulas verdaderas (falsos positivos) no supere un determinado valor. Otra ventaja añadida es su fácil interpretación, por ejemplo, si un estudio publica resultados estadísticamente significativos para un *FDR* del 10%, el lector tiene la seguridad de que como máximo un 10% de los resultados considerados como significativos son realmente falsos positivos.

Cuando un investigador emplea un nivel de significancia α, por ejemplo de 0.05, suele esperar cierta seguridad de que solo una pequeña fracción de los test significativos se correspondan con hipótesis nulas verdaderas (falsos positivos). Sin embargo, esto no tiene por qué ser así. La razón por la que un *false positive rate* bajo no tiene por qué traducirse en una probabilidad baja de hipótesis nulas verdaderas entre los test significativos (*false discovery rate*) se debe a que esta última depende de la frecuencia con la que la hipótesis nula contrastada es realmente verdadera. Un caso extremo sería el planteado en el ejemplo 1, en el que todas las hipótesis nulas son realmente ciertas por lo que el 100% de los test que en promedio resultan significativos son falsos positivos. Por lo tanto, la proporción de falsos positivos (*false discovery rate*) depende de la cantidad de hipótesis nulas que sean ciertas de entre todas los contrastes.

Los análisis de tipo exploratorio en los que el investigador trata de identificar resultados significativos sin apenas conocimiento previo se caracterizan por una proporción alta de hipótesis nulas falsas. Los análisis que se hacen para confirmar hipótesis, en los que el diseño se ha orientado en base a un conocimiento previo, suelen tener una proporción de hipótesis nulas verdaderas alta. Idealmente, si se conociera de antemano la proporción de hipótesis nulas verdaderas de entre todos los contrastes se podría ajustar con precisión el límite significancia adecuado a cada escenario, sin embargo, esto no ocurre en la realidad.

Inferencia sobre la probabilidad de hipótesis nula verdadera

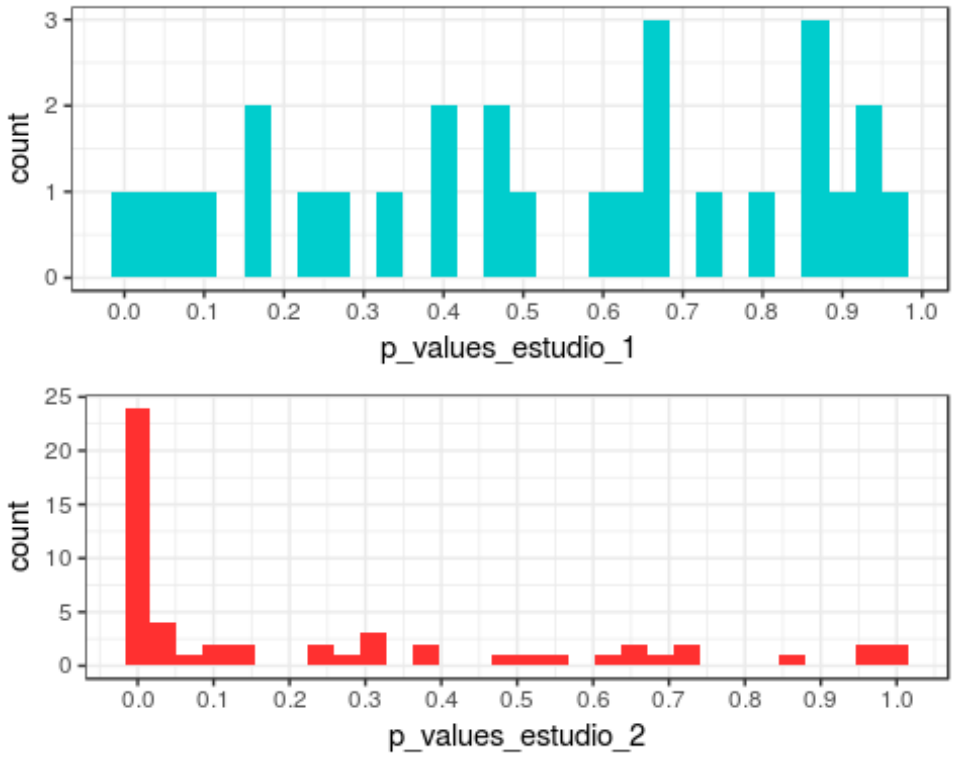
Cuando se realiza un único contraste de hipótesis, es difícil hacer inferencia sobre la probabilidad de que la hipótesis nula sea verdadera. Sin embargo, cuando se realizan múltiples contrastes, la distribución de los *p-values* proporciona información suficiente para hacerlo. Esto es así debido a que la distribución de *p-values* resultante de realizar múltiples test es una mezcla de dos componentes: la distribución de *p-values* de las hipótesis nulas verdaderas, que por definición es una distribución uniforme dentro del rango [0,1], y la distribución de *p-values* de las hipótesis nulas falsas, que es asimétrica con mayor densidad en valores cercanos al cero.

