# Comparaciones múltiples

Guillermo Ayala Gallego

2025-05-21

## Table of contents



### Comparaciones múltiples

- Una formulación ($i = 1, \dots, N$):
  - $H_i$: El gen i **no tiene** una **expresión diferencial** entre las condiciones consideradas.
  - $K_i$: El gen i **tiene** una **expresión diferencial** entre las condiciones consideradas.
- Quizás es mejor:

- $H_i$: La expresión del gen i no tiene asociación con la condición.
- $K_i$: La expresión del gen i tiene asociación con la condición.

### Benjamini and Hochberg (1995)

| Hipótesis nula | No rechazadas | Rechazadas | Total |
|---|---|---|---|
| Verdadera | $U$ | $V$ | $N_0$ |
| Falsa | $T$ | $S$ | $N - N_0 = N_1$ |
| Total | $N - R$ | $R$ | $N$ |

- ¿Qué conocemos? Solamente la variable $R$ y el número de hipótesis $N$.

## Tasas de error

### FWER: Family wise error rate

- Se define como:
$$FWER = P(V > 0) = P(V \geq 1).$$
- Es la tasa de error de uso clásico en Estadística.
- Correcta para un número pequeño de hipótesis.
- Muy exigente con un número grande de hipótesis.
- No queremos cometer al menos un error cuando tenemos miles de contrastes.
- **Posiblemente** genes con expresión diferencial no serán detectados.

### FDR: false discovery rate

- **Tasa de falsamente rechazados** o **tasa de falso rechazo**.
- Consideramos la proporción de test erróneamente rechazados
$$Q = \frac{V}{R}$$
si $R > 0$ y $Q = 0$ en otro caso.

- $$FDR = E(Q) = E\left(\frac{V}{R}\middle| R > 0\right) P(R > 0).$$
- Una modificación importante: **pFDR** (Positive false discovery rate)

$$pFDR = E\left(\frac{V}{R}\middle| R > 0\right).$$

## Relación entre FWER y FDR

$$FDR \leq FWER.$$

## Ajuste de p-valores

- Por contraste $H_i$ tenemos un p-valor $p_i$.
- Si $p_i \leq \alpha$ rechazamos $H_i$ (con nivel de significación $\alpha$).
- Cuando consideramos simultáneamente todos los tests el valor $\alpha$ cambia por gen.
- Tendríamos que comprobar $p_i \leq \alpha_i$.
- Ajustar un p-valor $p_i$ es transformarlo en otro valor $\tilde{p}_i$ de modo que:

$$p_i \leq \alpha_i \iff \tilde{p}_i \leq \alpha$$

.

## Método de Bonferroni

- Rechazamos $H_i$ si
$$p_i \leq \frac{\alpha}{N}.$$
- El p-valor ajustado sería:
$$\tilde{p}_i = \min\{Np_i, 1\}$$
- Rechazamos $H_i$ si
$$\tilde{p}_i \leq \alpha.$$

### Método de Benjamini-Hochberg

1. Fijamos la tasa de error $\alpha$.
2. Para cada i (gen) aplicamos un test y obtenemos un p-valor $p_i$.
3. Ordenamos los p-valores
$$p_{r_1} \leq ... \leq p_{r_N}.$$
4. Sea
$$i^* = \max\{i : p_{r_i} \leq \frac{i}{N}\alpha\}$$
5. Rechazamos $H_{r_i}$ para $i = 1, ..., i^*$.
6. Si no existe $i^*$ entonces no rechazamos ninguna hipótesis.

### Método de Benjamini y Yekutieli

- Como en el anterior, $p_{r_1} \leq ... \leq p_{r_N}$ son los p-valores originales ordenados.
- Los p-valores ajustados se definen como
$$\tilde{p}_{r_i} = \min_{k=i,...,N} \left\{ \min \{ \frac{N \sum_{j=1}^{N} 1/j}{k} p_{r_k}, 1 \} \right\}.$$

### Un ejemplo: gse21942

- Leemos datos.

```
pacman::p_load(Biobase)
data(gse21942,package="tamidata")
```

- Aplicamos los t-tests para cada gen.

```
tt = genefilter::rowttests(gse21942,
                           pData(gse21942)[,"FactorValue..DISEASE.STATE."])
```

Con un nivel de significación de $\alpha = 0.01$ tendríamos el siguiente número de características significativas.

```
table(tt$p.value<=.01)
```

```
FALSE  TRUE
17431  3927
```

- Utilizamos el método de Benjamini-Hochberg.

```
p.BH = p.adjust(tt$p.value,method = "BH")
```

- ¿Cuántos son significativos con $\alpha = 0.01$

```
table(p.BH<.01)
```

```
FALSE  TRUE
19178  2180
```

## q-valor

### Definición

- Propuesto por Storey (2002), Storey and Tibshirani (2003)
- Si consideramos un test determinado nos da la proporción esperada de falsos positivos en la que incurrimos cuando declaramos significativo ese test.

