

▪

Multiple Testing: the problem and some Bayesian and Frequentist solutions

M.J. Bayarri and J.O. Berger
Valencia, Duke and SAMSI

*MFO, Oberwolfach 2005,
October 16–22, 2005*

The Problem

1981

- Prince Charles get married
- Liverpool gains the European Championship League
- Pope dies

2005

- Prince Charles get married
- Liverpool gains the European Championship League
- Pope dies

The Solution

Soooo ...

If Prince Charles ever announces that he is getting married again and Liverpool gets into the Finals, go and alert the Pope!

The 'old' multiple comparison problem refers to detecting an excess of 'happenings' (signal, discoveries) when using the same data to perform many tests. Modern concerns arise even if each test 'uses its own data'

Many tests

- An increasingly common 'massive multiple comparison' situation is simultaneous screening of many (hundreds or thousands) of hypotheses to determine whether we have 'noise' or 'signals'.
- A typical example is in gene expression (microarrays), when many genes are tested for differential expression among different treatments
- Other examples occur in 'Anomaly Discovery'; for example, in 'Syndromic Surveillance' many countries perform daily tests on the 'excess' of some symptoms, the goal being early detection of the outbreak of epidemics or of bio-terrorist attacks.

- Assume observables $\mathbf{X} = (X_1, X_2, \dots, X_M)$. Want to perform M tests of hypotheses. For $i = 1, \dots, M$, test

$$H_{0i} : X_i \sim f_{0i} \quad \text{vs.} \quad H_{1i} : X_i \sim f_{1i}$$

X_i can be a vector. Often X_i is a test statistic or a p -value, and f_{1i} and f_{0i} involve unknown parameters.

- Let $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_M)$, with
 - $\gamma_i = 0$ if the i^{th} null is true
 - $\gamma_i = 1$ if the i^{th} alternative is true
- The multiple testing problem can thus be formulated as a model selection problem: to choose among the 2^M models indexed by the possible values of γ

- It has specific characteristics:
 - From a frequentist point of view, it is usually addressed using generalizations of hypothesis testing tools; rarely if ever using Model Selection tools
 - From a Bayesian point of view, and in contrast with most problems of model choice, the posterior probability of the models is not the object of main interest, but rather the probability that each γ_i is non 0 (inclusion probabilities)
- If the M tests are **independent** and each is tested at level α , then, even when $\gamma = \mathbf{0}$, we expect αM rejections; In simultaneous testing, this is perceived as ‘too many’, unduly masking detection of incorrect null

An example from Syndromic Surveillance

(Stoto et al., Chance, 2004)

“... Suppose every county in the USA has an independent statistical algorithm in place that is used daily and that has a 0.1% false-positive rate (Type I error = .001). Because there are approximately 3000 counties, on average three counties a day would have a false-positive alarm. The impact of false alarm is financial and psychological. Response to phantom events cost money and false alarms desensitize responders to real ones. While any particular county would experience a false-positive alarm only about once every three years, this nationwide false-positive rate would be unacceptable.”

The problem of multiplicity

- This is stated as the problem of ‘multiplicity in testing:’ as the number of simultaneous tests being conducted increases, the criterion for rejection must become more strict
- Note: this is not the ‘old’ multiplicity problem in which the *same* data is used for several tests; Here each test has its own (independent) data.
- In these massive screenings, there tends to be strong prior probability that there are few signals, that is, that many of the γ_i 's are 0. Recent frequentist analyses take this into account.

Controlling error rates

- The problem of multiple testing has to address:
 - What error-rate to “control”
 - What type of “control” is desired
- An error rate can be controlled:
 - Only when *all* nulls are true ($\gamma = \mathbf{0}$) \rightsquigarrow *weak* control
 - For all values of γ (all combinations of true and false nulls) \rightsquigarrow *strong* control

Reviews & references \rightsquigarrow Shaffer(95); Dudoit et al.(03); Yang & Rempala(04)

Per Comparison, Family-wise, FDR error-rates

For M tests, call:

	accept H_0	Reject H_0	
H_0 true ($\gamma_i = 0$)	U	V	$M_0 = M - M_1$
H_0 false ($\gamma_i = 1$)	T	S	$M_1 = \sum_{i=1}^M \gamma_i$
(observed \rightarrow)	W	R	M

$R \rightsquigarrow$ total number of rejections (discoveries)

$V \rightsquigarrow$ # false discoveries

(There is little concern about T in the non-Bayesian literature, a questionable omission when viewed decision-theoretically.)

