Multiple testing: FDR + Bayesian perspective

Mark van de Wiel
mark.vdwiel@vumc.nl

Department of Epidemiology & Biostatistics
Department of Mathematics
VUmc/VU, Amsterdam

vrije Universiteit

VU medisch centrum

Content

Simultaneous testing

How does multiple testing differ from single testing?

1. p-values not sufficient to control false positive rate

   $p \leq 0.05$ implies too many false positives

2. Different error control desirable when # tests large:

   control proportion of false positives instead of FWER= P(at least 1 false positive)

3. Opportunity to learn from other features

   When using a t-test: estimate the s.d. from all features (other lecture: limma)

Bonferroni

Bonferroni's solution

1. Multiply p-values with m, the number of tests
2. Reject null-hypothesis for feature i when

$$p_i^{bonf} = m \times p_i \leq \alpha \text{ (e.g. } \alpha = 0.05)$$

Probability on one or more false rejections, FWER, is smaller or equal to α.

Bonferroni

- m = 5
- α = 0.05.

Gene	p-values p_i	Bonferroni p_i^{Bonf}	Reject H_0
1	0.006	$p_i \times m = 0.03$	Yes
2	0.372	$p_i \times m > 1$	No
3	0.012	$p_i \times m = 0.06$	No
4	0.000	$p_i \times m = 0.00$	Yes
5	0.811	$p_i \times m > 1$	No

Gene 1 and 4 are declared differentially expressed.

Bonferroni

- Now suppose m=1000
- Consider the same 5 genes
- $\alpha = 0.05$.

Gene	p-values	Bonferroni rejection level	Reject H_0
1	0.006	$p_i \times m > 1$	No
2	0.372	$p_i \times m > 1$	No
3	0.012	$p_i \times m > 1$	No
4	0.000	$p_i \times m = 0.00$	Yes
5	0.811	$p_i \times m > 1$	No

Only gene 4 is declared differentially expressed.

FDR

Bonferroni's rule:

- Does not scale with the number of tests m
- Is too conservative for very high-dimensional data (e.g. m=100,000)

Alternative: False Discovery Rate (FDR)

FDR: expected proportion of false discoveries among all discoveries given a p-value threshold.

$$FDR = E(V/R) \approx E(V)/R = E(\#FD) / R.$$

FDR and False Discovery Proportion (FDP)