Los siguientes histogramas muestran la distribución de *p-values* de dos estudios distintos en los que se han realizado comparaciones múltiples (28 y 55 respectivamente). Se desconoce cuál es la probabilidad de que las hipótesis nulas contrastadas sean ciertas en cada uno de los escenarios, a pesar de ello, a partir de las distribuciones se puede inferir cual puede ser.

```
library(ggplot2)
library(gridExtra)
p_values_estudio_1 <- c(0.11, 0.8, 0.92, 0.68, 0.04, 0.15, 0.89, 0.47, 0.88,
    0.85, 0.17, 0.59, 0.4, 0.33, 0.97, 0.48, 0.85, 0.07, 0.66, 0.41, 0.64, 0.24,
    0.72, 0.004, 0.67, 0.51, 0.26, 0.94)
p1 <- ggplot(data = data.frame(p_values_estudio_1), mapping =
aes(p_values_estudio_1)) +
 geom_histogram(fill = "cyan3") +
 scale_x_continuous(breaks = seq(0, 1, by = 0.1)) +
 theme_bw()

p_values_estudio_2 <- c(0.001, 0.002, 1, 0.001, 0.001, 0.001, 0.25, 0.48, 0.09,
    0.51, 0.03, 0.01, 0.01, 0.15, 0.06, 0.85, 0.01, 0.003, 0.12, 0.01, 0.05,
    0.008, 0.38, 0.55, 0.95, 0.3, 0.3, 0.3, 0.66, 0.66, 0.38, 0.99, 0.26, 0.98,
    0.7, 0.72, 0.74, 0.001, 0.001, 0.009, 0.009, 0.25, 0.14, 0.61, 0.001, 0.001,
    0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.02, 0.02)
p2 <- ggplot(data = data.frame(p_values_estudio_2), mapping =
aes(p_values_estudio_2)) +
 geom_histogram(fill = "firebrick1") +
 scale_x_continuous(breaks = seq(0,1, by = 0.1)) +
 theme_bw()

grid.arrange(p1, p2)
```



En el primer estudio, la distribución de *p-values* es aproximadamente uniforme. Esta es la distribución de *p-values* esperada cuando la hipótesis nula es verdadera para todos los test. De hecho, el número de *p-values* significativos obtenidos (1) es muy próximo a la cantidad esperada ($0.05 * 28 = 1.4$) si se realizan 28 test empleando un $\alpha = 0.05$. En un escenario como este se tiene que establecer un nivel de significancia muy bajo para evitar rechazar hipótesis nulas verdaderas. El segundo estudio tiene una proporción de *p-values* próximos a cero mayor de lo que cabría esperar acorde a una distribución uniforme. Este tipo de distribución de *p-values* apunta que muchos de los *p-values* pequeños se corresponden con hipótesis nulas falsas. En escenarios como este, emplear niveles de significancia muy bajos provocaría que se rechazasen *p-values* que pertenecen a verdaderos positivos.

Teniendo en cuenta todas las observaciones anteriores, la mejor aproximación para determinar el nivel de significancia a emplear en cada estudio requiere de un método que se ajuste dependiendo de la probabilidad de hipótesis nulas verdaderas y no del número de test realizados.

FDR Benjamini & Hochberg (BH)

La primera aproximación para controlar el *FDR* fue descrita por Benjamini y Hochberg en 1995. Acorde a su publicación, si se desea controlar que en un estudio con n comparaciones el *FDR* no supere un porcentaje d hay que:

- Ordenar los n test de menor a mayor *p-value* ($p_1,\ p_2,\ \dots\ p_n$)
- Se define k como la última posición para la que se cumple que $p_i \leq d\frac{i}{n}$
- Se consideran significativos todos los *p-value* hasta la posición k ($p_1,\ p_2,\ \dots\ p_k$)

La principal ventaja de controlar el *false discovery rate* se hace más patente cuantas más comparaciones se realicen, por esta razón se suele emplear en situaciones con cientos o miles de comparaciones. Sin embargo, el método puede aplicarse también a estudios de menor envergadura.

El método propuesto por Benjamini & Hochberg asume a la hora de estimar el número de hipótesis nulas erróneamente consideradas falsas que todas las hipótesis nulas son ciertas. Como consecuencia, a estimación del *FDR* está inflada y por lo tanto es conservativa. A continuación se describen métodos más sofisticados que estiman la frecuencia de hipótesis nulas verdaderas a partir de la distribución de los *p-values*.

