9.5 Selección de genes expresados diferencialmente: determinación de un punto de corte.

Hemos visto diferentes propuestas para la elección del estadístico en base al cual se ordenarán los genes de acuerdo a la evidencia de expresión diferencial, desde la más débil a la más fuerte.

La importancia principal de este ordenamiento surge del hecho que solamente una cantidad limitada de genes puede seguirse en experimentos biológicos típicos para su confirmación y estudios posteriores.

En la mayoría de las veces será práctico continuar con una cantidad limitada de genes del orden de unos cientos. Por esta razón es importante identificar a los 100 candidatos más probables de estar diferencialmente expresados. La lista completa de genes que pueden considerarse DE estadísticamente significativos puede ser menos interesante si ésta es muy grande para su seguimiento (Smyth et al. 2003).

Una vez que se han ordenado los genes en base a un estadístico adecuado, el paso siguiente consiste en hallar un punto de corte por encima del cual los genes serán identificados como significativos.

La cuestión crucial en este punto es el control del nivel global inherente a la necesidad de realizar un test para cada gen.

9.6 Tipos de errores en tests de hipótesis, para un único test.

Consideremos, para un gen en particular, las siguientes hipótesis:

H_0: el gen no está DE vs. **H_1:** el gen sí está DE

	Decisión	
Realidad	**H_0**	**H_1**
H_0	bien!	error de tipo I
H_1	error de tipo II	bien!

Cuando se testea una única hipótesis como **H_0** se pueden cometer dos tipos de errores:

Rechazar **H_0** cuando **H_0** es verdadera: error de tipo I ó falso positivo
No rechazar **H_0** cuando **H_0** es falsa: error de tipo II ó falso negativo

Habitualmente se controla la probabilidad de error de tipo I, es decir se fija *un nivel* α de manera que

$$P(\text{error de tipo I}) \leq \alpha$$

y dentro de una familia de posibles tests con nivel α se elige el que tiene menor probabilidad de error de tipo II (mayor potencia). Fijado un test y el nivel es necesario aumentar el tamaño de la muestra para controlar la probabilidad de cometer errores de tipo II (falsos negativos), de manera de asegurarse suficiente potencia para detectar verdaderos DE.

Estadístico del test: es el estadístico en base al cual se toma la decisión, lo llamamos T.

Región de rechazo a nivel α, son los valores del estadístico que resultan en rechazo:

$$|T| \geq c(\alpha),$$

con $P(|T| \geq c(\alpha)$ / cuando $\mathbf{H_0}$ es verdadera) = α.

> Este tipo de regiones de rechazo son las utilizadas para detectar genes estadísticamente DE. Son llamadas a *dos colas*, porque tanto valores positivos como negativos son evidencia a favor de la hipótesis alternativa $\mathbf{H_1}$**.** Un valor negativo del estadístico corresponderá a un gen expresado diferencialmente hacia abajo (down-regulated) y un valor positivo a uno expresado diferencialmente hacia arriba (up-regulated).

p-valor, es el menor nivel para el cual el test resultaría en rechazo para los datos observados.

¿Cómo se calcula el p-valor?

Llamemos $t_{observado}$ al valor que resulta de reemplazar los datos en la expresión del estadístico *T*, entonces

$$\text{p-valor } (t_{observado}) = P(|T| \geq |t_{observado}|).$$

El p-valor es la probabilidad de obtener un valor tan ó más extremo que el valor observado del estadístico del test, cuando la $\mathbf{H_0}$ es verdadera.

Decisión utilizando p-valores El test resulta en rechazo si p-valor $(t_{observado}) \leq \alpha$

10. Tests múltiples

Consideremos un experimento de microarreglos que provee datos sobre los niveles de expresión de *m* genes (estos son las variables o features) para *n* muestras de mRNA (es decir observaciones). Supongamos también que se registra para cada una de las muestras una covariable de interés.

Ejemplo: de un estudio sobre el nivel de expresión de genes en biopsias tumorales de pacientes con leucemia la covariable o variable respuesta es el tipo de tumor y el objetivo es identificar genes que están expresados diferencialmente entre los diferentes tipos de tumores.

Para la muestra *i* los datos consisten de una variable respuesta o covariable y_i y un perfil de expresiones de genes $\mathbf{x}_i = (x_{1i}, \ldots, x_{mi})$, donde x_{ji} denota la medida de expresión del gen *j* en la muestra *i*, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

Los niveles de expresión están guardados en una matriz $X = (x_{ji})$ de $m \times n$, en la cual las filas (j:1,...., m) corresponden a genes y las columnas (i = 1,...,n) corresponden a las muestras individuales de mRNA.