	True Null	False Null	
Rejected	V	U	R
Non-rejected	m_0 - V	m_1 - U	m-R
	m_0	m_1	m

$$FDR = E(FDP) = E(V/R) \approx E(V) /R$$

FDR and FDP

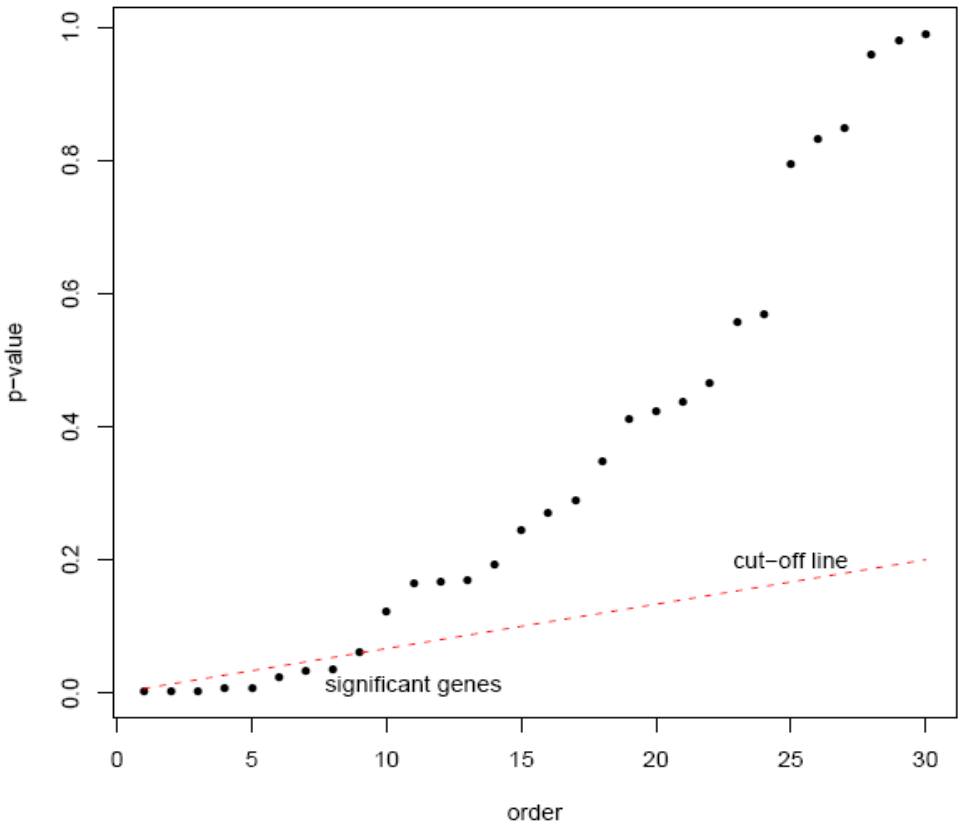
FDR control vs FDP estimation

- FDR control: guarantee that $E(V/R) \leq \alpha$
- FDP estimation: $\hat{FDP} \leq \alpha$
- FDR control is more strict than FDP estimation

FDR, Benjamini-Hochberg

Control [hand] of FDR, Benjamini-Hochberg's step-up [tut] procedure

- Select a desired limit α of the FDR.
- Order the p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
- Find largest rank r for which: $m \times p_{(r)}/r \leq \alpha$.
- All genes with a lower or equal rank are declared significant.



The cut-off line is given by $y = \alpha \times r / m$.

FDR

Benjamini-Hochberg: adjusted p-values

- Order the *p*-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
- $p'_{(r)} = m \times p_{(r)}/r$
- $p_{(r)}^{BH} = \min(p'_{(r)}, p'_{(r+1)})$

m=100, five genes with lowest p-values

$p_{(r)}$	$p'_{(r)}$	$p_{(r)}^{BH}$
0.00010	0.010	0.005
0.00011	0.005	0.005
0.0052	0.173	0.173
0.024	0.600	0.600
0.066	>1	>1

Last step: guarantees that adjusted p-values are in the same order as the raw ones

FDR

Conditions for Benjamini-Hochberg (BH)

- Proven FDR-control for independent p-values [hand] or specific type of correlation (PDSS [tut])

- Simulations, however, show that the BH procedure is very robust against many types of correlations

- Alternative when one suspects very strong correlations:

Replace m/r by: $(m/r)\sum_{j=1}^{m} 1/j \approx (m/r)\log(m)$

- Known as Benjamini-Yekutieli rule

R

Benjamini-Hochberg and Benjamini-Yekutieli in R

```
load("C:\\Synchr\\Onderwijs\\HighDimensional\\Slides\\FDR\\Exercises\\pvals.Rdata")

pvaladjBH <- p.adjust(pvals,method="BH")
pBH50 <- sort(pvaladjBH)[1:50]

pvaladjBY <- p.adjust(pvals,method="BY")
pBY50 <- sort(pvaladjBY)[1:50]

pvaladjBonf <- p.adjust(pvals,method="bonferroni")
pBonf50 <- sort(pvaladjBonf)[1:50]

cbind(pBH50, pBY50,pBonf50)
```

FDP/FDR estimation

- Estimation of FDP = V/R. Literature not consistent about FDP/FDR terminology.

- Given p-value threshold t:

$$\hat{FDP}(t) = \hat{V}(t) / \hat{R}(t) = \sum_{i=1}^{m} P(p_i \le t \mid H_{0i} = \text{true})\, \hat{P}(H_{0i} = \text{true}) \Big/ \#(p_i \le t)$$

$$= \hat{\pi}_0 mt / \#(p_i \le t)$$

- Methods differ in estimation of the proportion of non-rejections: π_0 [hand,tut]

Equivalence Theorem [hand]

When using $\pi_0 = 1$, the BH rule is equivalent to rejecting all H_{0i} for which $p_i \leq t$, with

$$t = \max_u(\hat{FDP}(u) \leq \alpha)$$

- Equivalence Theorem is nice, but crux of FDP estimators is estimation of π_0
- No guarantee for control of FDR, but generally more powerful than BH