Ejemplo 1

Supóngase un estudio en el que se han realizado n=10 contrastes de hipótesis. Identificar que comparaciones son significativas controlando mediante BH que el *false discovery rate* no supere el 5% (d=0.05).

```
require(dplyr)
require(knitr)
# p_values obtenidos
p_values <- c(0.52, 0.07, 0.013, 1e-04, 0.26, 0.04, 0.01, 0.15, 0.03, 2e-04)
test <- paste("test", 1:10, sep = "_")
resultados <- data.frame(test, p_values)

# ordenación por p_values en sentido creciente
resultados <- arrange(resultados, p_values)

# añadir índices para utilizarlos en el cálculo
resultados <- mutate(resultados, indice = 1:length(p_values))

# cálculo de d*i/n
resultados <- mutate(resultados, `d*i/n` = 0.05 * (indice/length(p_values)))
```

```
# identificar p-values significativos
resultados <- mutate(resultados, significancia = p_values <= `d*i/n`)
kable(resultados, align = "c")
```

test	p_values	indice	d*i/n	significancia
test_4	0.0001	1	0.005	TRUE
test_10	0.0002	2	0.010	TRUE
test_7	0.0100	3	0.015	TRUE
test_3	0.0130	4	0.020	TRUE
test_9	0.0300	5	0.025	FALSE
test_6	0.0400	6	0.030	FALSE
test_2	0.0700	7	0.035	FALSE
test_8	0.1500	8	0.040	FALSE
test_5	0.2600	9	0.045	FALSE
test_1	0.5200	10	0.050	FALSE

Los 4 primeros test (test_4, test_10, test_7 y test_3) son significativos para un *false discovery rate* del 5%.

Las funciones de R p.adjust() y pairwais.t.test() permiten incorporar distintas correcciones, entre ellas la de *Benjamini & Hochberg (BH)*. La primera, recibiendo directamente un vector con los *p-values* y la segunda realizando primero los test de hipótesis mediante t-test. En ambos casos se devuelven los *p-values* originales multiplicados por el factor correspondiente acorde a la corrección. De esta forma se pueden comparar directamente frente al nivel de significancia α elegido inicialmente.

```
p_values_BH <- p.adjust(p = p_values, method = "BH")
names(p_values_BH) <- p_values_BH <= 0.05
p_values_BH
```

```
##      FALSE      FALSE       TRUE       TRUE      FALSE      FALSE
## 0.52000000 0.10000000 0.03250000 0.00100000 0.28888889 0.06666667
##       TRUE      FALSE      FALSE       TRUE
## 0.03250000 0.18750000 0.06000000 0.00100000
```

Si se consideran como significativos todos aquellos test cuyo valor devuelto por el método *BH* esté por debajo de un determinado límite apha, se tiene certeza de que el FDR$\leq \alpha$.

Ejemplo 2

Se emplean los *p-values* descritos gráficamente en la introducción del concepto de *FDR* para estudiar cómo se comporta este método en un escenario en el que la probabilidad de hipótesis nulas falsas es muy baja (estudios de confirmación de hipótesis) y en otro en el que es muy alta (estudios exploratorios). Se compara el número de eventos significativos obtenidos si no se aplica corrección alguna ($\alpha = 0.05$), con corrección de *Bonferroni* y con control *BH* para un *FDR* = 0.05.

```
p_values_estudio_1 <- c(0.11, 0.8, 0.92, 0.68, 0.04, 0.15, 0.89, 0.47, 0.88,
    0.85, 0.17, 0.59, 0.4, 0.33, 0.97, 0.48, 0.85, 0.07, 0.66, 0.41, 0.64, 0.24,
    0.72, 0.004, 0.67, 0.51, 0.26, 0.94)

n_1 <- length(p_values_estudio_1)
sin_correccion_1 <- sum(p_values_estudio_1 <= 0.05)
bonferroni_1 <- sum(p.adjust(p = p_values_estudio_1, method = "bonferroni") <=
    0.05)
BH_1 <- sum(p.adjust(p = p_values_estudio_1, method = "BH") <= 0.05)

kable(data.frame(Numero_de_p.values = n_1, Sin_corrección = sin_correccion_1,
    Bonferroni = bonferroni_1, Benjamini_Hochberg = BH_1), align = "c")
```

Numero_de_p.values	Sin_corrección	Bonferroni	Benjamini_Hochberg
28	2	0	0

```
p_values_estudio_2 <- c(0.001, 0.002, 1, 0.001, 0.001, 0.001, 0.25, 0.48, 0.09,
    0.51, 0.03, 0.01, 0.01, 0.15, 0.06, 0.85, 0.01, 0.003, 0.12, 0.01, 0.05,
    0.008, 0.38, 0.55, 0.95, 0.3, 0.3, 0.3, 0.66, 0.66, 0.38, 0.99, 0.26, 0.98,
    0.7, 0.72, 0.74, 0.001, 0.001, 0.009, 0.009, 0.25, 0.14, 0.61, 0.001, 0.001,
    0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001, 0.02, 0.02)

n_2 <- length(p_values_estudio_2)
sin_correccion_2 <- sum(p_values_estudio_2 <= 0.05)
bonferroni_2 <- sum(p.adjust(p = p_values_estudio_2, method = "bonferroni") <=
    0.05)
BH_2 <- sum(p.adjust(p = p_values_estudio_2, method = "BH") <= 0.05)

kable(data.frame(Numero_de_p.values = n_2, Sin_corrección = sin_correccion_2,
    Bonferroni = bonferroni_2, Benjamini_Hochberg = BH_2), align = "c")
```

