Multiple Testing

Alex Lewin

Department of Epidemiology and Public Health, Imperial College

Introduction

In biostatistics, many large data sets

- Large data sets in epidemiology, spatial data
- Gene expression data (microarrays)
- SNP (single nucleotide polymorphism) arrays

Same hypothesis test performed simultaneously on 1000s or 10,000s of statistics

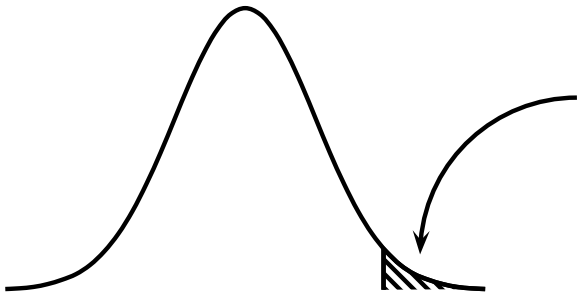
Contents

I Why Correct for Multiple Tests?

Each test i has null hypothesis H_{0i}

Test statistic T_i, observed value t_i

P-value (one-sided) $p_i = \mathbb{P}(T_i > t_i | H_{0i})$



Decision Rule: reject null hypothesis if $p_i < p^{cut}$

For one test threshold p^{cut} chosen to be pre-specified Type I error rate (e.g. $\alpha = 0.05$ or 0.01)

$\mathbb{P}(p_i < \alpha | H_{0i}) = \alpha$ since p-values are Uniform under the null hypothesis

I Why Correct for Multiple Tests?

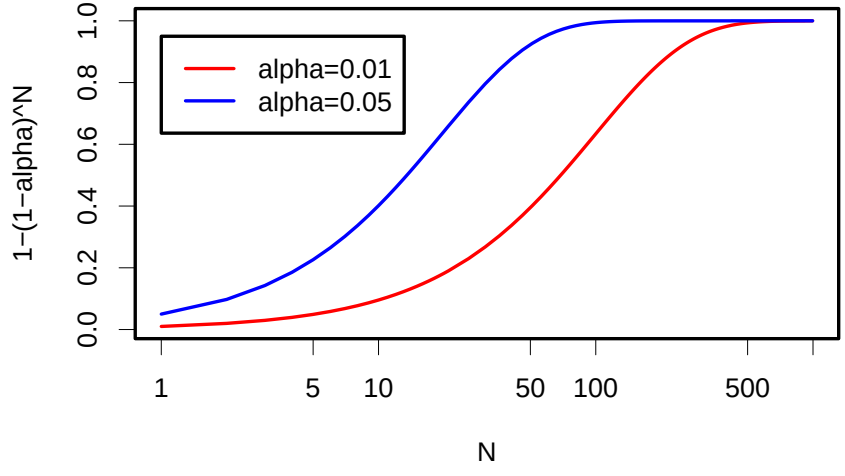
Now testing m hypotheses together (m may be 1000s).

What is the Decision Rule? Now we have $\{p_1, ..., p_m\}$.

For large m, very likely that one (or many) of $\{p_1, ..., p_m\} < \alpha$ even if all null hypotheses true.

If tests are independent,

$$\mathbb{P}(\text{at least one of } \{p_1, ..., p_m\} < \alpha | H_{01}, ..., H_{0m}) = 1 - (1 - \alpha)^m$$



I Example: Bonferroni Correction

Even if tests are dependent,

$$\begin{aligned}\mathbb{P}(\text{at least one of } \{p_1, ..., p_m\} < p^{cut} | H_{01}, ..., H_{0m}) &\leq \sum_{i=1}^{m} \mathbb{P}(p_i < p^{cut} | H_{0i}) \\ &= mp^{cut}\end{aligned}$$

Bonferroni Correction:

Set $p^{cut} = \alpha/m$

Decision Rule: for each i, reject H_{0i} if $p_i < \alpha/m$

This rule means $\mathbb{P}(\text{at least one of } \{p_1, ..., p_m\} < p^{cut} | H_{01}, ..., H_{0m}) \leq \alpha$.

II Different Error Rates

Given a decision rule, can specify various error rates based on the 2x2 table:

	Accepted	Rejected	
Null True	U	V	m_0
Null False	T	S	m_1
	$m - R$	R	m

V = false positives
T = false negatives
R = rejections
m_0 = true null hypotheses

Aim is to decide on a decision rule which controls or estimates a particular error rate.

II Different Error Rates

Reference: Dudoit, Shaffer and Boldrick 2003

- Family-Wise Error Rate FWER = $\mathbb{P}(V \geq 1)$
- Per Comparison Error Rate PCER = E(V/m)
- False Discovery Rate FDR = E(V/R)

	Accepted	Rejected	
Null True	U	V	m_0
Null False	T	S	m_1
	$m-R$	R	m

II Formulae

Suppose know number of true null hypotheses (m_0 out of m). Can calculate actual error rates.

Assume tests independent, then use Binomial distribution

FWER, PCER: Only need to know about true hypotheses:

$$FWER = \mathbb{P}(V \geq 1) = 1 - (1-\alpha)^{m_0}$$

$$PCER = E(V/m) = m_0\alpha/m$$

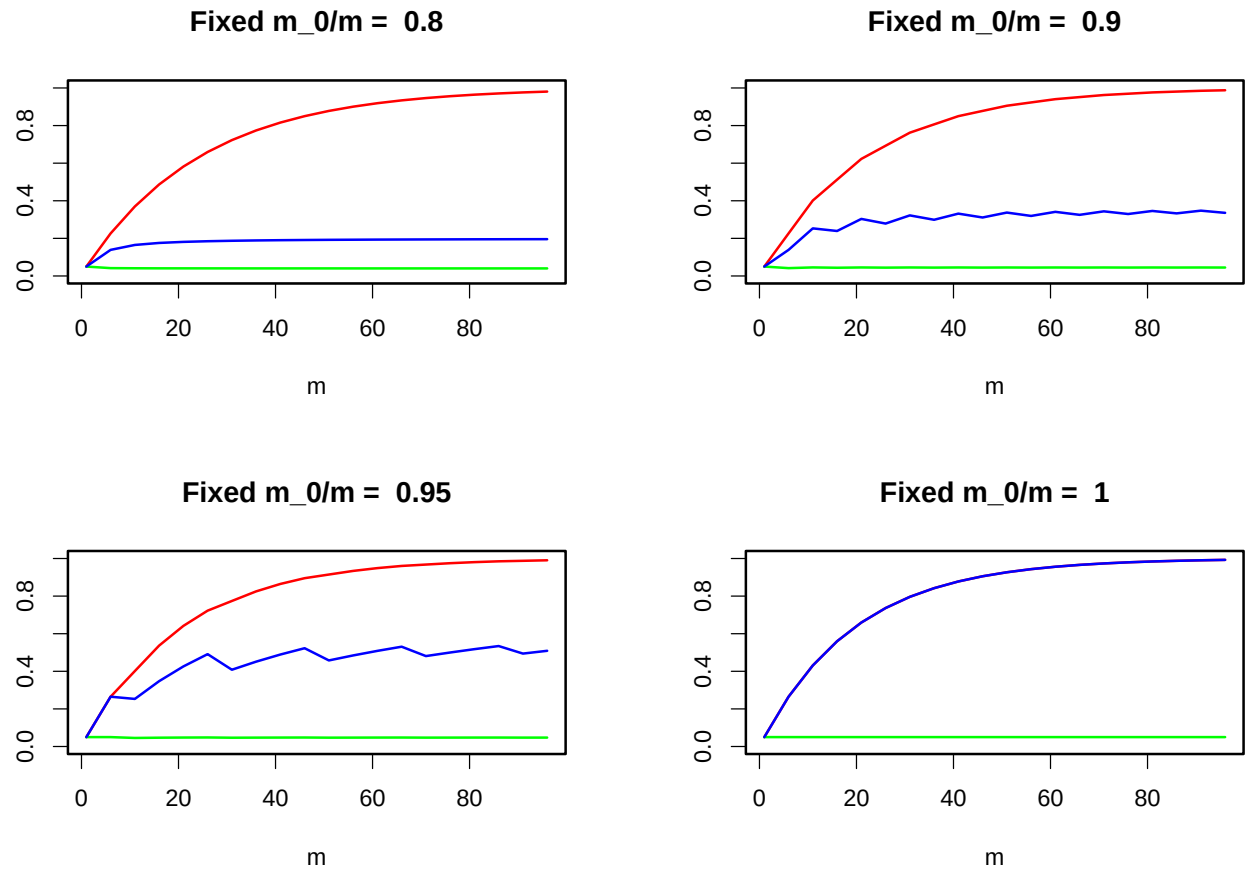
FDR: Need to know about true and false hypotheses:

$$E(V/R) = \sum_{v=1}^{m_0} \sum_{s=0}^{m-m_0} \frac{v}{(v+s)} \binom{m_0}{v} \alpha^v (1-\alpha)^{m_0-v} \binom{m-m_0}{s} \beta^s (1-\beta)^{m-m_0-s}$$

where β is the power ($\mathbb{P}(T_i > t_i | H_{1i})$, H_{1i} is the alternative)

II Error Rates as Functions of Number of Tests

red=FWER, green=PCER, blue=FDR



- PCER $\leq$ FDR $\leq$ FWER
- FDR stabilises as N increases

II Positive False Discovery Rate

Problem when $R = 0$
$V = 0$ when $R = 0$ so V/R is undefined

Possible Solutions

- Define $V/R = 0$ when $R = 0$ FDR = $E(V/R|R > 0)\mathbb{P}(R > 0)$
- Condition on $R > 0$ pFDR = $E(V/R|R > 0)$

Interesting Property of the pFDR
If test statistics $T_1, ..., T_m$ are i.i.d. as mixture of null and alternative hypotheses, pFDR can be written as

$$pFDR = \mathbb{P}(H_{0i}|T_i > t_i)$$

Note this is not a Bayesian quantity.

Reference: Storey (2003)

II Explanation of Property

Assume test statistics $T_1, ..., T_m$ are i.i.d. from mixture of null and alternative hypotheses.

$R = 1 \Longrightarrow V = 0$ or $V = 1$

Bernoulli with probability $\mathbb{P}(H_{0i}|T_i > t_i)$

Since test statistics independent,

$V|R \sim \text{Bin}(R, \mathbb{P}(H_{0i}|T_i > t_i))$

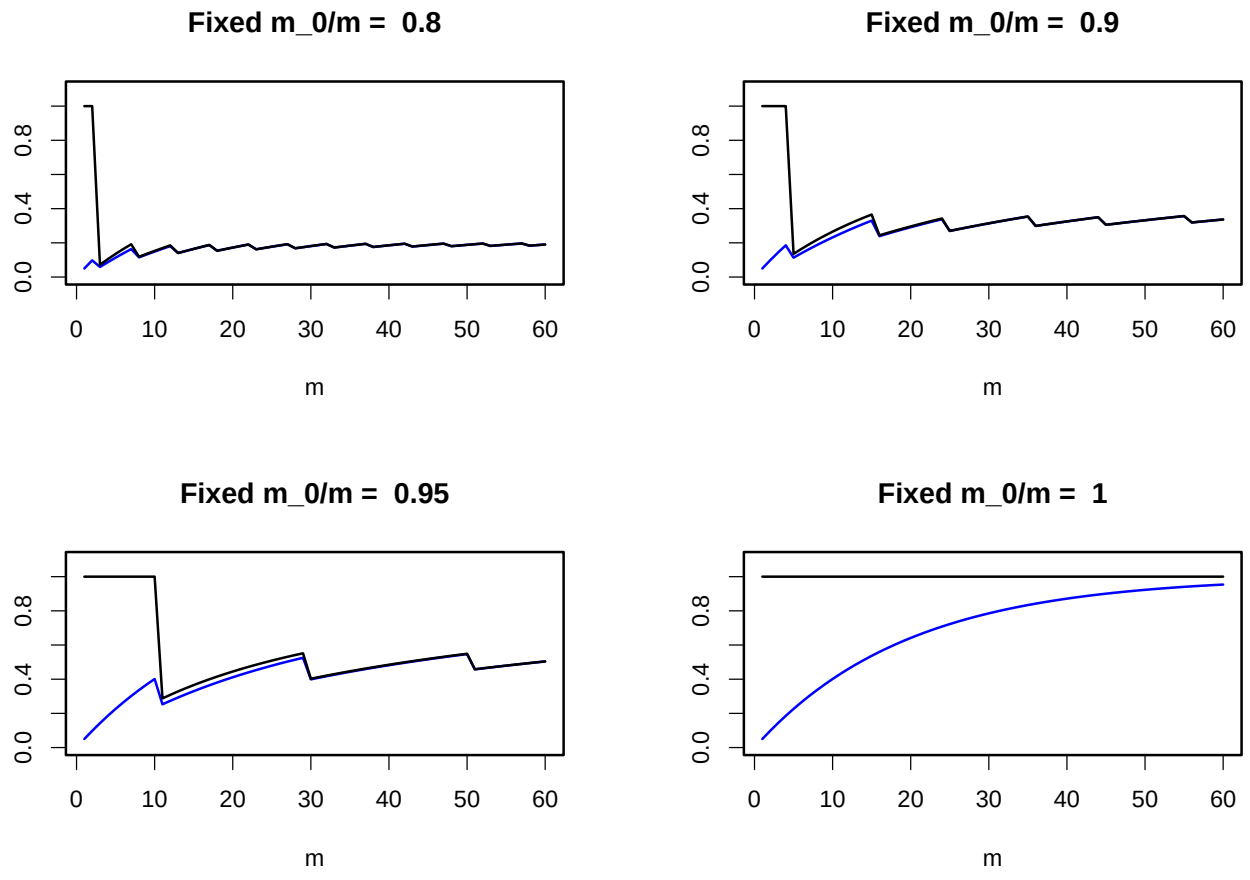
$\Longrightarrow E(V|R = r) = r\mathbb{P}(H_{0i}|T_i > t_i)$