Hemos visto que en un experimento típico el total *n* de muetras se encuentra entre 10 y algunos cientos, mientras que la cantidad *m* de genes es de varios miles. Las medidas de los niveles de expresión *x*, son en general variables continuas. Las variables respuesta o covariables *y*, pueden ser categóricas (ej. tratamiento/control, tipo de célula) ó continuas (dosis de una droga, tiempo). Las respuestas podrían ser censuradas como los tiempos de supervivencia o algún otro resultado clínico.

Para cada muestra i , los pares $\{(\mathbf{x}_i, y_i)\}$ $i=1,...,n$, formados por los perfiles de expresión $\mathbf{x}_i$ y las respuesta o covariables yi, se consideran una muestra aleatoria de una población de interés.

Sean X_j e Y respectivamente, las variables aleatorias que corresponden a la medición de expresión del gen j , $j = 1, \ldots, m$, y la respuesta o covariable.

El problema de determinar qué genes están expresados diferencialmente puede plantearse como un problema de testear múltiples hipótesis.

Para cada gen j interesa testear la hipótesis H_{0j} de no asociación entre la medida de expresión X_j y la variable respuesta Y .

Por ejemplo la hipótesis nula puede ser la igualdad de los niveles medios de expresión en dos poblaciones de células.

Para realizar el procedimiento de testeo múltiple es necesario
1) calcular un estadístico del test T_j para cada gen j, y
2) aplicar un procedimiento de testeo múltiple para determinar que hipótesis rechazar pero controlando una tasa de error de tipo I adecuada.

No trataremos el tema de elección del estadístico del test en esta sección. Consideramos que para cada gen j se realiza un test para la hipótesis nula H_{0j} basado en el estadístico T_j que es una función de X_j e Y . Una realización de la variable aleatoria T_j se denota por las letras minúsculas t_j como es habitual. Para simplificar, supondremos en lo que sigue que la hipótesis nula H_{0j} es rechazada para valores grandes de of $|T_j|$ (hipótesis bilateral).

10.1 Planteo del problema

Consideremos el problema de testear simultáneamente *m* hipótesis nulas

$H_{0\,j}$, $j = 1, \ldots, m$,

e indiquemos por *R* la cantidad de hipótesis rechazadas. La situación puede resumirse en la Tabla 1 (Benjamini and Hochberg, 1995). La cantidad *m* hipótesis a testear es fija y conocida de antemano. Los conjuntos de subíndices que corresponden a hipótesis nulas verdaderas y falsas, $\Delta_0 = \{j : H_{0j} \text{ es verdadera}\}$ y $\Delta_1 = \{j : H_{0j} \text{ no es verdadera}\}$ son parámetros desconocidos. El conjunto total de índices es $\Delta = \{1,2,\ldots,m\} = \Delta_0 \cup \Delta_1$.

Las cantidades de hipótesis nulas verdaderas, $m_0 = \#\ \Delta_0$ y falsas, $m_1 = m - m_0 = \#\ \Delta_1$, son cantidades fijas pero desconocidas.

Tabla 1

Realidad **Cantidad de**	Decisión		Total
	H_0 **Cantidad de hipótesis no rechazadas**	**H_1** **Cantidad de hipótesis rechazadas**	
H_0 Hipótesis nulas verd: m_0	U	V (Falsos +)	m_0
H_1 Hipótesis nulas falsas: m_1	T (Falsos -)	S	m_1
Total	**m-R**	**R**	**m**

La cantidad de hipótesis nulas rechazadas R y no rechazadas m-R son variables aleatorias observables y S, T, U y V son variables aleatorias no observables.

Tenemos una hipótesis nula H_{0j} para cada gen j y el rechazo de H_{0j} corresponde a declarar que el gen j está expresado diferencialmente. Idealmente querríamos minimizar la cantidad V de *falsos positivos*, o *errores de tipo I* y la cantidad T de *falsos negativos*, o *errores de tipo II.*

Lo habitual en tests de hipótesis es preespecificar un nivel α para la probabilidad de Error de Tipo I y hallar tests que minimicen la probabilidad de Error de Tipo II, es decir maximizar la potencia (tests uniformemente más potentes) dentro de una clase de tests con probabilidad de Error de Tipo I a lo sumo α.

Cuando se testea una única hipótesis, digamos H_{01}, el control de la probabilidad de Error de Tipo I se logra en general eligiendo un valor crítico $c\alpha$ tal que
$P\ (|T_1| \geq c\alpha \mid$ cuando H_{01} es verdadera$) \leq \alpha$ y rechazando H_{01} cuando $|T_1| \geq c\alpha$.