Numero_de_p.values	Sin_corrección	Bonferroni	Benjamini_Hochberg
55	28	0	26

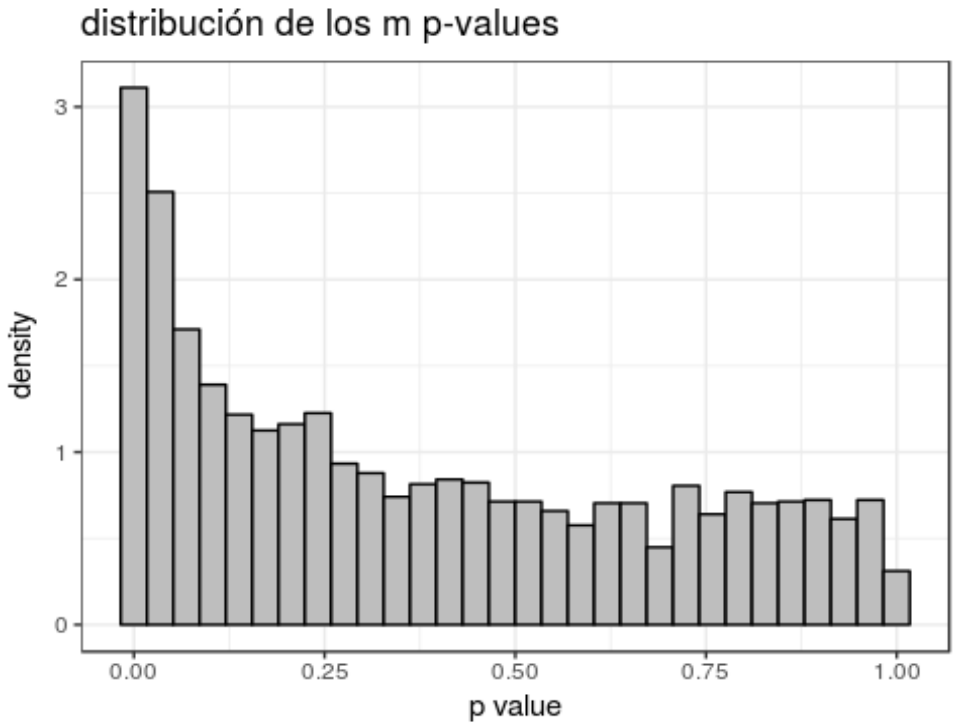
En el primer estudio, en el que la distribución de *p-values* es aproximadamente uniforme, tanto la corrección de *Bonferroni* como la de *BH* resultan en 0 contrastes significativos. Los *p-values* bajos tienen una alta probabilidad de corresponderse con hipótesis nulas verdaderas que, por azar, han resultado en valores pequeños. Por lo tanto, tiene sentido que no consideren como significativos. Para el segundo estudio, la corrección de *Bonferroni* no considera ninguno de los test como significativos, mientras que el resultado de la corrección de *BH* resulta en un número muy próximo al obtenido si no se aplica ninguna corrección. Dado que la frecuencia de *p-values* pequeños es alta, el método de *BH* reconoce que el ajuste tiene que ser mínimo ya que la gran mayoría de *p-values* significativos se corresponden con hipótesis nulas falsas.

Método q-Value

Este método para controlar el *FDR* fue desarrollado por John D. Storey y Robert Tibshirani en 2003. Tal y como se ha descrito anteriormente, el *FDR* es una medida de la precisión conjunta de todas las características/test consideradas significativas. El *q-value* es una extensión del concepto de *FDR* que, en lugar de estar asociado al conjunto de eventos significativos, puede asociarse a cada evento de forma individual (al igual que hace un *p-value*). El *q-value* se define como la proporción esperada de falsos positivos entre todas los test iguales o más extremos que el observado.

Supóngase que se realizan 3170 *t-test* obteniendo $m = 3170$ *p-values* cuya distribución se muestra en el siguiente histograma.

```
require(qvalue)
require(ggplot2)
data("hedenfalk")
ggplot(data = as.data.frame(hedenfalk$p), aes(hedenfalk$p)) +
 geom_histogram(aes(y = ..density..), fill = "grey", color = "black") +
 labs(title = "distribución de los m p-values", x = "p value") +
 theme_bw()
```



El primer paso es estimar el *FDR* obtenido si se consideran significativos aquellos test cuyo *p-value* es menor o igual a un determinado límite α, que tiene que estar comprendido en [0,1].

$$FDR = \frac{falsos\ positivos}{total\ positivos}$$

$$FDR = \frac{pvalue\ \leq\ \alpha\ , siendo\ la\ hipótesis\ nula\ verdadera}{pvalue\ \leq\ \alpha}$$

Dado que se desconoce tanto el número total de eventos significativos esperado como el número esperado de los mismos que son falsos positivos, es necesario recurrir a estimaciones. Una estimación simple y directa de *E(pvalue* $\leq \alpha$*)* es el número de *p-values* observados $\leq$ al límite de significancia α. El número de falsos positivos esperado *E(pvalue* $\leq \alpha$*, siendo la hipótesis nula cierta)* es por definición *número de hipótesis nulas verdaderas*$* \alpha$. Es aquí donde se complica la cosa, ya que de las 3170 hipótesis nulas contrastadas (m) se desconoce cuántas de ellas se corresponden con hipótesis nulas verdaderas (m_0).