Per-Comparison (PCER)

	d_0	d_1	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

- Based on % of false rejections $\frac{V}{M}$
- Controls $\frac{E[V]}{M} = \frac{M_0\alpha}{M} \leq \alpha$
- Controlled (strong) by testing each H_{0i} at level α
- accused of 'ignoring the multiplicity problem' (too liberal)

Family-wise (FWER)

	d_0	d_1	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

- Classical solution to deal with multiplicity
- Based on total number of false rejections (discoveries) V
- Controls $\Pr(V \geq 1)$
- Bonferroni \rightsquigarrow controlled (strong) at level $\leq \alpha$ by testing each H_{0i} at level $\frac{\alpha}{M}$
- results in *very conservative* tests which can result in very low power.

False Discovery rate (FDR)

	d_0	d_1	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

- Benjamini and Hochberg (95) argued that the interesting quantity is the % of false discoveries (erroneous rejections) **among the rejected hypotheses**
- Ideally based on $\frac{V}{R}$
- But this is not defined for $R = 0$ (all M nulls accepted)
- Several associated 'error rates' to assess or control:

Some possibilities \rightsquigarrow to control:

	d_0	d_1	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

- False discovery rate:

$$\text{FDR} = E \left[\frac{V}{\max\{R, 1\}} \right] = E \left[\frac{V}{R} \mid R > 0 \right] Pr(R > 0)$$

- Positive false discovery rate: $\text{pFDR} = E \left[\frac{V}{R} \mid R > 0 \right]$
- Proportion of false discoveries

$$\text{PFP} = \frac{E[V]}{E[R]}$$

Also \leadsto to 'control' an error rate (ER):

- Fix the acceptable ER (α) beforehand and get a (data dependent) thresholding rule such that $ER \leq \alpha$ (this is really what it is meant by 'controlling' the ER)
- Fix the thresholding rule (critical region) and form a conservative estimate of the ER (larger than the true ER over the rejection region). 'Control' is then based on this estimate.

Note: Bayesians would, for a fix thresholding rule, find the expected loss; the optimal thresholding (decision rule) would be the one minimizing the expected loss.

(Expected) False Discovery Rate

	d_0	d_1	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

Benjamini and Hochberg (1995) was the pioneering work that started the interest in the FDR-type error rates.

They argue that, when all the nulls are true, ($M_0 = M$):

$$\text{pFDR} = E\left[\frac{V}{R} \mid R > 0\right] = 1 = \frac{E[V]}{E[R]} = \text{PFP}$$

Thus neither pFDR nor PFP can be controlled, so they choose FDR as a ‘good’ error rate which can be controlled

$\text{FDR} \leq \text{FWER}$, with $\text{FDR} = \text{FWER}$ when all nulls are true

\rightsquigarrow control of FDR produces less conservative tests

B&H algorithm:

1. compute ordered p-values $p_{(1)} \leq p_{(2)}, \dots, \leq p_{(M)}$
2. compute $D = \max \{1 \leq i \leq M : p_{(i)} \leq \alpha \frac{i}{M}\}$
3. if D exists, reject all nulls corresponding to $p_{(1)} \leq p_{(2)}, \dots, \leq p_{(D)}$, otherwise reject nothing.

Properties:

- Simes(86) shows *weak* control of FWER
- B&H(95) show *strong* control of FDR
- Genovese and Wasserman(02, 03) study asymptotic behavior, operating characteristics and many properties of FDR procedures. They also show that the effective α is between the unadjusted α and Bonferroni's α/M .

Some Relations

	d_0	d_1	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

- asymptotically $FDR \approx pFDR \approx PFP = \frac{E[V]}{E[R]}$
- let p^* s.t. reject in tests for which $p_{(j)} \leq p^*$, then

$$\widehat{pFDR} = \hat{p}_0 M p^* / R$$

$$FDR = M p^* / R$$

where $p_0 =$ proportion of true nulls ($R = \#$ rejected). Hence

- If same p^* in both $\leadsto \widehat{pFDR} = \hat{p}_0 \times FDR$
pFDR 'controlled' at lower level than FDR
- If same level on both \leadsto pFDR can reject more nulls
increased power due to inclusion of \hat{p}_0