Bayesian interpretation

- Model distribution of p-values by $F(t) = \pi_0 F_0(t) + (1- \pi_0)F_1(t)$
- bFDR = Posterior probability *given p-value* p_i:

$$\begin{aligned} bFDR &= P(H_{0i} = true \mid p_i \leq t) \\ &= P(p_i \leq t \mid H_{0i} = true)P(H_{0i} = true)/P(p_i \leq t) \\ &= \pi_0 F_0(t)/F(t) \\ &\approx \hat{\pi}_0 mt/\#(p_i \leq t) \end{aligned}$$

- Local FDR (lfdr) concept [tut,hand]:

$$lfdr = P(H_{0i} = true \mid p_i = t) = \pi_0 f_0(t)/f(t)$$

Bayesian interpretation

Relationship between lfdr and bFDR [hand]

$$E_f\left[\text{lfdr}(U) \mid U \leq v\right] = \text{bFDR}(v)$$

So, at threshold v, bFDR is the mean lfdr over all thresholds $U \leq v$

Why bFDR and lfdr are not really “Bayesian”?

$$bFDR = P(H_{0i} = true \mid p_i \leq t)$$

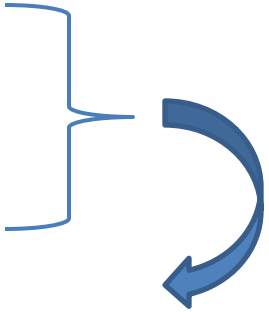
$$lfdr = P(H_{0i} = true \mid p_i = t)$$

These quantities still rely on p-values!

True Bayesian analogues of lfdr and bFDR: BFDR

Recap: posterior probabilities

- Simple regression (or GLM-type) model: $Y_j = \alpha + \beta X_j + \varepsilon_j$
- Prior on β: $\pi(\beta)$
- Likelihood: $\pi(\mathbf{Y}|\beta)$