10.2 Tasas de error de tipo I - falsos positivos- Type I error rates

$$V = \sum_{j \in \Lambda_0} V_j, \quad V_j = \begin{cases} 1 & \text{rechazo} \quad H_{0,j} \text{ con } H_{0,j} \text{ verdadera} \\ 0 & \text{no rechazo} \quad H_{0,j} \quad \text{con } H_{0,j} \text{ verdadera} \end{cases}, \; p_{Vj} = P(V_j = 1, H_{0,j} \text{ verdadera})$$

donde Δ_0 identifica al subconjuto de índices de las hipótesis nulas verdaderas, $\#\ \Delta_0 = m_0$.

Existe una gran variedad de generalizaciones para tests múltiples de la probabilidad de error de tipo I. Describiremos las más habituales (Shaffer, 1995).

- *Tasa de error por familia. The per-family error rate* (PFER). Se define como la cantidad esperada de errores tipo I. PFER = E(V): número esperado de falsos positivos

- *Tasa de error por comparación. The per-comparison error rate* (PCER). Se define como la cantidad esperada de errores tipo I dividida por la cantidad de hipótesis, esto es PCER = E(V)/m: proporción esperada de falsos positivos
- *Family-wise Error Rate* - probabilidad de tener por lo menos un error de tipo I, esto es FWER = P(V ≥ 1) (probabilidad de al menos un falso positivo)
- *False Discovery Rate* (FDR) de Benjamini y Hochberg (1995) es la proporción esperada de errores de tipo I entre las hipótesis rechazadas, esto es FDR = E(Q), siendo Q = V/R si R > 0 y 0 si R=0. Es la proporcion esperada de falsos positivos entre las pruebas que fueron significativas: E[V/R |R>0] P(R>0)
- *positive False Discovery Rate* (pFDR). Si interesa únicamente estimar una tasa de error cuando han ocurrido descubrimientos positivos, entonces es adecudado utilizar el pFDR de Storey (2002): E[V/R |R>0]

10.3 Control de la tasa de error

Se dice que un procedimiento de testeo múltiple controla una Tasa de error de tipo I particular, a *nivel* α, si esa tasa de error es menor o igual a α cuando el procedimiento es aplicado para producir una lista de R hipótesis rechazadas.

Es importante observar que las probabilidades y esperanzas anteriores son condicionales a cuales son las hipótesis nulas verdaderas, es decir que depende del subconjunto $\Delta_0 \subseteq \{1, \ldots, m\}$.

10.3 Tipos de control de la tasa de error en hipótesis múltiples

- **Control exacto**. Si se controla la tasa de error cuando las hipótesis nulas están dadas por el subconjunto $\Delta_0 \subseteq \{1, \ldots, m\}$ lo llamaremos control exacto.

Por ejemplo para el "familywise error rate" tenemos que es la probabilidad de rechazar (en forma equivocada) por lo menos una de las hipótesis entre todas las hipótesis nulas verdaderas:

FWER = P ($V \geq 1 \mid \cap_{j \in \Delta 0} H_{0j}$), donde "$\cap_{j \in \Delta 0} H_{0j}$" se refiere al subconjunto de hipótesis nulas verdaderas. Como Δ_0 en general es desconocido terminamos controlando esa tasa de error bajo la hipótesis nula completa - control débil -

- **Control débil**. Se controla la tasa del error de tipo I suponiendo que todas las hipotesis nulas son ciertas. Es decir bajo una hipótesis nula completa

$$H_0^C = \bigcup_{j=1}^{m} H_{0j}$$

para la cual mo=m. No habría genes expresados diferencialemente. En general la hipótesis nula completa no es realista y el control débil no es satisfactorio.
En realidad algunas hipótesis nulas pueden ser verdaderas y otras falsas pero el subconjunto Δ_0 es desconocido.

- **Control fuerte**. Se controla el error tipo I bajo cualquier combinacion de hipotesis nulas y falsas. Algunos genes estarán diferencialmente expresados y otros no.

Observe que los conceptos de control débil y fuerte se aplican a cada una de las tasas de errores definidas arriba PCER, PFER, FWER y FDR.

En general, control fuerte implica control exacto y control débil. En el contexto de los microarrays es muy raro que ninguno de los genes estén expresados diferencialmente. De manera que el control débil sólo es no satisfactorio y es importante tener control exacto o fuerte de las tasas de error de tipo I. La ventaja del control exacto es una mayor potencia.

10.4 Potencia

Dentro de la clase de los procedimientos de tests múltiples que controlan la tasa del error de tipo I a un nivel α, se busca un procedimiento que maximice la *potencia* esto es minimizar la tasa de un error de tipo II bien definido. Tal como ocurre con las tasas de error de tipo I (Type I error rates), el concepto de potencia puede ser generalizado en varias maneras al pasar de una hipótesis a hipótesis múltiples.