- Increasing the power of FDR (Benjamini and Hochberg, 2000; Black, 2004)
 - B&H algorithm controls FDR at level $p_0\alpha$ (Finner and Roters, 04) \rightsquigarrow use and estimate of p_0 to increase power (keeping control at level α)
 - ‘Adaptive’ modification of B&H algorithm: Compute $D^* = \max\{i : p_{(i)} \leq \frac{\alpha}{\hat{p}_0} \frac{i}{M}\}$ and reject for $p_{(i)} \leq p_{(D^*)}$
 - Adaptive modification practically identical to fixed rejection region method of Storey (Black 04)
- For summary and references on frequentist estimates of p_0 see Langaas et al. (2005). **Note:** In Bayesian analyses, estimation of p_0 , if of interest, is a byproduct.

Models for Bayes/Empirical Bayes analyses

(Efron et.al.; Storey; Genovese & Wasserman; Müller ...)

- Consider *independent* tests
- Let $p_0 = Pr(\gamma_i = 0)$; assume $\gamma_i \mid p_0 \stackrel{i.i.d}{\sim} Ber(1 - p_0)$
note: marginally γ_i are exchangeable but *not* independent
- Also, $X_i \mid \gamma_i = 0 \sim f_0$ (null distribution),
 $X_i \mid \gamma_i = 1 \sim f_1$ (alternative distribution)

Marginally, $X_i \sim f$, the mixture model:

$$f(x_i) = p_0 f_0(x_i) + (1 - p_0) f_1(x_i)$$

- Often p_0 and f_1 are considered “unknown”.

A simple Bayesian example (Scott and Berger, 06)

(More sophisticated Bayesian analyses are in Newton et al. 01,04; Newton and Kendzioriski 03; Gönen et al. 03; Müller et al. 04; Do, Müller, and Tang 05; House, Clyde and Huang, 06.)

- Observe $X_i \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, M$, (σ^2 unknown).
- To determine which μ_i are nonzero \rightsquigarrow we have M (conditionally) independent tests, each testing

$$H_{0i} : \mu_i = 0 \quad \text{vs} \quad H_{1i} : \mu_i \neq 0$$

- $p_0 =$ prior probability that μ_i is zero, $p_0 = \Pr(\gamma_i = 0)$.
- *Crucial* point in multiple testing: let data estimate p_0

S&B Use the hierarchical model

1. $X_i \mid \mu_i, \sigma^2, \gamma_i \stackrel{iid}{\sim} N(\gamma_i \mu_i, \sigma^2)$
2. $\mu_i \mid \tau^2 \stackrel{iid}{\sim} N(0, \tau^2), \quad \gamma_i \mid p_0 \stackrel{iid}{\sim} Ber(1 - p_0)$
3. $(\tau^2, \sigma^2) \sim \pi(\tau^2, \sigma^2), \quad p_0 \sim \pi(p_0)$

Prior for (τ^2, σ^2)

$$\pi(\tau^2, \sigma^2) = \pi(\tau^2 \mid \sigma^2) \pi(\sigma^2) = (\tau^2 + \sigma^2)^{-2} \frac{1}{\sigma^2}$$

Note: The usual objective prior $\pi(\tau^2 \mid \sigma^2) \propto (\tau^2 + \sigma^2)^{-1}$ can not be used because it is improper and τ^2 does not occur in all models (in particular, when $\gamma = \mathbf{0}$).

prior for p_0

- A ‘neutral’ (objective) prior $\rightsquigarrow p_0 \sim Un(0, 1)$
- An easy to assess prior that allows for beliefs that p_0 (\approx % of true nulls) is likely to be large

$$p_0 \sim Beta(a, 1) \quad \text{with} \quad a = \frac{\log(.5)}{\log(\hat{p}_0)} - 1$$

where \hat{p}_0 is a ‘best guess’ for p_0 (interpreted as the prior median)

- Note that $a = 1$ gives the Uniform(0, 1) density.
- In spite of the improper (joint) prior, the joint posterior distribution of all unknowns can be shown to be proper

the quantities of interest here are

- the probabilities that each of the hypothesis is true (the *inclusion probabilities*)
- the distributions of the ‘signals’ if the corresponding null is not true.
- curiously, the posterior probability of models $\pi(\gamma \mid \mathbf{x})$, is not of main interest