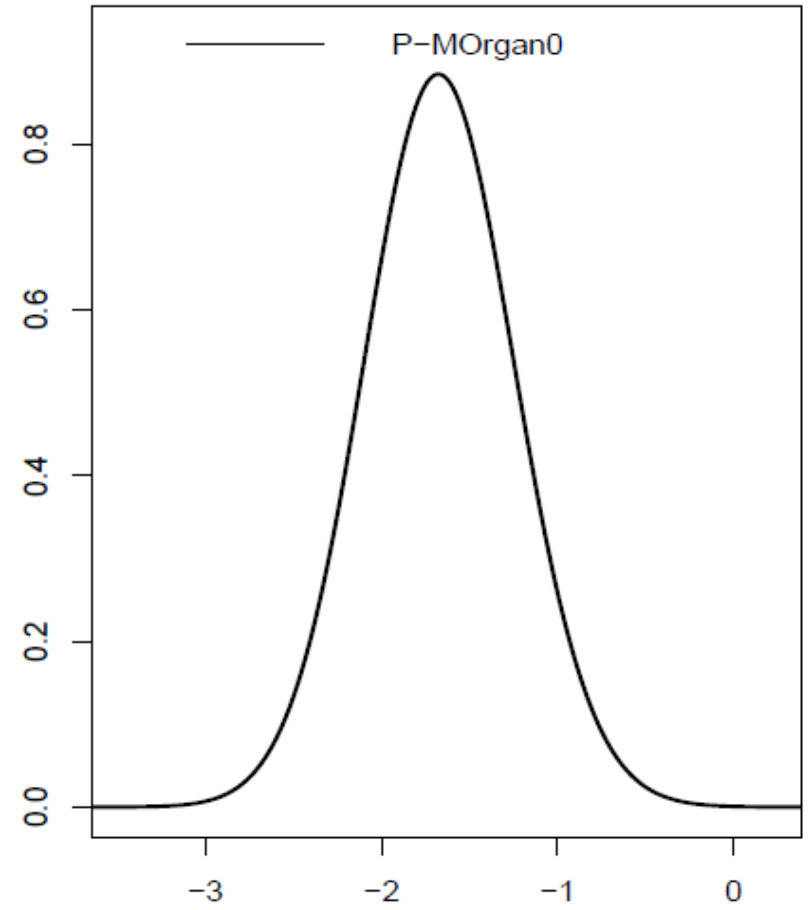
Bayes' rule

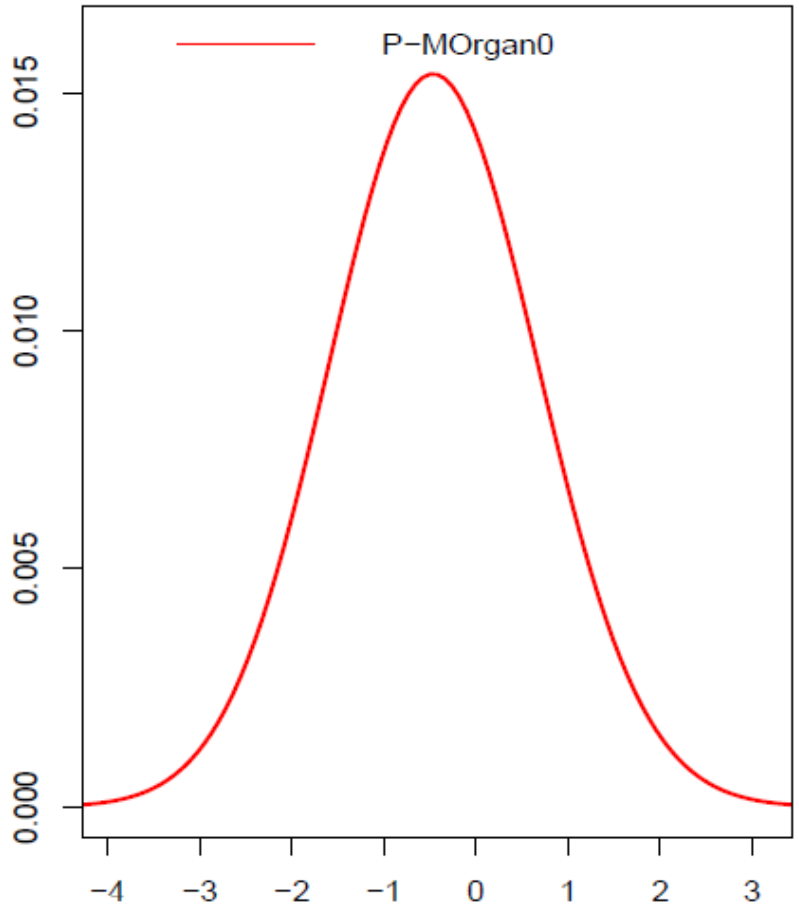
Posterior: $$\pi(\beta|\mathbf{Y}) = \frac{\pi(\beta)\pi(\mathbf{Y}|\beta)}{\pi(\mathbf{Y})} = \frac{\pi(\beta)\pi(\mathbf{Y}|\beta)}{\int \pi(\beta)\pi(\mathbf{Y}|\beta)d\beta}$$

Bayesian inference, single test

Parameter of interest: gene expression difference between two conditions: primary and metastasis. Model:

$$Y_j = \alpha + \beta_{P\text{-}M} I_{\{j \in P\}} + \varepsilon_j$$





Bayesian FDR: BFDR

Hypothesis testing

Null-hypothesis H_0: $|\beta| \leq \delta$

Posterior null-probability: $\pi_0(\mathbf{Y}) = P(H_0 \mid \mathbf{Y}) = \int_{-\delta}^{\delta} \pi(\beta \mid \mathbf{Y}) d\beta$

Recall lfdr: $\text{lfdr} = P(H_0 \mid p = t) = P(H_0 \mid p(\mathbf{Y}) = t)$

$\pi_0(\mathbf{Y})$ is a Bayesian version of lfdr

Bayesian FDR: BFDR

Point null-hypothesis in a Bayesian setting

What if one aims to test H_0: $\beta=0$?

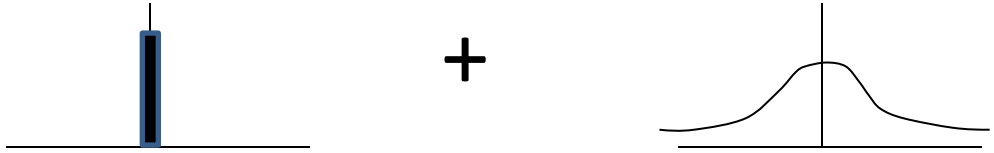
$$\pi_0(\mathbf{Y}) = P(H_0 \mid \mathbf{Y}) = \pi(0 \mid \mathbf{Y}) \quad = 0????$$