Recordemos que S mide la cantidad de rechazos entre todas las hipótesis nulas falsas, o sea:

$$S = \sum_{j=1}^{m_1} S_j, \quad S_j = \begin{cases} 1 & \text{rechazo} \quad H_{0,j} \text{ con } H_{0,j} \text{ falsa} \\ 0 & \text{no rechazo} \quad H_{0,j} \text{ con } H_{0,j} \text{ falsa} \end{cases}, \; p_{Sj} = P(S_j = 1, H_{0,j} \text{ falsa})$$

Tres son las generalizaciones de la definicón de potencia habituales:

(1) Probabilidad de rechazar por lo menos una hipótesis nula falsa

$$\mathrm{P}(\, S \geq 1 \,) = \mathrm{P}\,(\, T \leq m_1 - 1 \,)$$

(2) $E(S)/m1$, *potencia media.*

$$E(S)/m1 = \sum_{j=1}^{m_1} p_{S_j} / m_1$$

(3) Probabilidad de rechazar todas las hipótesis nulas falsas (Shaffer, 1995)

$$\mathrm{P}(\, S = m_1 \,) = \mathrm{P}\,(\, T = 0 \,)$$

Cuando la familia de tests consisten en comparaciones de pares de medias estas cantidades han sido llamadas: potencia para cualquier par, potencia por cada par y potencia para todos los pares respectivamente (Ramsey, 1978).

Con un espírutu análogo al del FDR, se podría definir potencia como el valor esperado de la proporción de aciertos:

$E(S/R \mid R > 0)\, P(R > 0) = P(R > 0) - FDR$

Cuando todas las hipótesis nulas son falsas, $m = m1$, esto resulta la potencia de cualquier par: $P(S \geq 1)$.

Destacamos nuevamente que todas las probabilidades dependen del subconjunto particular $\Delta_0 \subseteq \{1, \ldots, m\}$ de hipótesis nulas que son ciertas que por supuesto es desconocido ($m0 = \#\Delta_0$ y $m1 = m - \#\Delta_0$)

10.5 *Comparación de tasas de Error de Tipo I.*

Dado el mismo procedimiento de testeo múltiple, basado en el rechazo de m estadísticos ($T_1, T_2, \ldots, T_m$) sobre la misma región m- dimensional vale que las tasas de Error de Tipo I satisfacen las siguientes ecuaciones

$$PCER \leq FDR \leq FWER \leq PFER$$
$$FDR \leq pFDR$$

Por lo tanto para un valor fijo α que acota las tasas de Error de Tipo 1 la cantidad de rechazos R tendrá el orden inverso.

Demostración: Es fácil ver que $0 \leq V \leq R \leq m$ y que $R = 0 \Rightarrow V = 0$

Por lo tanto $\frac{V}{m} \leq \frac{V}{R} 1_{\{R>0\}} \leq 1_{\{V>0\}} \leq V$

La afirmación resulta de tomar esperanzas.

Es más dificil describir la relación entre el pFDR y FWER. En las aplicaciones de microarrays se espera que $pFDR \leq FWER$ salvo para el caso en que $m_0 = m$ y que $1 = pFDR \geq FDR = FWER$. Es poco probable que esto ocurra en experimentos de microarrays en los que se espera que por lo menos uno de los genes esté expresado diferencialmente. También $P(R > 0) \to 1$ cuando $m \to \infty$, en cuyo caso pFDR = FDR. Por lo tanto se espera que en general se satisfagan

$$PCER \leq FDR \leq pFDR \leq FWER \leq PFER$$

Observación: Si se utiliza $PFER \leq \alpha$ y se rechaza con este criterio seguro también se rechazará con FWER y PCER. Decimos entonces que PFER es un critero generalmente más *conservativo*, esto significa que produce menos rechazos que cuando se controla el FWER o el PCER. A su vez los procedimientos que controlan el FWER son más conservativos que aquellos que controlan el PCER.

10.5.1 Ejemplo ilustrativo general de las diferentes tasas de Error de Tipo I

Supongamos que en un procedimiento de tests múltiple cada hipótesis H_{0j} se testea individualmente al nivel α_j y que la decisión de rechazar o no rechazar esa hipótesis está basada únicamente en ese test.

Bajo la hipótesis nula completa PCER es simplemente el promedio de los α_j y el PFER es la suma de los α_j .

En cambio el FWER es una función no solo de los α_j, sino que interviene la distribución *conjunta* de los estadísticos *Tj*:

$$PCER = \frac{\alpha_1 + \cdots + \alpha_m}{m} \leq \max(\alpha_1, \cdots, \alpha_m)$$
$$\leq FWER \leq PFER = \alpha_1 + \cdots + \alpha_m$$

El FDR también depende de la distribución conjunta de los estadísticos y para un procedimiento fijo FDR **≤** FWER, con FDR **=** FWER bajo la nula completa (Benjamini and Hochberg, 1995).