Una forma de estimar la proporción de hipótesis nulas verdaderas (m_0) de entre todas las hipótesis nulas contrastadas (m), valor al que denominaremos π_0, es identificando en el histograma de *p-values* el valor λ a partir del cual la distribución es aproximadamente uniforme. En el ejemplo de *Hedenfalk et all* se podría considerar que esto ocurre a partir del valor 0.5, indicando que la región superior alberga principalmente p-values nulos.

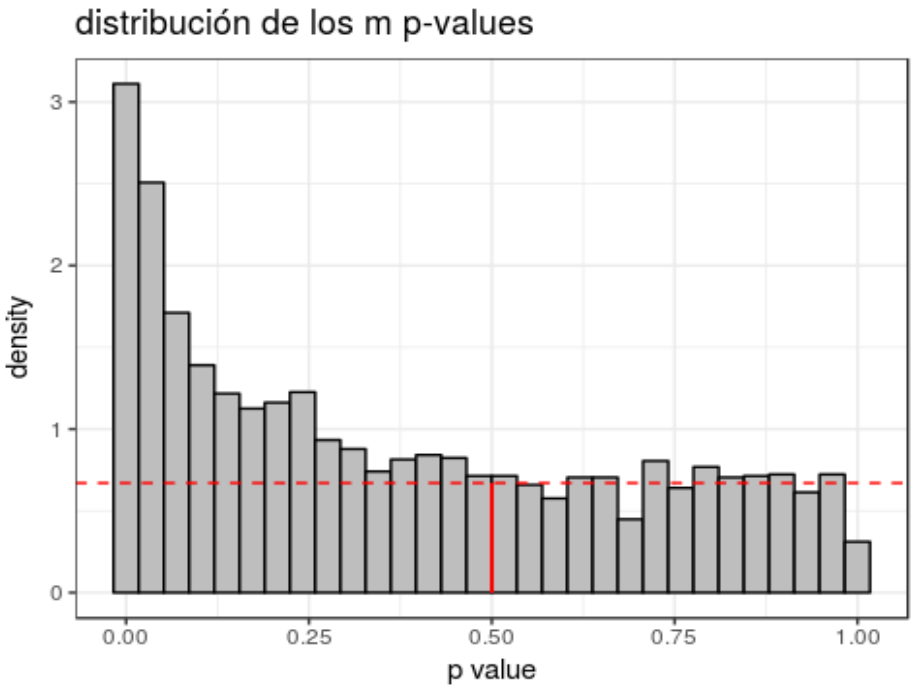
$$\hat{\pi}_0 = \frac{pvalues > \lambda}{m(1-\lambda)}$$

```
pi_0 <- sum(hedenfalk$p > 0.5)/(length(hedenfalk$p) * 0.5)
pi_0
```

```
## [1] 0.6763407
```

Acorde a la estimación de π_0 para $\lambda = 0.5$, el 67% de los genes no están diferencialmente expresados. Siendo del todo estrictos, en la región comprendida entre 0 y λ también hay unos pocos *p-values* que se corresponden con hipótesis nulas verdaderas, de ahí que $\hat{\pi}_0$ sea un estimador sesgado.

```
ggplot(data = as.data.frame(hedenfalk$p), aes(hedenfalk$p)) +
 geom_histogram(aes(y = ..density..), fill = "grey", color = "black") +
 labs(title = "distribución de los m p-values", x = "p value") +
 geom_hline(yintercept = 0.67, linetype = "dashed", color = "red") +
 geom_segment(aes(x = 0.5, y = 0, xend = 0.5, yend = 0.67), colour = "red") +
 theme_bw()
```



Introduciendo $\hat{\pi}_0$ en la ecuación de *FDR* se obtiene su estimación si se emplea α como nivel de significancia:

$$F\hat{D}R = \frac{\hat{\pi}_0 * m * \alpha}{pvalue \ \leq \ \alpha}$$

Finalmente, el *q-value* de cada test i se obtiene como el mínimo $F\hat{D}R$ que se puede obtener si se considerase a ese test y por consiguiente a todos lo que tengan un *p-value* menor como significativos.

$$\hat{q}\,(p_i) = min\, F\hat{D}R\,(\alpha)$$

Interpretación

Si el *q-value* de un gen *ZX* es de 0.013, significa que aproximadamente un 1.3% de los genes cuya diferencia de expresión es igual o más extrema que la observada para *ZX* son falsos positivos. Es importante no confundir esta última afirmación con la de que existe un 1.3% de probabilidad de que *ZX* sea un falso positivo.