### Estimación del q-valor

- Fijamos un valor $t$ y consideremos que rechazamos $H_i$ cuando $P_i \leq t$.
- Definimos:

$$V(t) = |\{P_i : P_i \leq t; H_i \text{ es cierta}; i = 1, \dots, N\}|$$

y

$$R(t) = |\{P_i : P_i \leq t; i = 1, \dots, N\}|$$

- pFDR se puede aproximar con

$$E\left[\frac{V(t)}{R(t)}\right] \approx \frac{EV(t)}{ER(t)}.$$

- Estimamos
$$\hat{R}(t) = |\{p_i : p_i \leq t\}|$$
- Además
$$EV(t) = N_0 t.$$
- Estimamos $\pi_0 = N_0/N$ con
$$\hat{\pi}_0 = \frac{|\{p_i : p_i > \lambda; i = 1, \ldots, N\}|}{N(1-\lambda)}.$$

- Podemos pues estimar pFDR con
$$\widehat{pFDR} = \frac{\hat{\pi}_0 N t}{|\{p_i : p_i \leq t\}|}.$$
- El q-valor asociado a un contraste sería el mínimo valor de pFDR que se alcanza cuando el contraste es rechazado.
- El q-valor asociado al test $i$-ésimo sería
$$q(p_i) = \min_{t \geq p_i} pFDR(t)$$
y su estimador sería
$$\hat{q}(p_i) = \min_{t \geq p_i} \widehat{pFDR}(t).$$

### q-valor y datos tamidata::gse21942

- Calculamos p-valores originales.

```
tt = genefilter::rowttests(gse21942,
                           pData(gse21942)[,"FactorValue..DISEASE.STATE."])
pvalue = tt$p.value
```

- Calculamos los q-valores

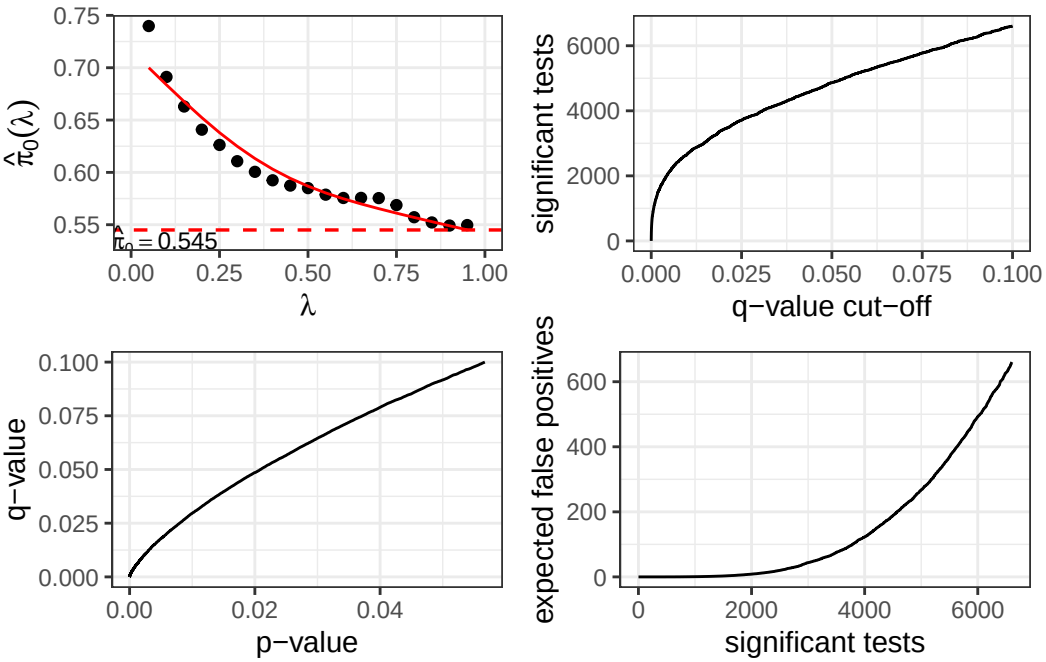
```
library(qvalue)
aa = qvalue(pvalue)
```

- Si simplemente queremos los q-valores los obtenemos con

```
q.value = qvalue(pvalue)$qvalues
```

### Dibujos asociados al q-valor

```
plot(aa)
```




### Bibliografía


Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. http://www.jstor.org/stable/2346101.

Storey, John D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3): 479–98. https://doi.org/10.1111/1467-9868.00346.

Storey, John D., and Robert Tibshirani. 2003. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences* 100 (16): 9440–45. https://doi.org/10.1073/pnas.1530509100.