El enfoque clásico de los procedimientos de tests múltiples utiliza el control fuerte del FWER (por ejemplo el procedimiento de Bonferroni). Un procedimiento más reciente de Benjamini and Hochberg (1995) controla el FWER en el sentido débil en cuyo caso FDR **=** FWER y puede ser menos conservativo (rechaza más) que el FWER. Los procedimientos que controlan el PCER son generalmente menos conservativos que aquellos que controlan el FDR o el FWER, pero tienden a ignorar el problema de la multiplicidad.

10.5.2 Ejemplo simple:

Utilizaremos un ejemplo simple para describir el comportamiento de los distintas tasas de error de tipo I cuando la cantidad de hipótesis *m* (una para cada gen) y la proporción de hipótesis nulas verdaderas m_0/m varían.

Consideremos un vector aleatorio de longitud *m*, tal que cada componente X_j , $1 \leq j \leq m$,
$X_j \sim N(\mu_j, 1)$ independientes. Por lo tanto el vector tiene media $\mu = (\mu_1, \ldots, \mu_m)$ y matriz de covarianza I_m

Supogamos que queremos testear simultaneamente las *m* hipótesis nulas $H0j : \mu_j = 0$ contra las alternativas bilaterales $H'_j : |\mu_j| \neq 0$

Dada una muestra aleatoria de *n* vectores de longitud *m,* un procedimiento simple de testeo múltiple podría rechazar H_0j si $|\overline{X}_j| \geq z_{\alpha/2}/\sqrt{n}$, donde $\overline{X}_j = \sum_{i=1}^{n} X_j^{(i)}/n$ es el promedio de la coordenada *j* de los *n* vectores de long. *m*, donde $z_{\alpha/2}$ satisface $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ siendo $\Phi(.)$ la función de distribución acumulada de la Normal

estándar. Sea $R_j = I(|\overline{X}_j| \geq z_{\alpha/2}/\sqrt{n})$ donde $I(.)$ es la función indicadora que vale 1 cuando la condición entre paréntesis es verdadera y cero si es falsa. Supongamos sin pérdida de generalidad que las m_0 hipótesis nulas verdaderas son las primeras $H_{01}, \ldots,$ H_{0m0}, esto es $\Delta_0 = \{1, \ldots, m_0\}$. Entonces $V = \sum_{j=1}^{m_0} R_j$ y $R = \sum_{j=1}^{m} R_j$.

Sea $\gamma_j = E(R_j) = P(R_j = 1) = 1 - \Phi(z_{\alpha/2} - \mu_j\sqrt{n}) + \Phi(-z_{\alpha/2} - \mu_j\sqrt{n})$, o sea que γ_j es la probabilidad de rechazar H_{0j} cuando el verdadero valor de la media es μ_j. Es fácil ver que las tasas de Error de Tipo I en este caso son:

$$PFER = \sum_{j=1}^{m_0} \gamma_j,\ PCER = \sum_{j=1}^{m_0} \gamma_j / m,\ FWER = 1 - \prod_{J=1}^{m_0} (1-\gamma_j) \quad \text{y}$$

$$FDR = \sum_{r_1=0}^{1} \cdots \sum_{r_m=0}^{1} \frac{\sum_{j=1}^{m_0} r_j}{\sum_{j=1}^{m} r_j} \prod_{j=1}^{m} \gamma_j^{r_j} (1-\gamma_j)^{1-\gamma_j}$$

Las 3 primeras expresiones resultan inmediatamente de las definiciones de las tasas de error y las propiedades de la esperanza. Veamos la demostración de la expresión que tenemos para el FDR.

Dem: recordemos que FDR = E(Q), siendo Q = V/R si R > 0 y 0 si R=0. Además en este caso $V = \sum_{j=1}^{m_0} R_j$ y $R = \sum_{j=1}^{m} R_j$. Como Q = g (R_1,..., R_m), de la definición de esperanza, por la independencia de los R_j y la convención de FDR que 0/0 es 0, resulta

$$E(Q) = \sum_{r_1=0}^{1} \cdots \sum_{r_m=0}^{1} \frac{\sum_{j=1}^{m_0} r_j}{\sum_{j=1}^{m} r_j} P(R_1 = r_1, \cdots, R_m = r_m) =$$

$$\sum_{r_1=0}^{1} \cdots \sum_{r_m=0}^{1} \frac{\sum_{j=1}^{m_0} r_j}{\sum_{j=1}^{m} r_j} \prod_{j=1}^{m} P(R_j = r_j) =$$

$$\sum_{r_1=0}^{1} \cdots \sum_{r_m=0}^{1} \frac{\sum_{j=1}^{m_0} r_j}{\sum_{j=1}^{m} r_j} \prod_{j=1}^{m} \gamma_j^{r_j} (1-\gamma_j)^{1-\gamma_j}$$