Si se consideran como significativos todos aquellos test cuyo *q-value* esté por debajo de un determinado límite α, se tiene certeza de que el FDR$\leq \alpha$.

Comparación frente a Benjamini & Hochberg (BH)

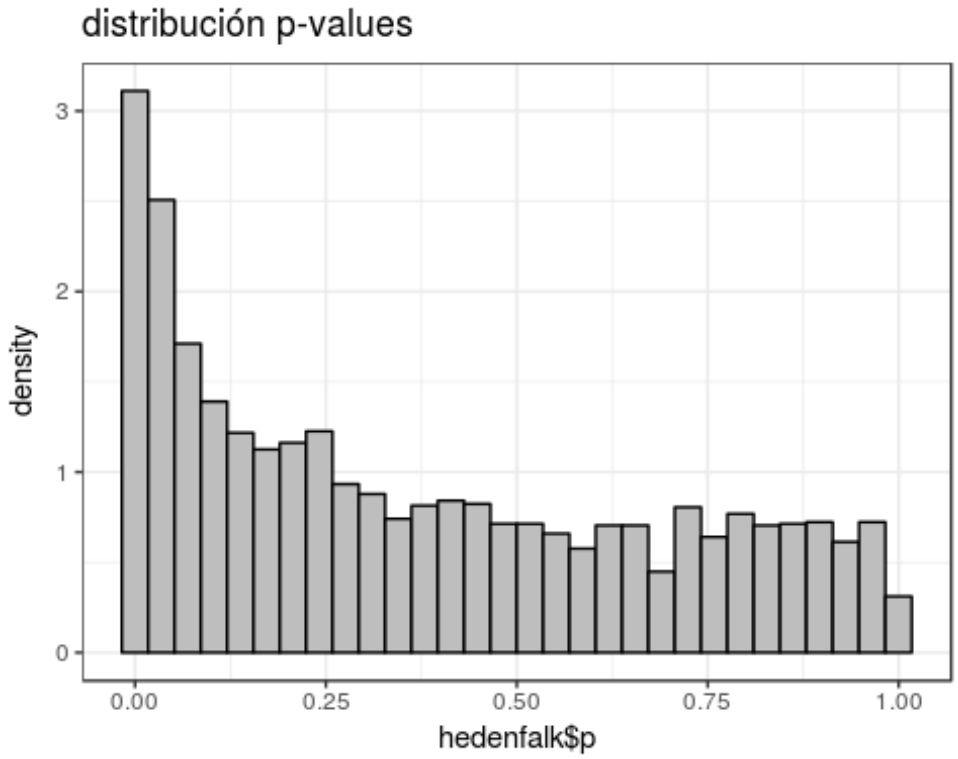
El método original descrito por Benjamini & Hochberg (BH) considera $\hat{\pi}_0 = 1$ por lo que es más conservador, llegando a ser excesivo para estudios genómicos con miles de comparaciones. Otra desventaja del método de BH frente al uso de *q-values* es que obliga al investigador a decidir un nivel de FDR antes de ver los resultados.

Ejemplo q-value

El paquete de R *qvalue* permite estimar a partir de un listado de *p-values* 3 tipos de valores: los *q-values*, la proporción de hipótesis nulas verdaderas (π_0) y *local false discovery rates*. Para mostrar cómo se emplea este paquete se van a emplear los datos publicados por *Hedenfalk et all, Gene expression profiles in hereditary Breast Cancer* https://research.nhgri.nih.gov/microarray/NEJM_Supplement/. El objetivo del estudio era identificar genes diferencialmente expresados entre tumores BRCA1 mutados y BRCA2 mutados. Para ello se midió mediante microarray la expresión de 3170 genes en 7 tumores BRCA1 y 8 BRCA2. Se realizó un *t-test* para cada uno de los genes generando los $m = 3170$ *p-values* que se encuentran almacenados en `hedenfalk$p`.

Dado que el cálculo de *q-values* se basa en gran medida en la distribución de los *p-values*, es recomendable representar estos últimos mediante un histograma y confirmar que se distribuyen de forma aproximadamente uniforme excepto en la región cercana a cero, donde hay un incremento de la frecuencia.

```
require(ggplot2)
require(qvalue)
ggplot(data = as.data.frame(hedenfalk$p), aes(hedenfalk$p)) +
 geom_histogram(aes(y = ..density..), fill = "grey", color = "black") +
 ggtitle("distribución p-values") +
 theme_bw()
```



Una vez confirmado que los *p-values* se distribuyen de forma adecuada, se emplea la función `qvalue().` Esta función tiene diferentes argumentos dependiendo de los cuales varía el resultado obtenido, los más importantes son:

- *p*: vector de *p-values*. Es el único argumento obligatorio.
- *fdr.level*: el nivel al que se quiere controlar el *false discovery rate*. En caso de utilizarlo, la función devuelve un vector lógico (TRUE o FALSE) especificando si cada *q-value* es menor o no que el valor de FDR introducido.
- *lambda* y *pio.method*: Estos dos parámetros determinan el valor y el método con el que se estima $\hat{\pi}_0$ (proporción total de hipótesis nulas verdaderas). Por defecto se emplea el método *smoother*.