$\Longrightarrow E(V/R|R = r) = \mathbb{P}(H_{0i}|T_i > t_i)$

$\Longrightarrow E(V/R|R > 0) = \mathbb{P}(H_{0i}|T_i > t_i)$

II Comparison of pFDR and FDR

blue=FDR, black=pFDR



- pFDR = 1 in situations where no hypotheses should be rejected
- For large m, pFDR=FDR

III Controlling Error Rates

- Decision rule: Reject H_{0i} if $p_i < p^{cut}$
- Calculate **a bound for** the error rate in terms of p^{cut},

$$\begin{aligned} \text{e.g. } FWER &= \mathbb{P}(V \geq 1 | m_0 \text{ true}, m - m_0 \text{ false}) \\ &= \mathbb{P}(\text{at least one } p_i < p^{cut} | m_0 \text{ true}, m - m_0 \text{ false}) \\ &\leq mp^{cut} \end{aligned}$$

- So to control the FWER at level α, we can set $mp^{cut} = \alpha$

NB: Using a bound for the error rate means it is **conservative** (expect to miss some hypotheses which should be rejected).

III Assumptions on Test Statistic Distributions

There are lots of different methods of controlling the FWER.
Why? Because different assumptions are made on the distributions of test statistics.

- **Strong/Weak Control**
  In previous section (II) we knew how many null hypotheses were true (m_0). What do we do in real experiments, where m_0 is unknown?
  - **Weak Control**
    Error rates controlled only if conditioned on all hypotheses being true ($m_0 = m$).
  - **Strong Control**
    Error rates conditioned on any combination of hypotheses being true or false ($m_0 \neq m$).
- **Dependence Assumptions**
  Many methods assume p-values are independent. Some make no assumptions about dependence, others somewhere in between.

III Examples of Distributional Assumptions

Method	Strong/weak control	Dependence
Bonferroni $FWER \leq mp^{cut}$	strong ($m_0 \neq m$)	can be dependent
Sidak $FWER \leq 1-(1-p^{cut})^m$	strong ($m_0 \neq m$)	independent
Westfall and Young (re-sampling based method to control FWER)	weak ($m_0 = m$)	can be dependent

Trade-off between assumptions and conservativeness

III Step-wise Procedures

More powerful (less conservative), without changing assumptions on distributions of test statistics.

Holm Step-down method for controlling FWER at level α

Order p-values: $p_1^{ord}, ..., p_m^{ord}$

$$
\begin{array}{rcll}
p_1^{ord} & \leq & \alpha/m & \Longrightarrow \text{ reject} \\
p_2^{ord} & \leq & \alpha/(m-1) & \Longrightarrow \text{ reject} \\
 & \vdots & & \\
p_{(j^*-1)}^{ord} & \leq & \alpha/(m-j^*+2) & \Longrightarrow \text{ reject} \\
p_{j^*}^{ord} & > & \alpha/(m-j^*+1) & \Longrightarrow \text{ don't reject} \\
 & & & \text{All others don't reject}
\end{array}
$$

Analogue of the Bonferroni correction

III Step-wise Procedures

Holm Step-down method for controlling FWER at level α

- Order p-values: $p_1^{ord}, ..., p_m^{ord}$
- j^* = label of minimum ordered p-value with $p_j^{ord} > \alpha/(m - j + 1)$
- $p^{cut} = p_{(j^*-1)}^{ord}$

Can also define step-down procedures for the Sidak and Westfall and Young methods (making the same respective distributional assumptions).

IV Control of the False Discovery Rate

FDR = $E(V/R)$

Benjamini and Hochberg Step-up procedure (controls FDR at level α):

- Order p-values: $p_1^{ord}, ..., p_m^{ord}$
- j^* = label of maximum ordered p-value with $p_j^{ord} \leq \alpha j/m$
- $p^{cut} = p_{j^*}^{ord}$
- Strong control at level α

Other procedures have been proposed which use permutations/re-sampling to allow for dependence between p-values.

IV Estimating the positive False Discovery Rate

pFDR = $E(V/R|R>0)$

This cannot be controlled because it is 1 when $m_0 = m$. Instead it is estimated.

Recall the 'interesting property': $pFDR = \mathbb{P}(H_{0i}|T_i > t_i)$. This can be used to provide a method to estimate the pFDR:

$$\begin{aligned}\mathbb{P}(H_{0i}|T_i > t_i) &= \mathbb{P}(T_i > t_i|H_{0i})\mathbb{P}(H_{0i})/\mathbb{P}(T_i > t_i) \\ &= p^{cut}\pi_0/(R/m)\end{aligned}$$

where π_0 is the proportion of true null hypotheses.