Además en este ejemplo $\gamma_j = \alpha$ para $j = 1, \ldots, m_0$. Si además suponemos que para $j = m_0 + 1, \ldots, m$ $\mu_j = d/\sqrt{n}$ las expresiones de las tasas de error se simplifican a

$$PFER = m_0 \alpha,\ PCER = m_0 \alpha / m,\ FWER = 1 - (1-\alpha)^{m_0}$$

$$FDR = \sum_{s=0}^{m_1} \sum_{v=0}^{m_0} \frac{v}{v+s} \binom{m_0}{v} \alpha^v (1-\alpha)^{m_0-v} \binom{m_1}{s} \beta^s (1-\beta)^{m_1-s}$$

donde

$$\beta = 1 - \Phi(z_{\alpha/2} - d) + \Phi(-z_{\alpha/2} - d)$$

Teniendo en cuenta que pFDR = FDR / P(R>0) resulta en este caso que

$$pFDR = \frac{\sum_{s=0}^{m_1} \sum_{v=0}^{m_0} \frac{v}{v+s} \binom{m_0}{v} \alpha^v (1-\alpha)^{m_0-v} \binom{m_1}{s} \beta^s (1-\beta)^{m_1-s}}{\sum_{k=1}^{m} \sum_{v=0}^{\min(m_0,k)} \binom{m_0}{v} \alpha^v (1-\alpha)^{m_0-v} \binom{m_1}{k-v} \beta^{k-v} (1-\beta)^{m_1-(k-v)}}$$

Observación

A diferencia del PCER, PFER y FWER, el FDR y el pFDR dependen de la distribución del estadístico bajo las hipótesis alternativas H'_j, para $j = m_0 + 1, \ldots, m$, a través de la variable aleatoria S. Además únicamente para el cálculo de FDR y pFDR hemos utilizado el supuesto de independencia entre los estadísticos de los tests. En el ejemplo el FDR y pFDR son funciones de β, la probabilidad de rechazo bajo la hipótesis alternativa.
En general, FDR y pFDR son mucho más dificultosos que las otras tres tasas de error.

10.6 *p*-valores

p-valores sin ajustar .
Consideremos primero un test para una única hipótesis H_{01} basado en el estadístico T_1. Si el valor observado del estadístico es t_1, el correspondiente p-valor es

$$p_1 = P(|T_1| \geq |t_1| \mid H_{01})$$

o sea es la probabilidad de observar un valor del estadístico del test tan o más extremo que el observado en la dirección de rechazo cuando la hipótesis nula es verdadera.

Cuanto menor sea el *p*-valor $p1$, mayor es la evidencia en contra de la hipótesis nula H_{01}. Rechazar H_{01} cuando $p1 \leq \alpha$ provee un control de la probabilidad de error de tipo I a nivel α.

La extensión del concepto de *p*-valor al contexto de tests múltiples lleva a la muy útil definición de *p*-valor ajustado.

p-valores ajustados.

El problema aquí es cómo deben modificarse los niveles de los tests individualmente para obtener un nivel global, dado por alguna de las tasas de error dadas anteriormente,

deseado. Y equivalentemente cómo se podifican los p-valores individuales para cada test.

Sean t_j y $p_j = P(|T_j| \geq |t_j| \mid H_{0j})$ el valor del estadístico del test y el p-valor *no ajustado*, respectivamente para la hipótesis H_{0j} (gen j), $j = 1, \ldots, m$. De igual manera que para una única hipótesis un procedimiento múltiple puede definirse en términos de los valores críticos ó los p-valores de las hipótesis individuales: por ejemplo rechace H_{0j} si $|t_j| \geq c_j$ o si $p_j \leq \alpha_j$, donde los valores críticos c_j and α_j son elegidos para controlar alguna tasa de error (FWER, PCER, PFER o FDR) a un nivel α preespecificado.

Dado cualquier procedimiento de tests múltiples, el *p-valor ajustado* que corresponde al test para la hipótesis individual H_{0j} puede ser definido como el nivel nominal para el cual el procedimiento completo resultaría justo en rechazo dados todos los valores de los estadísticos observados. (Hommel and Bernhard, 1999; Shaffer, 1995; Westfall and Young, 1993; Wright, 1992; Yekutieli and Benjamini, 1999).

Si interesa controlar el FWER, el p-valor ajustado para la hipótesis H_{0j} , dado un nivel especificado nominal α del procedimiento de tests múltiples éste será

$$\tilde{p}_j = \inf\{\alpha : H_{0j} \text{ es rechazada al } FWER = \alpha\}$$

Observación: tanto los p-valores como los p-valores ajustados son variables aleatorias.