Si se selecciona *lambda = 0* (por lo que $\hat{\pi}_0 = 0$) y *fdr.level = 0.05*, los resultados obtenidos son equivalentes a los obtenidos con el método de Benjamini & Hochberg (1995).

```
objeto_q <- qvalue(p = hedenfalk$p)
str(objeto_q)
```

```
## List of 8
##  $ call      : language qvalue(p = hedenfalk$p)
##  $ pi0       : num 0.67
##  $ qvalues   : num [1:3170] 0.0882 0.2094 0.668 0.1616 0.6325 ...
##  $ pvalues   : num [1:3170] 0.0121 0.075 0.9949 0.0418 0.8458 ...
##  $ lfdr      : num [1:3170] 0.168 0.413 1 0.309 1 ...
##  $ pi0.lambda: num [1:19] 0.852 0.807 0.782 0.756 0.731 ...
##  $ lambda    : num [1:19] 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 ...
##  $ pi0.smooth: num [1:19] 0.814 0.796 0.778 0.761 0.745 ...
##  - attr(*, "class")= chr "qvalue"
```

El objeto devuelto por la función `qvalue()` almacena la siguiente información:

- call: La llamada a la función.
- pi0: Estimación de la proporción de *p-values* nulos $\hat{\pi}_0$ (procedentes de hipótesis nulas verdaderas).
- qvalues: Vector con la estimación de los *q-values*.
- pvalues: Vector con los *p-values* originales.
- lfdr: Vector con la estimación de los FDR locales.
- significant: Si se ha especificado el argumento *fdr.level*, un indicador lógico de si el *q-value* estimado está por debajo de nivel de FDR elegido (si se consideran todos los *q-values* que cumplen esta condición como significativos, se controla el FDR a ese nivel.)
- pi0.lambda: Estimación de la proporción de *p-values* nulos para cada valor de lambda [0,1].
- lambda: Vector con los valores de lambda empleados para estimar $\hat{\pi}_0$.

El `summary()` de un objeto *qvalue* muestra un resumen muy informativo sobre la estimación de π_0 así como el número de genes que resultan significativos para diferentes *cutoffs*.

```
summary(objeto_q)
```

```
##
## Call:
## qvalue(p = hedenfalk$p)
##
## pi0: 0.669926
##
## Cumulative number of significant calls:
##
##           <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
## p-value       15     76   265    424   605  868 3170
## q-value        0      0     1     73   162  319 3170
## local FDR      0      0     3     30    85  167 2241
```

Dada la definición de *q-value*, mínimo FDR que se logra al considerar un test significativo, se recomienda reportar el *q-value* para cada test. Sin embargo, en muchas ocasiones es de interés conocer qué FDR se alcanza si se elige un determinado valor α como límite de significancia de *p-values* o bien el valor α que se tiene que elegir para no superar un determinado FDR.

Si en el caso de estudio aquí expuesto se quisiera estimar el *false discovery rate* que se tiene al consideran significativos los *p-values* < 0.01, solo hay que identificar el mayor *q-value* de entre todos los test cuyo p-value es ≤ 0.01.

```
max(objeto_q$qvalues[objeto_q$pvalues <= 0.01])
```

```
## [1] 0.07932935
```

Si se consideran significativos todos aquellos genes cuyo *p-value* es menor que 0.01, se espera que un 7.9% de ellos sean falsos positivos.