Therefore, $\hat{pFDR} = mp^{cut}\hat{\pi}_0/R$

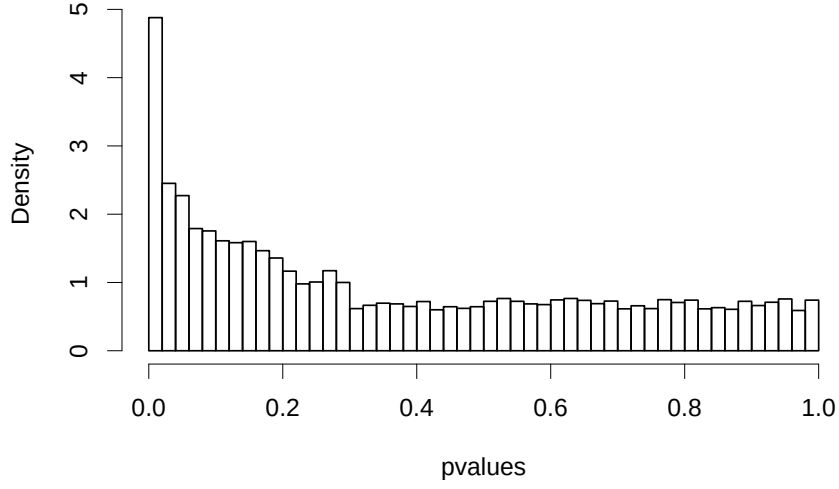
Here the p^{cut} used can be determined by the required $pFDR$, or the other way round.

Remaining problem is $\hat{\pi}_0$.

IV Estimating π_0

Assuming test statistics $T_1, ..., T_m$ are i.i.d. from mixture of null and alternative hypotheses,
$T_i \sim \pi_0 H_{0i} + (1 - \pi_0) H_{1i}$

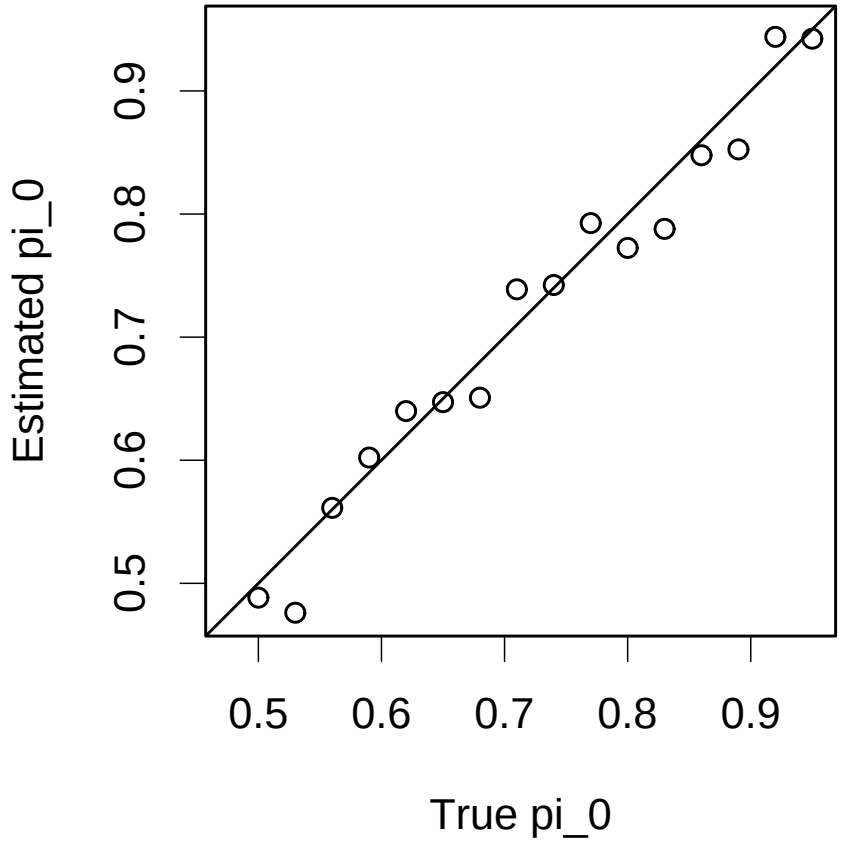
Under the null, p-values are Uniform. If $\pi_0 \neq 1$ expect an excess of small p-values (from those individuals where the null hypothesis is false).



π_0 is estimated from the flat part of the histogram (various methods, including spline-fitting).

IV Estimation of π_0

Simulated data: null pvalues from Uniform distribution, alternative from Beta distribution



Curve not smooth due to variability in simulated data.

Storey method often conservative in estimation of π_0

IV Connection Between Benjamini and Hochberg and Storey Procedures

Storey

For given cut-off p^{cut},

Estimate of pFDR is $p\hat{FDR} = \hat{\pi}_0 m p^{cut}/R$

where R = no. p-values $\leq p^{cut}$ (so R is determined by p^{cut})

Benjamini and Hochberg

To control FDR at level α

$p^{cut} = \alpha j^*/m$ where j^* is the label of the largest rejected p-value

Here $R = j^*$

Switch this around:

$FDR = m p^{cut}/R$ where R = no. p-values $\leq p^{cut}$

IV Connection Between Benjamini and Hochberg and Storey Procedures

Storey

$$\hat{pFDR} = \hat{\pi}_0 m p^{cut}/R$$

Benjamini and Hochberg

$$FDR = m p^{cut}/R$$

If same p^{cut} used in the two methods, $\hat{pFDR} = \hat{\pi}_0 FDR$

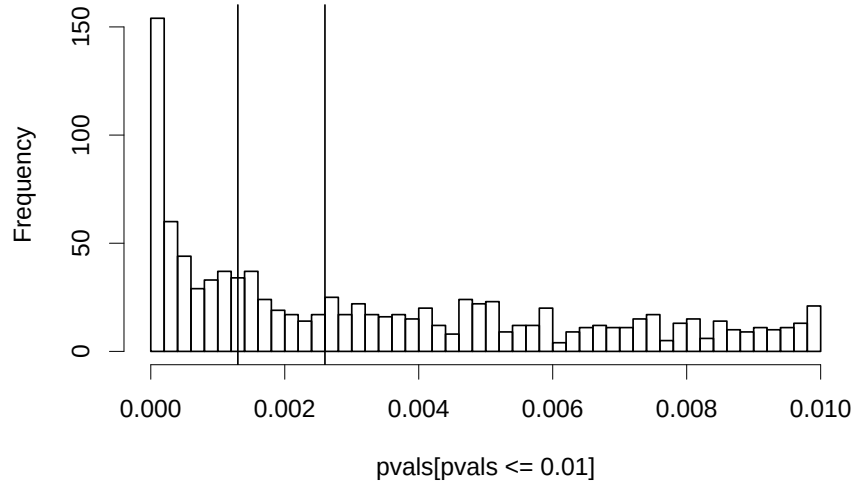
so Storey pFDR is controlled at a lower level than BH FDR.