Las variables aleatorias correspondientes a p-valores ajustados y sin ajustar serán indicados por $\tilde{P}_j$ y P_j respectivamente.

La hipótesis H_{0j} es rechazada, esto es el gen j es declarado diferencialmente expresado a nivel nominal FWER α si $\tilde{p}_j \leq \alpha$

Para muchos procedimientos, tal como el de Bonferroni que veremos en la próxima sección, el nivel *nominal* es generalmente mayor que el nivel *real*, resultando así un procedimiento conservativo.

Los p-valores ajustados para procedimientos que controlan otros tipos de errores se definen en forma similar. Por ejemplo, para procedimientos que controlan el FDR

$$\hat{p}_j = \inf\{\alpha \in [0,1] : H_{0j} \text{al nivel nominal FDR} = \alpha\}$$

Tal como ocurre en el caso de una única hipótesis, la ventaja que tiene reportar p-valores ajustados en vez de simplemente rechazar o no rechazar la hipótesis, es que no es necesario especificar el nivel del test previamente.

Ya veremos que algunos procedimientos de testeo múltiple se describen mejor en base a los p-valores ajustados y que estos a su vez pueden estimarse por métodos de remuestreo (resampling methods, Westfall and Young, 1993).

10.7 Procedimientos que controlan el FWER-family-wise error rate

Tres son las clases de procedimientos que son utilizados habitualmente:

- *Un paso*, single-step: se realizan ajustes por multiplicidad a todas las hipótesis por igual, sin depender de el orden de los valores de los estadísticos ni de los p-valores originales.

Procedimientos en pasos

Se pueden lograr mejoras en la potencia, manteniendo el control de la tasa de error de tipo I por procedimientos en pasos en los que el rechazo de una hipótesis en particular está basado en la cantidad total de hipótesis a testear sino también en los resultados de los tests para las demás hipótesis.

- *Pasos hacia abajo*, step-down: se ordenan los p-valores (o los estadísticos de los tests) comenzando por los más significativos.
- *Pasos hacia arriba*, step-up: se ordenan los p-valores (o los estadísticos de los tests) comenzando por los menos significativos.

En esta clase veremos unicamente

10.7.1 Procedimientos de un paso

Método de Bonferroni

El método de Bonferroni es tal vez uno de los procedimientos más conocidos de testeo múltiple. Propone rechazar cada una de las hipótesis nulas H_{0j} , j=1,...,m con un nivel individual $\alpha^* = \alpha/m$ para obtener un control fuerte del FWER a nivel α. Esto es equivalente a rechazar H_{0j} si $p_j \le \alpha/m \Rightarrow mp_j \le \alpha$, llamando p-valor ajustado $\tilde{p}_j = \text{mín}\{mp_j, 1\}$

Dem:
El control del FWER en el sentido fuerte depende del conjunto de hipótesis nulas verdaderas es decir depende de $\Delta_0 = \{j : H_{0j} \text{ es verdadera}\}$ con $\#\Delta_0 = m_0$

$$FWER = P(V > 0) = P(\text{rechazar alguna } H_{oj} \text{ verdadera}) = P\left[\bigcup_{j\in\Delta_0}\{P_j \le \alpha^*\}\right] \le$$

$$\sum_{j\in\Delta_0} P\{P_j \le \alpha^*\} = \sum_{j\in\Delta_0} \alpha^* = m_0\alpha^* \le \alpha; \text{ alcanza con tomar } \alpha^* = \alpha/m$$

equivalentemente

$$FWER = P(V > 0) = P(\text{rechazar alguna } H_{oj} \text{ verdadera}) = P\left[\bigcup_{j\in\Delta_0}\{\tilde{P}_j \le \alpha\}\right] \le$$

$$\sum_{j\in\Delta_0} P\{P_j \le \alpha/m\} = \sum_{j\in\Delta_0} \alpha/m = m_0\alpha/m \le \alpha$$

Corolario. La corrección propuesta por el método de Bonferroni tambíen controla el PFER

Dem: El control del *PFER* en el sentido fuerte depende del conjunto de hipótesis nulas verdaderas es decir depende de $\Delta_0 = \{j : H_{0j} \text{ es verdadera}\}$ con $\#\Delta_0 = m_0$, por lo tanto si consideramos que todos los tests se realizan al mismo nivel α tenemos que $PFER = \sum_{j \in \Delta_0} \alpha^* = m_0 \alpha^* \leq \alpha$ si $\alpha^* = \alpha / m$

Observaciones, para Bonferroni.

Se obtiene un FWER de a lo sumo α. Los p-valores ajustados por Bonferroni, no son p-valores ajustados en sentido estricto. En cambio son cotas inferiores conservativas de los p-valores ajustados que son muy difíciles o imposibles de calcular sin realizar más supuestos.