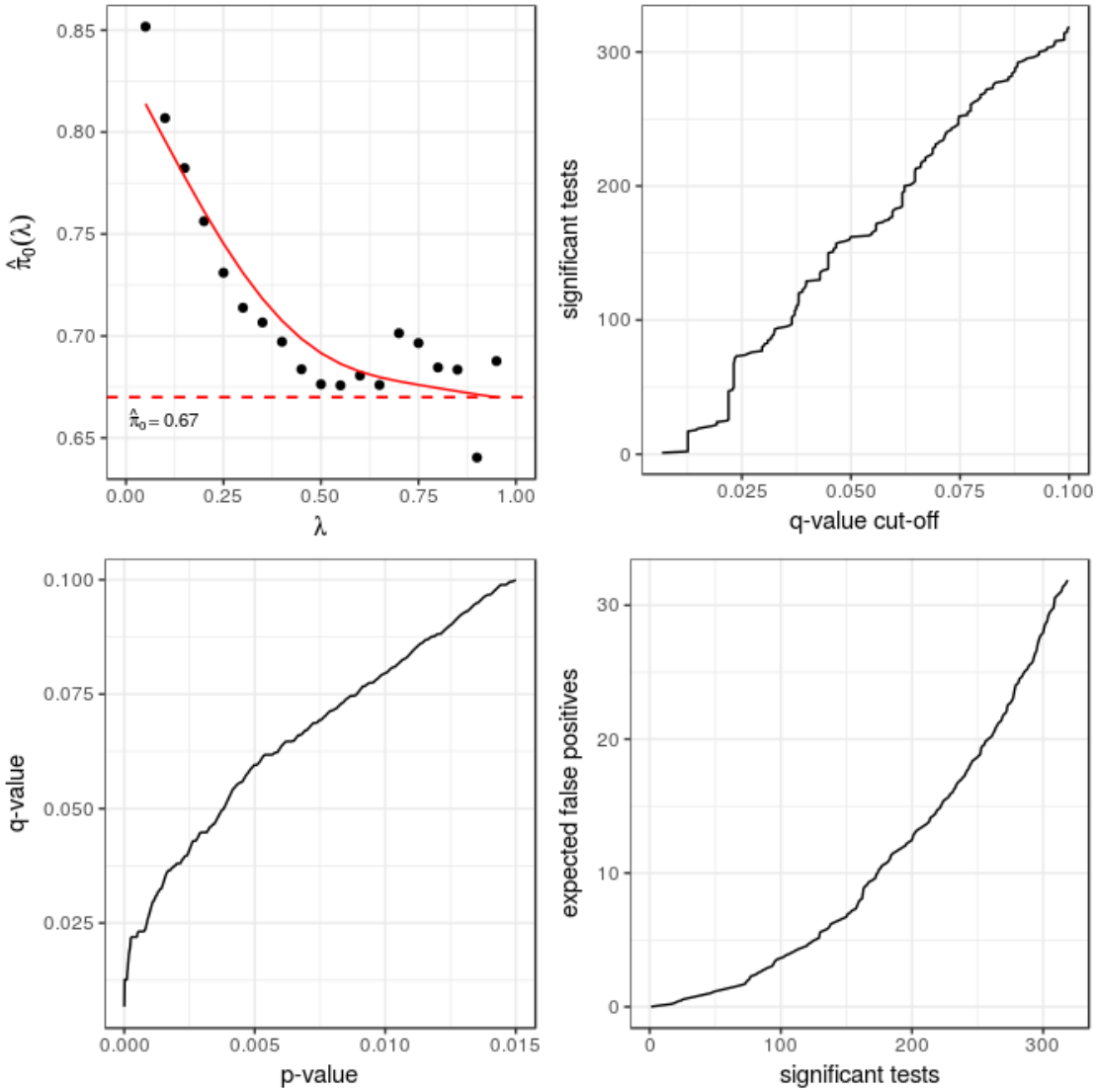
Mantener el FDR por debajo de un determinado nivel α se consigue considerando únicamente como significativos aquellos test para los que el *q-value* calculado sea menor o igual a α. Especificando en el argumento *fdr.level* el valor α se obtiene directamente el vector lógico `$significant` que indica cuales cumplen la condición.

```
head(qvalue(p = hedenfalk$p, fdr.level = 0.2)$significant)
```

```
## [1]  TRUE FALSE FALSE  TRUE FALSE FALSE
```

Por último, la representación gráfica de los resultados generados por la función `qvalue()` permite evaluar la estimación de π_0, el número de test significativos para cada límite de *q-value*, la evolución de los *q-value* vs *p-value* y el número de falsos positivos esperados en función del límite de *q-value* seleccionado.

```
plot(objeto_q)
```



Comparación entre controlar FWER y FDR

A la hora de decidir qué tipo de control (*family wise error* o *false discovery rate*) se quiere llevar a cabo, es importante diferenciar entre estudios exploratorios y estudios de confirmación de hipótesis. En los estudios exploratorios es de esperar que la proporción de hipótesis nulas falsas, es decir, de test que son realmente significativos, sea alta. Es frecuente

que estos estudios exploratorios estén seguidos de estudios de confirmación de modo que si se incluyen inicialmente falsos positivos estos serán detectados en los sucesivos estudios. Por lo tanto, lo más adecuado es controlar el *false discovery rate*. Los estudios de confirmación de hipótesis se emplean para contrastar ideas que están fundadas en conocimientos o datos previos, por lo que se suelen estar diseñados de forma que solo resulten significativos si se confirma esa idea y que aseguren que el riesgo de un falso positivo sea mínimo. Para estos es adecuado ser más estricto por lo que se prefiere controlar el *family wise error*.

Bibliografía

Wikipedia, Multiple comparisons problem

False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies, Mark E. Glickman

J. D. Storey. False discovery rates.Springer, 2011. http://genomine.org/papers/Storey_FDR_2011.pdf.

Statistical significance for genomewide studies, John D. Storey and Robert Tibshirani

What's wrong with Bonferroni adjustments, Thomas V Perneger, medical epidemiologist

Local False Discovery Rates, Bradley Efron

Bioconductor's qvalue package Version 2.7.0, John D. Storey and Andrew J. Bass Princeton University