Prior $\pi(\beta)$ needs to have mass on 0 to avoid this

Bayesian FDR: BFDR

Point null-hypothesis in a Bayesian setting

If: $\pi(\beta) = p_0\delta(0) + (1 - p_0)\pi'(\beta)$



Then[Exer]:

$$\pi_0(\mathbf{Y}) = \frac{p_0\pi(\mathbf{Y}\,|\,0)}{p_0\pi(\mathbf{Y}\,|\,0) + (1 - p_0)\Pi(\mathbf{Y})} = \frac{p_0\pi(\mathbf{Y}\,|\,0)}{p_0\pi(\mathbf{Y}\,|\,0) + (1 - p_0)\int \pi'(\beta)\pi(\mathbf{Y}\,|\,\beta)d\beta}$$

$\pi(\mathbf{Y}|0)$ and $\Pi(\mathbf{Y})$ are called *marginal likelihoods.*

Bayesian FDR: BFDR

BFDR

$$BFDR(t) = E\left[\pi_0(\mathbf{Y}) \mid \pi_0(\mathbf{Y}) \le t\right] \hat{=} \frac{\sum_{i=1}^{p} \pi_0(\mathbf{Y}) I_{\{\pi_0(\mathbf{Y}) \le t\}}}{\sum_{i=1}^{p} I_{\{\pi_0(\mathbf{Y}) \le t\}}}$$

pi0(Y) = lfdr = t	sum	div	BFDR
0.001	0.001	1	0.001
0.003	0.004	2	0.002
0.008	0.012	3	0.004
0.012	0.024	4	0.006

Bayesian FDR: BFDR

BFDR vs bFDR

$$bFDR(u) = E_f\left[lfdr(t) \mid t \le u\right]$$

$$BFDR(u) = E\left[\pi_0(\mathbf{Y}) \mid \pi_0(\mathbf{Y}) \le u\right]$$

Bayesian FDR: BFDR

Some discussion

- Bayesian FDR is only an estimate; Bayesians general are less concerned with error control

- Prior very important for performance (see ShrinkBayes lectures)

- Bayesian FDR and (frequentistic) FDR are different entities BUT can be interpreted similary for the purpose of estimation

- Very useful in practice: Biologists, medical researchers etc. do not have intuition for posterior probabilities, but do for (B)FDR

- The more true null features are separated from non-null features the more similar BFDR and FDR are.

FDR: Pitfalls

Subsetting property: for each subset S of all hypotheses, error control on the entire set of hypotheses implies error control for S

FWER has the subsetting property (trivial)

FDR does generally not have the subsetting property

Very important to *a priori* fix the relevant set of hypotheses to be tested

FDR: Pitfalls

m = 1000 genes on the array, α = 0.05.

Case 1

gene	p-values	rank	p^{BH}	Reject H_0
1	0.006	10	0.6	No
3	0.012	15	1	No
4	0.000	1	0.000	Yes ←

Case 2

gene	p-values	rank	p^{BH}	Reject H_0
1	0.006	200	0.03	Yes ←
3	0.012	300	0.045	Yes ←
4	0.000	1	0.000	Yes ←

Significance of gene 1 & 3 depends also on p-values of the *other* genes (the more small p-values, the better…)

FDR: Pitfalls

Setting: researcher knows that apoptosis-related genes are likely to be differentially expressed given two conditions
Interest: which *other* genes are differentially expressed?

Example: 500 apoptosis genes, m_1 =10.000 genes, m_2 = 10.000-500 = 9.500 non-apoptosis genes

p-value	Apoptosis	p_1^{BH}	p_2^{BH}
$2.3*10^{-6}$	yes	$p*m_1/1=0.023$	
$5.2*10^{-6}$	yes	$p*m_1/2=0.026$	
$7.9*10^{-6}$	no	$p*m_1/3=0.026$	$p*m_2/1=0.075$
$11.8*10^{-6}$	yes	$p*m_1/4=0.030$	
$21.7*10^{-6}$	no	$p*m_1/5=0.043$	$p*m_2/2=0.103$

Removal of apoptosis genes leads to less discoveries!

Multiple testing

FWER or FDR?

- FDR usually preferable when screening is the pupose and validation on *independent* samples can be performed using a robust technique

- Interpretation of FDR is restricted to the context of the set of features tested. Good definition of this set is very critical

- FWER is preferable when the number of tests is small and no further validation is available

- FWER can be generalized to more powerful and liberal procedures, e.g. kFWER: $P(V > k) \leq \alpha$.