- Para un gen (j) el p-valor ($t^{j)}_{observado}$) = $p^{(j)}$, entonces

p- valor ajustado para el gen j
$p_A^{(j)} = \text{mín}(m^* p^{(j)}, 1)$

- Todos los p-valores se multiplican por el mismo número m, para ajustarse.
- Se eligen como estadísticamente DE aquellos genes para los que $p^{(j)} \leq \alpha / m$.

Esta propuesta puede ser *demasiado conservativa* y cuando la cantidad de tests es grande los niveles corregidos resultan demasiado bajos, equivalentemente los p-valores demasiado altos. Esto significa que se *seleccionarán pocos genes* como candidatos a estar DE.

El método de Bonferroni garantiza que la probabilidad de rechazar como mínimo una hpótesis nula verdadera sea menor o igual a α para una distribución arbitraria de *p*-valores. Sin embargo la cota al PFER puede ser mayor a 1 ya que nos da la cantidad esperada de falsos positivos. Tomando un α más grande (que no tiene sentido que sea mayor a 1 al controlar el FWER que es una probabilidad) tendremos una cota menos restrictiva que dará un procedimiento con mayor potencia.

Este es el enfoque adoptado por Alexander Gordon, Galina Glazko, Xing Qiu, Andrei Yakovlev (2007). Proponen utilizar el método de Bon Ferroni pero controlando el valor esperado de falsos positivos PFER = E(V). El procedimiento de Bonferroni a "nivel α" controla el PFER (Lee 2004) y como FWER ≤ PFER, también controla el FWER. Pero si interesa controlar el PWER α este valor puede ser mayor a 1. Muestran que este procedimiento es más estable que controlar el FDR.

Método de Sidák

El método de Sidák provee un control exacto del FWER bajo la hipótesis nula completa cuando los p-valores sin ajustar son independientes y están uniformemente distribuidos en el intervalo [0,1]. También provee un control fuerte del FWER para cualquier combinación de hipótesis nulas verdaderas. Los p-valores ajustados están dados por

$$\widetilde{p}_j = 1-(1-p_j)^m$$

Dem:

$$P(V=0) = P(\text{no rechazar todas las } H_{oj} \text{ verdaderas}) = P\left[\bigcap_{j\in\Delta_0}\left\{\widetilde{P}_j > \alpha\right\}\right] \text{ por indep. tenemos}$$

$$= \prod_{j\in\Delta_0} P\left\{\widetilde{P}_j > \alpha\right\} = \prod_{j\in\Delta_0} P\left\{(1-(1-P_j)^m > \alpha\right\} = \prod_{j\in\Delta_0} P\left\{P_j > 1-(1-\alpha)^{1/m}\right\} \text{ por Unif. resulta}$$

$$= \prod_{j\in\Delta_0} P\left\{P_j \le (1-\alpha)^{1/m}\right\} = (1-\alpha)^{m_0/m} > 1-\alpha$$

Por lo tanto

$$FWER = P(V>0) = 1-P(V=0) = 1-(1-\alpha)^{m_0/m} \le \alpha$$

Muchas veces los estadísticos de los tests están correlacionados. Esto ocurre en el caso de los experimentos de microarreglos en los cuales grupos de genes tienden a tener altas correlaciones debido a la corregulación.

Una buena descripción de los procedimientos de tests múltiples de uno y varios pasos se encuentra en Dudoit et al 2003.

Referencias

Alexander Gordon, Galina Glazko, Xing Qiu, Andrei Yakovlev (2007), "Control of the mean number of false discoveries, Bonferroni and stability of multiple testing", Annals of Applied Statistics, 1(1):179-190.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society B, 57, 289 -300.

Benjamini, Y and Yekutieli, D. (2001) The Control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29(4):1165-1188, 2001. Preprint verison available at http://www.math.tau.ac.il/~ybenja/ under "Papers"

GE, Y., DUDOIT, S., SPEED, T. (2003). Resampling-based Multiple Testing for Microarray Data Analysis. *Sociedad de Estadística e Investigación Operativa. Test* **12**,1, 1–77.

Multiple Hypothesis Testing in Microarray Experiments
Sandrine DUDOIT, Juliet Popper SHAFFER and Jennifer C. BOLDRICK
Statistical Science 2003, **18**, 1, 71–103 Institute of Mathematical Statistic

LEE, M.-L. (2004). *Analysis of Microarray Gene Expression Data*. Kluwer, Boston.

RAMSEY, P. H. (1978). Power differences between pairwise multiple comparisons. *J. Amer. Statist. Assoc.* **73** 479–485.

STOREY, J. D. (2002a). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**:479–498.

SHAFFER, J. P. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology* **46** 561–584. Bajado

SHAFFER, J. P. (1986).Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81** 826–831.