

Comparaciones múltiples

Guillermo Ayala Gallego

2024-04-09

Table of contents

Comparaciones múltiples	1
Benjamini and Hochberg (1995)	2
Tasas de error	2
FWER: Family wise error rate	2
FDR: false discovery rate	2
Relación entre las tasas de error tipo I	3
p-valor ajustado	3
Método de Bonferroni	3
Método de Benjamini-Hochberg	4
Método de Benjamini y Yekutieli	4
Lo aplicamos con dos grupos	4
q-valor	7
Definición	7
Estimación del q-valor	7
q-valor y datos tamidata::gse21942	8
Dibujos asociados al q-valor	8

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.
- Es un criterio muy exigente si tenemos un número de hipótesis muy grande.
- Notemos que nos fijamos en cometer al menos un error cuando con frecuencia tendremos decenas de miles de contrastes.
- Como (mala) contrapartida muchos genes que tienen una expresión diferencial realmente no serán detectados.

FDR: false discovery rate

- O tasa de falsamente rechazados
- Definimos: $Q = \frac{V}{R}$ si $R > 0$ y $Q = 0$ en otro caso.
- Proporción de tests erróneamente rechazados.

-

$$FDR = E(Q) = E\left(\frac{V}{R} \mid R > 0\right)P(R > 0).$$

- Una modificación importante: **pFDR** (Positive false discovery rate)

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

Relación entre las tasas de error tipo I

- Se verifican las siguientes desigualdades:

$$FDR \leq FWER.$$

p-valor ajustado

- Para cada contraste H_i tendremos un p-valor p_i .
- Podemos contrastar el contraste H_i con un nivel de significación α_i rechazando la hipótesis nula H_i si $p_i \leq \alpha_i$ y no rechazando en otro caso.
- Cuando consideramos simultáneamente todos los tests los valores α_i serán distintos y por ello tendríamos que ir comprobando si la desigualdad $p_i \leq \alpha_i$ se verifica o no.
- La idea del p-valor ajustado es transformar el p-valor p_i en otro valor \tilde{p}_i , el i-ésimo p-valor ajustado, de modo que sean equivalentes:

$$p_i \leq \alpha_i$$

y

$$\tilde{p}_i \leq \alpha$$

siendo α el valor que especificamos para controlar alguna de las tasas de error tipo I.

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 α .
2. Para cada i (gen) aplicamos un test y obtenemos un p-valor p_i .
3. Ordenamos los p-valores

$$p_{r_1} \leq \dots \leq p_{r_N}.$$

4. Sea

$$i^* = \max\{i : p_{r_i} \leq \frac{i}{N}\alpha\}$$

5. Rechazamos H_{r_i} para $i = 1, \dots, i^*$.
6. Si no existe i^* entonces no rechazamos ninguna hipótesis.

Método de Benjamini y Yekutieli

- Como en el anterior, $p_{r_1} \leq \dots \leq p_{r_N}$ son los p-valores originales ordenados.
- Los p-valores ajustados se definen como

$$\tilde{p}_{r_i} = \min_{k=i, \dots, N} \left\{ \min \left\{ \frac{N \sum_{j=1}^N 1/j}{k} p_{r_k}, 1 \right\} \right\}.$$

Lo aplicamos con dos grupos

- Leemos datos.

```
pacman::p_load(Biobase)
data(gse21942, package="tamidata")
y = pData(gse21942)[, "FactorValue..DISEASE.STATE."]
head(exprs(gse21942))
```

	GSM545846.CEL	GSM545845.CEL	GSM545844.CEL	GSM545843.CEL	GSM545842.CEL
1007_s_at	6.701185	6.770585	6.896115	6.913170	7.055230
1053_at	6.904241	7.160595	6.986157	7.252437	6.820581
117_at	8.347887	8.309832	8.042841	7.973542	7.905076
121_at	7.527261	7.732676	7.534203	7.582493	7.624345
1255_g_at	2.741237	2.785420	2.705875	2.712574	2.874974
1294_at	8.512001	8.697264	8.471296	8.609792	8.932694
	GSM545841.CEL	GSM545840.CEL	GSM545839.CEL	GSM545838.CEL	GSM545837.CEL
1007_s_at	7.090447	7.161786	6.990521	7.143398	7.019538
1053_at	6.983571	6.801948	7.131738	6.941465	6.829286
117_at	7.656960	7.661755	8.849415	7.781422	7.431034