Conversely, if control both methods at the same level, Storey method can reject more null hypotheses than BH method.

Difference between BH and Storey only due to inclusion of π_0, (recall $pFDR = FDR$ for large m)

IV Connection Between Benjamini and Hochberg and Storey Procedures

Example for p-values shown previously, with FDR, pFDR controlled/estimated at level $\alpha = 5\%$:



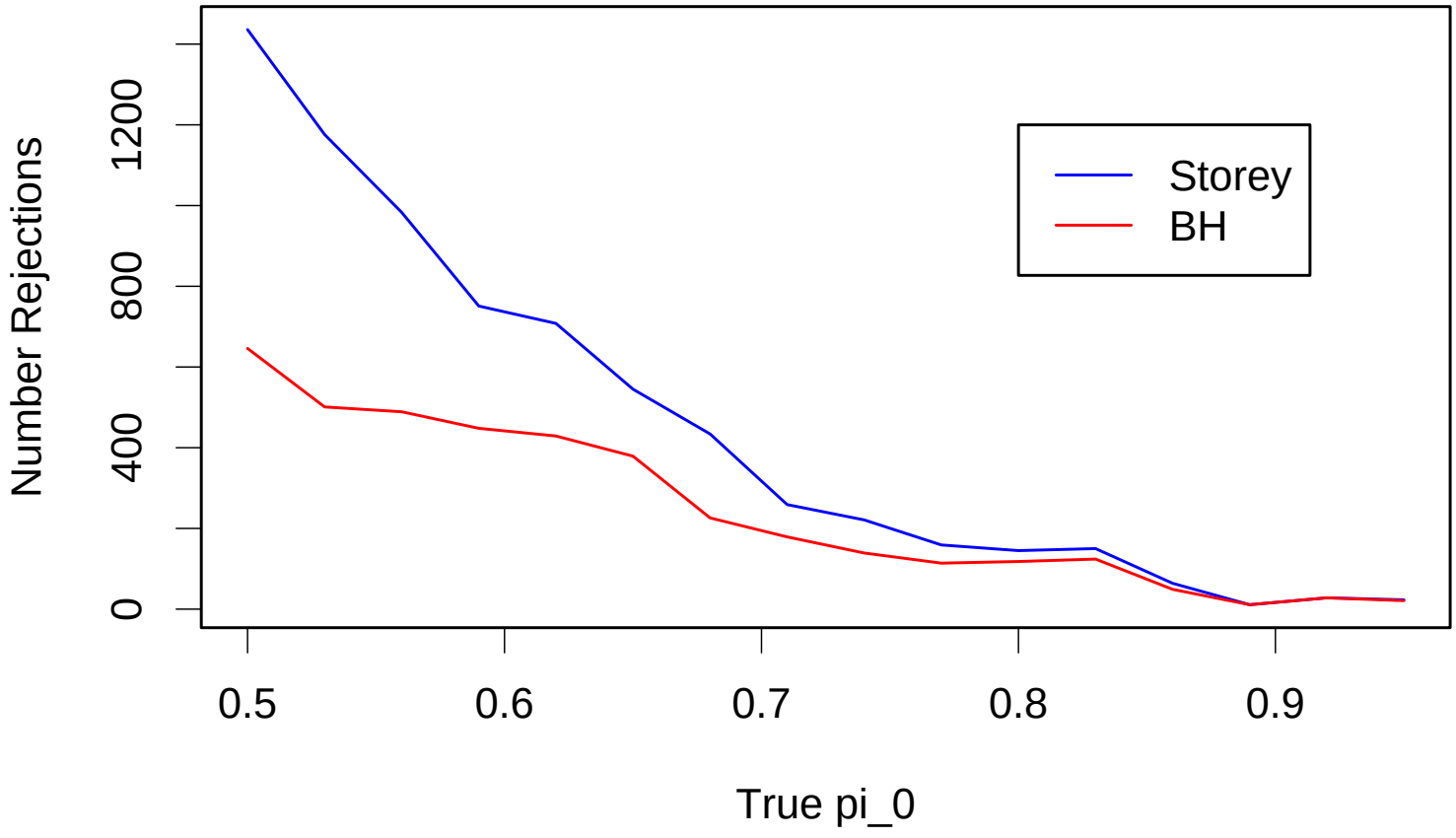
Vertical lines show the p-value thresholds for BH and Storey methods ($p^{cut}_{BH} < p^{cut}_{Storey}$).

Storey method estimates $\hat{\pi}_0 = 0.68$. The true value is 0.69.

- Difference between BH and Storey is only due to inclusion of π_0
- This depends on a good estimate for π_0

IV Compare BH and Storey Procedures

Same simulation set-up as before (null Uniform, alternative Beta). 10,000 tests.