The posterior probability π_i that $\mu_i \neq 0$ (a ‘signal’) is

$$\pi_i = 1 - \frac{\int_0^1 \int_0^1 p_0 \prod_{j \neq i} \left(p_0 + (1 - p_0) \sqrt{1 - w} e^{wx_j^2 / (2\sigma^2)} \right) dp_0 dw}{\int_0^1 \int_0^1 \prod_{j=1}^M \left(p_0 + (1 - p_0) \sqrt{1 - w} e^{wx_j^2 / (2\sigma^2)} \right) dp_0 dw}.$$

computed numerically or by importance sampling (large M)

Example:

- Consider the following ten 'signal' observations:
-5.65, -5.56, -2.62, -1.20, -1.01, -.90, -.15, 1.65, 1.94, 3.57
generated as $\mu_i \sim N(0, 3^2)$, and then $X_i \sim N(\mu_i, 1)$.
- Consider increasingly 'worse' multiple comparisons scenarios: Generate $n = 25, 100, 500,$ and 5000 noise observations $X_i \sim N(0, 1)$, and mix in with the signals.
- Consider the uniform prior for p_0 and the prior $\pi(p_0) = 11p_0^{10}$ (subjective prior median for p_0 of 0.93).
- Compute posterior probabilities of the 'signal' means being nonzero.

	7 of the Signal Observations						
n	-5.56	-2.62	-1.20	-0.90	-0.15	1.94	3.57
25	0.97	0.71	0.31	0.26	0.20	0.51	0.88
100	0.99	0.47	0.21	0.19	0.16	0.31	0.75
500	1	0.34	0.07	0.06	0.04	0.15	0.79
5000	1	0.11	0.02	0.02	0.01	0.04	0.42

Posterior inclusion probabilities

$$\pi_i = Pr(\gamma_i = 1 \mid \mathbf{x}) = Pr(\mu_i \neq 0 \mid \mathbf{x})$$

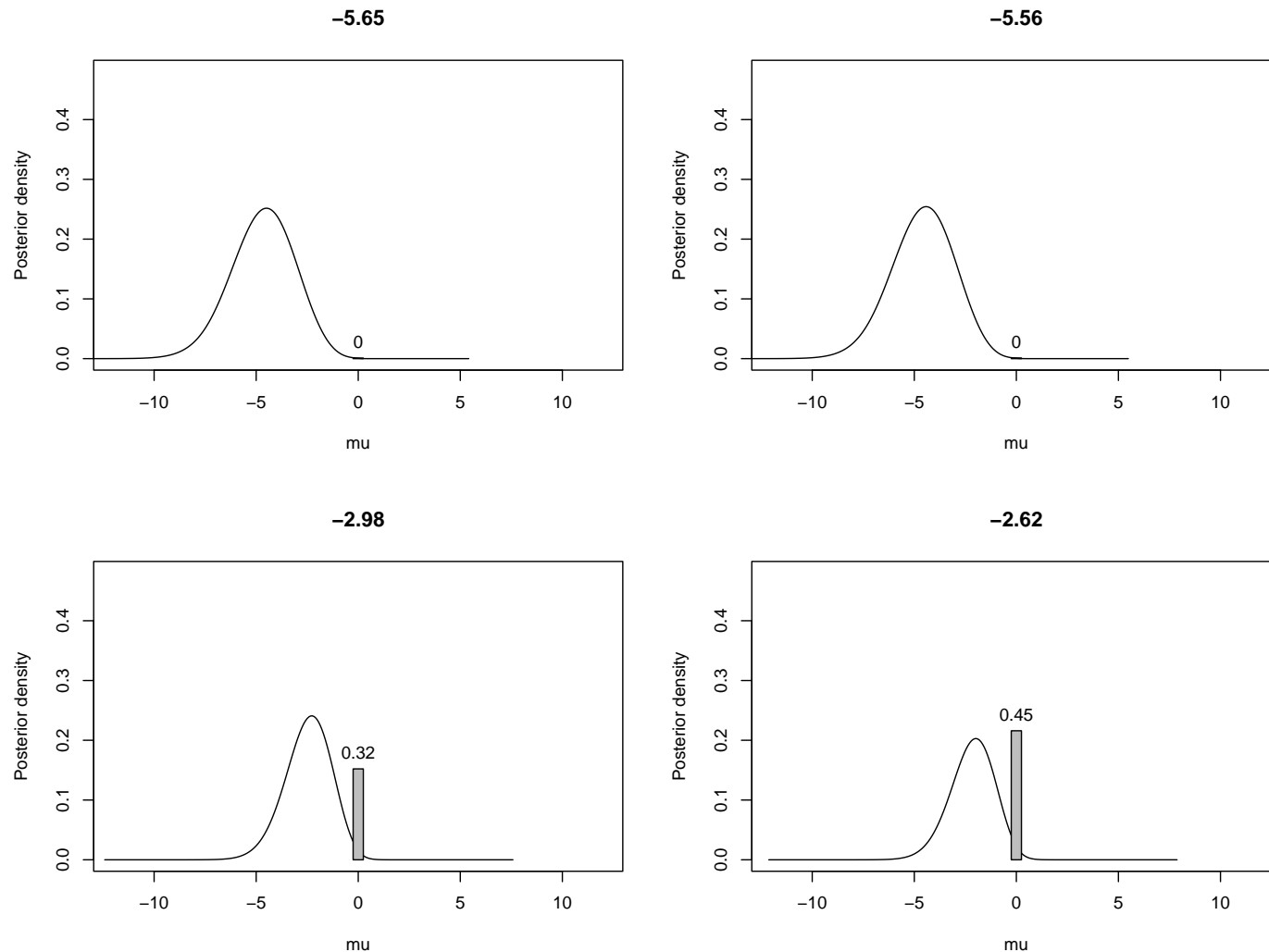
The prior on p_0 is uniform.