121_at	7.462491	7.757673	7.771086	7.728931	7.507087
1255_g_at	2.731671	3.032517	2.957291	2.908258	2.801843
1294_at	8.466908	8.613721	8.996203	8.312664	8.444103
GSM545836.CEL GSM545835.CEL GSM545834.CEL GSM545833.CEL GSM545832.CEL					
1007_s_at	7.263345	7.183820	7.068422	7.018434	7.134651
1053_at	7.227540	6.779245	7.141372	6.908670	7.484561
117_at	8.128334	7.638504	8.336013	8.466568	8.495304
121_at	7.604226	7.721856	7.660064	7.735517	7.514174
1255_g_at	2.949537	2.747506	2.634829	2.995889	2.631472
1294_at	9.093238	9.048817	9.141685	8.805013	8.557823
GSM545831.CEL GSM545830.CEL GSM545829.CEL GSM545828.CEL GSM545827.CEL					
1007_s_at	7.164015	7.079536	6.879690	6.577880	6.928483
1053_at	6.997132	6.890387	7.179972	7.303770	7.410144
117_at	8.036675	8.228937	7.237569	8.042661	8.141817
121_at	7.553159	7.623796	7.507285	7.698743	7.696693
1255_g_at	2.708215	2.840438	2.723045	2.651039	2.906055
1294_at	8.761252	8.737342	8.528587	8.587533	8.761697
GSM545826.CEL GSM545825.CEL GSM545824.CEL GSM545823.CEL GSM545822.CEL					
1007_s_at	6.653279	6.900510	6.436211	7.008971	6.834951
1053_at	7.120752	7.109658	7.085381	7.046270	7.324719
117_at	8.033017	7.955324	8.457568	7.863781	7.965000
121_at	7.522826	7.725928	7.733425	7.846924	7.534118
1255_g_at	2.803210	2.889357	2.781139	2.868890	2.632601
1294_at	8.669705	8.470591	8.596620	8.660778	8.730843
GSM545821.CEL GSM545820.CEL GSM545819.CEL GSM545818.CEL					
1007_s_at	6.747209	7.052194	6.785745	6.715101	
1053_at	7.176329	7.224084	7.369654	7.277721	
117_at	7.776154	7.825469	8.342164	7.998972	
121_at	7.443340	7.714347	7.434026	7.699178	
1255_g_at	2.947213	2.765574	2.898040	2.709309	
1294_at	8.648626	8.640482	8.730575	8.606553	

```
dim(gse21942)
```

```
Features Samples
54675      29
```

- Aplicamos los t-tests para cada gen.

```
tt = genefilter::rowttests(gse21942,y)
head(tt)
```

	statistic	dm	p.value
1007_s_at	-2.29869931	-0.15981906	0.029492008
1053_at	3.44084102	0.20940361	0.001901206
117_at	-0.08505071	-0.01110444	0.932848609
121_at	-0.53362792	-0.02274832	0.597965094
1255_g_at	-1.01536731	-0.04339509	0.318943716
1294_at	-1.05030996	-0.07873761	0.302886126

- Los p-valores originales los guardamos en p0.

```
p0 = tt$p.value
head(p0)
```

```
[1] 0.029492008 0.001901206 0.932848609 0.597965094 0.318943716 0.302886126
```

o bien

```
p0 = tt[,"p.value"]
```

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

```
table(p0<=.01)
```

```
FALSE TRUE
44014 10661
```

- Utilizamos el método de Benjamini-Hochberg.

```
p.BH = p.adjust(p0,method = "BH")
```

- Los p-valores ajustados son los siguientes

```
View(cbind(p0,p.BH))
head(p.BH)
```

```
[1] 0.11049371 0.01513959 0.96675733 0.76778314 0.53372043 0.51746083
```

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

```
FALSE TRUE  
48702 5973
```

q-valor

Definición

- Propuesto por Storey (2002)
- 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

$$\hat{\pi}_0 = \frac{|\{p_i : p_i > \lambda; i = 1, \dots, N\}|}{N(1 - \lambda)}.$$

- Podemos pues estimar pFDR con

$$p\widehat{FDR} = \frac{\hat{\pi}_0 N t}{|\{p_i : p_i \leq t\}|}$$

- 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

$$\hat{q}(p_i) = \min_{t \geq p_i} p\widehat{FDR}(t).$$

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)
```