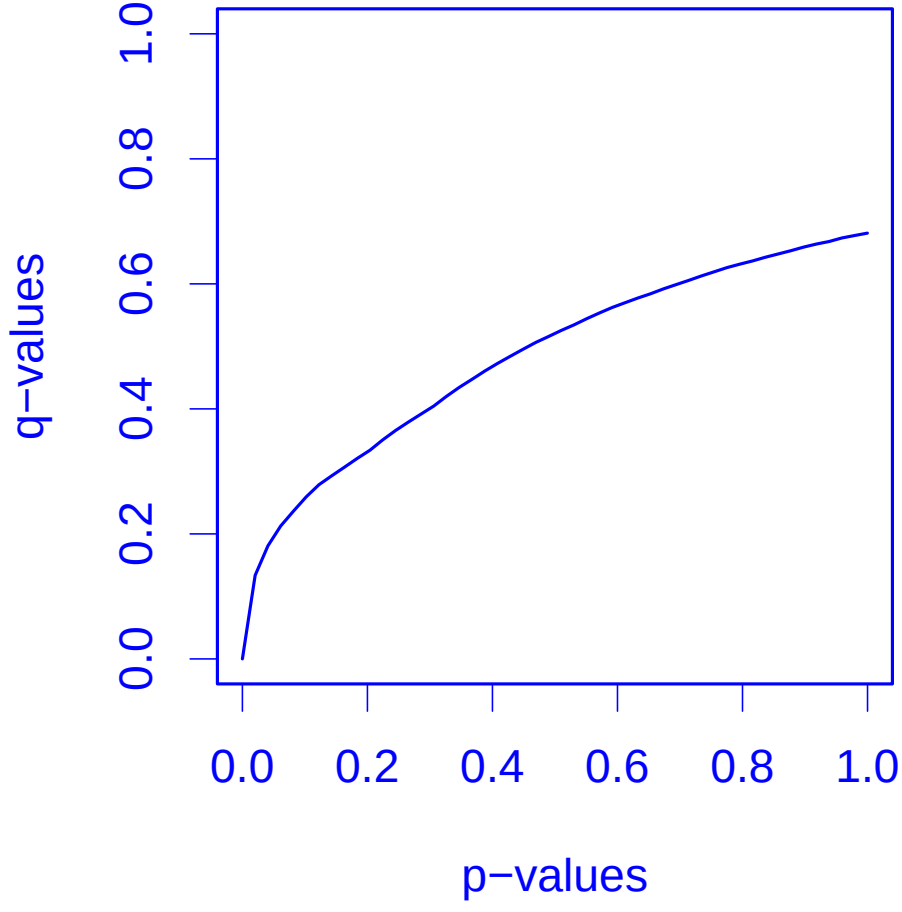


For large values of $\pi_0 \equiv m_0/m$ BH and Storey procedures give very similar results.

IV Q-Values

Storey proposed defining a quantity called the q-value:

For feature i, the q-value q_i is the $pFDR$ for the list of significant features obtained by rejecting all hypotheses with p-values $p_{i'} < p_i$.



IV False Non-Discovery Rate

Define analogous quantity for hypotheses which are not rejected:

$pFDR = E(V/R|R > 0)$

$pFNR = E(T/(m-R)|(m-R) > 0)$ where $T = (m - m_0) - (R - V)$

	Accepted	Rejected	
Null True	U	V	m_0
Null False	T	S	m_1
	$m - R$	R	m

IV Bayes Rule

Storey method estimates $\hat{\pi}_0 = 68\%$, but only rejects 4% of hypotheses. Why?

Because set $pFDR = 5\%$, $pFNR$ much higher

Balance pFDR and pFNR

Define loss function for mis-classification in 2x2 table

$$\begin{aligned} E(\text{loss}) &= \lambda \mathbb{P}(\text{false negative})\mathbb{P}(\text{accept}) + (1-\lambda)\mathbb{P}(\text{false positive})\mathbb{P}(\text{reject}) \\ &= \lambda pFNR(m - R/m) + (1-\lambda)pFDR(R/m) \end{aligned}$$

$\lambda = 1/2$ is Bayes Rule

This can be used with Storey method, and in the Bayesian framework.

V Bayesian Estimate of pFDR

Storey's result

$$pFDR = \mathbb{P}(H_0 \text{ true} \mid \text{reject } H_0)$$

suggests Bayesian method where estimate for pFDR is a posterior probability, conditioned on the data:

$$pFDR_{Bayes} = \mathbb{P}(H_0 \text{ true} \mid \text{reject } H_0, \text{data})$$

References:
Broët, Lewin, Richardson, Dalmasso & Magdelenat (2004)
Newton, Noueiry, Sarkar & Ahlquist (2004)

V Bayesian Mixture Model

Data (or function of data) $t_i; i = 1, ..., m$ from mixture of null and alternative:

$$t_i \sim \pi_0 f_0 + (1 - \pi_0) f_1$$

For all features can calculate posterior probability of allocation to null component:

$$\begin{aligned} \pi_i &\equiv \mathbb{P}(\text{null hypothesis true for } i | \text{data}) \\ &= \frac{\pi_0 f_0(t_i)}{\pi_0 f_0(t_i) + (1 - \pi_0) f_1(t_i)} \end{aligned}$$

Given that the null hypotheses in set S are rejected,

$$pFDR_{Bayes} = \frac{1}{|S|} \sum_{j \in S} \pi_j$$

Set S can be any set of hypotheses, not just those based on ranking statistics.

V Comparison of Storey and Bayesian Estimation

Bayesian method

Starts with posterior probabilities π_i

$\pi_i = \mathbb{P}(H_{0i} \text{ true} | \text{data})$

$\pi_0 = \mathbb{P}(H_{0i} \text{ true}) \ \forall i$ can also be estimated in model

Estimate $\pi_i \ \forall i \longrightarrow$ can calculate pFDR for any set of features