The penalty for multiple comparisons is automatic :-)

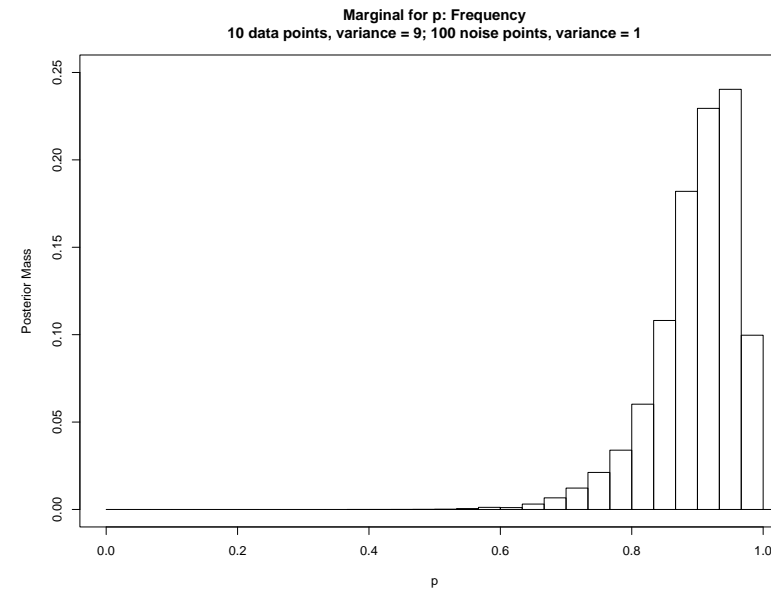
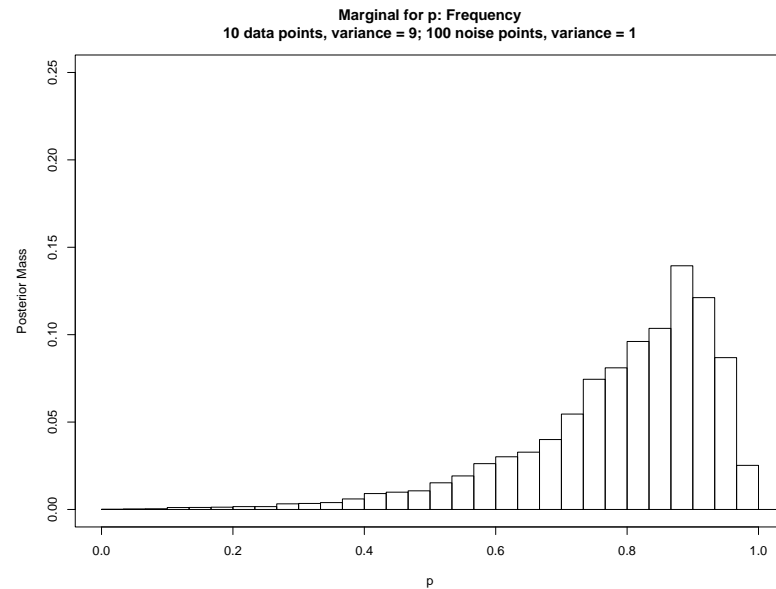
Below: π_i for the same 7 signal observations but with prior $\pi(p_0) = 11p_0^{10}$ ($p_0^{med} = .93$).