Storey method

Starts with p-values $\mathbb{P}(\text{reject } H_{0i} | H_{0i} \text{ true})$

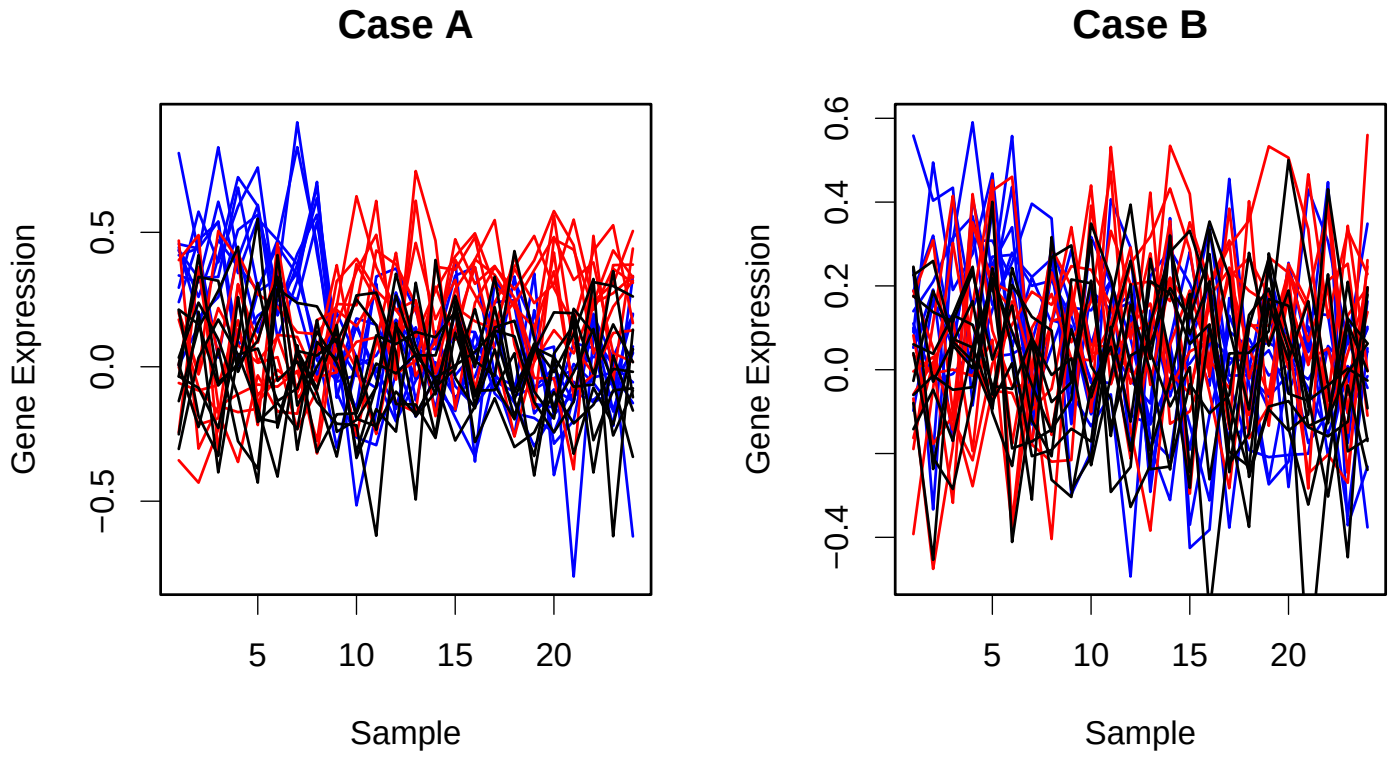
Estimates $\pi_0 = \mathbb{P}(H_{0i} \text{ true}) \ \forall i$

No estimate of $\mathbb{P}(H_{0i} \text{ true} | \text{data})$ for individual features i.

No posterior information on individual features $\longrightarrow$ pFDR must be estimated on monotonic rejection regions

V Example: Simulated Gene Expression Profiles

Each gene has 8 measurement of expression for each of 3 experimental conditions

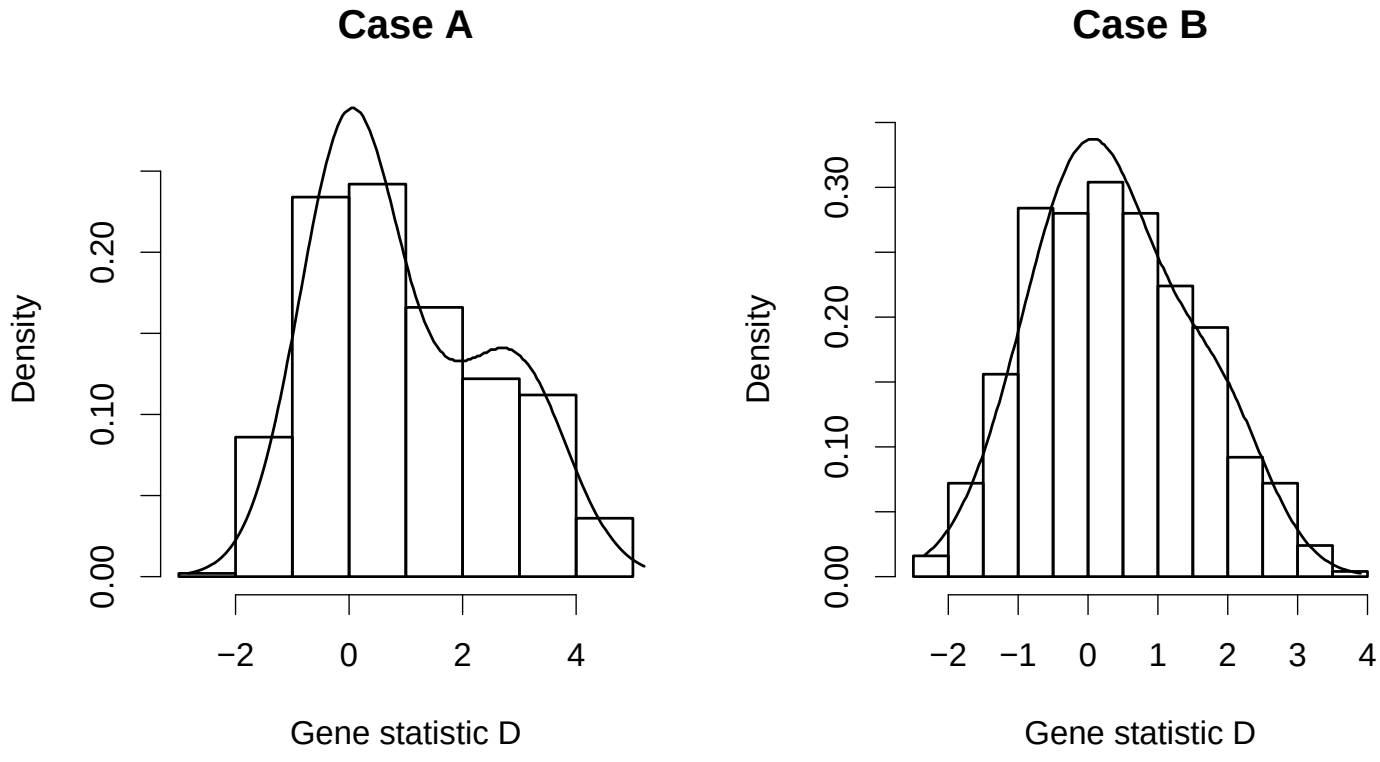


Null hypothesis: no change in expression across 8 conditions

Summarize profile by F-statistic (one for each gene)

V Example: Simulated Gene Expression Profiles

Transform F-statistics to D-statistics, D approx. Normal if no change across conditions



V Estimate pFDR

Bayesian mixture model:

$$D_i \sim \pi_0 N(0, \sigma_0^2) + \sum_{j=1}^{k} \pi_j N(\mu_j, \sigma_j^2)$$

- μ_j ordered, Uniform on upper range
- No. mixture components in alternative (k) unknown (estimated by reversible jump MCMC) + results integrated over k
  $\Longrightarrow$ alternative modelled semi-parametrically
- pFDR uses posterior probabilities of genes being allocated to null component

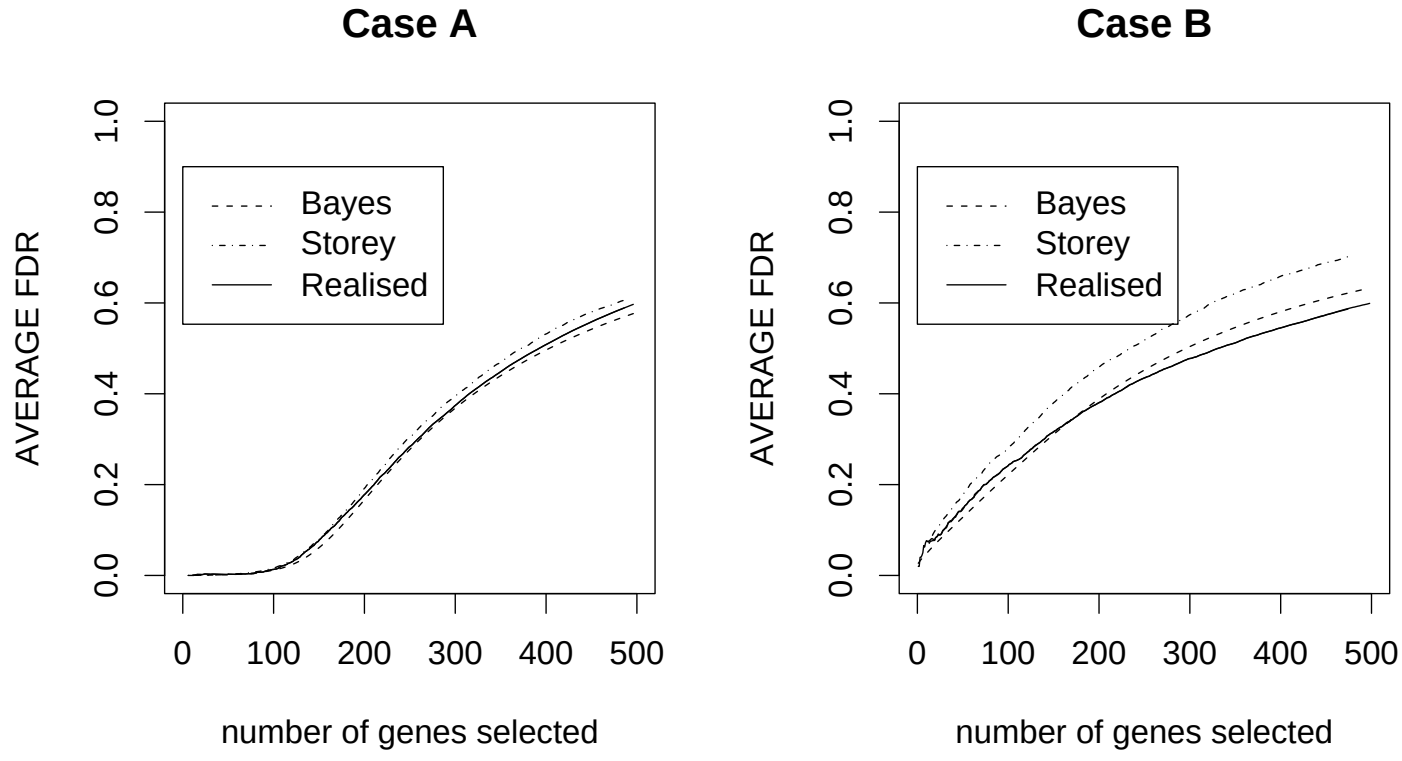
Storey method:

- pFDR uses the usual p-values obtained by ANOVA on profile data.

So both method start with the F-statistics.

V Results for Simulated Expression Profiles

Simulation study: 50 simulations of the 2 sets of profiles.



Both methods estimate pFDR very well.

Bayesian method does slightly better when profiles are closer together (Case B).

V $pFDR_{Bayes}$ for any set of hypotheses

Breast Cancer Data

Study of gene expression changes among 3 types of tumour: BRCA1, BRCA2 and sporadic tumours (Hedenfalk et al.)

Bayesian mixture model (as above) gives posterior probabilities $\pi_i = \mathbb{P}(H_{0i} \text{ true} | \text{data})$ for all genes.

Can estimate $pFDR_{Bayes}$ for **any** group of genes.

3 pre-defined groups of genes with known functions:

- apoptosis $\quad pF\hat{D}R_{Bayes} = 70\%$
- cell cycle regulation $\quad pF\hat{D}R_{Bayes} = 26\%$
- cytoskeleton $\quad pF\hat{D}R_{Bayes} = 66\%$

Lower $pF\hat{D}R_{Bayes}$ for cell cycle regulation suggests this group of genes more differentially expressed between the different tumour types.

Summary

- Controlling Error Rates

  Many different methods, single step, step-wise, re-sampling,...

- Estimating False Discovery Rate

  Use estimate of proportion of null hypotheses to estimate the FDR.

  Less conservative than controlling.

- Bayesian Methods

  Can calculate a posterior quantity analogous to the pFDR.

  Can be used for any group of hypotheses.

References

- Review of frequentist methods, controlling error rates:

  Dudoit, Shaffer & Boldrick (2003), *Statistical Science 18, p 71.*

  http://www.stat.berkeley.edu/ sandrine/

- pFDR:

  Storey (2003), *Annals of Statistics, 31, p 2013.*

  And other papers ...

  http://faculty.washington.edu/ jstorey/

- Bayesian estimation of pFDR:

  Newton, Noueiry, Sarkar & Ahlquist (2004). *Biostatistics 5, p 155.*

  http://www.stat.wisc.edu/ newton/

  Broët, Lewin, Richardson, Dalmasso & Magdelenat (2004), *Bioinformatics 20(16), p 2562.*

  http://www.bgx.org.uk/alex/