	Same 7 Signal Observations						
n	-5.56	-2.62	-1.20	-0.90	-0.15	1.94	3.57
25	0.90	0.37	0.10	0.08	0.06	0.20	0.65
100	0.98	0.23	0.06	0.05	0.04	0.11	0.58
500	1	0.26	0.04	0.03	0.02	0.10	0.74
5000	1	0.08	0.01	0.01	0.01	0.03	0.36

- The posterior **inclusion** probabilities $\pi_i = \Pr(\gamma_i = 1)$ or Prob. that the *i*th observation is a signal is, in multiple testing, one of the most interesting quantities
- Other quantities of interest:
 - Distribution of the μ_i if indeed it is not zero (strength of the 'signal')
 - Posterior distribution of p_0 , % of true nulls ('noise')
 - Posterior distribution of τ^2 , variance of the μ_i 's that are signals, and of σ^2 , variance of the observations
- Note, posterior probability of the γ 's (models) is of not direct interest

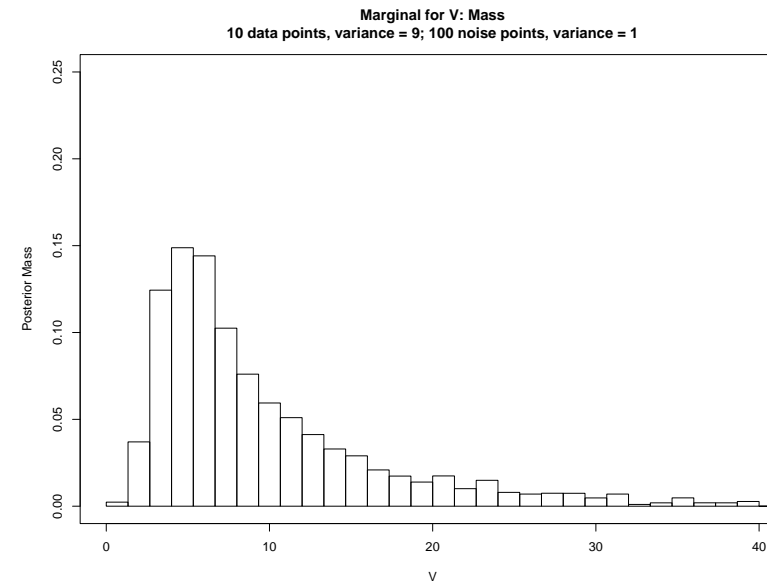
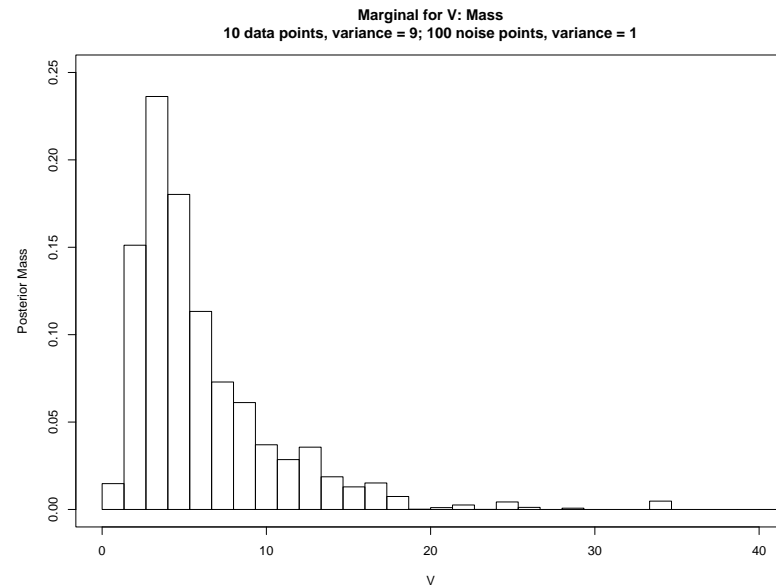


For four of the observations, $1 - \pi_i = \Pr(\mu_i = 0 | \boldsymbol{x})$ (the vertical bar), and posterior densities for $\mu_i \neq 0$.



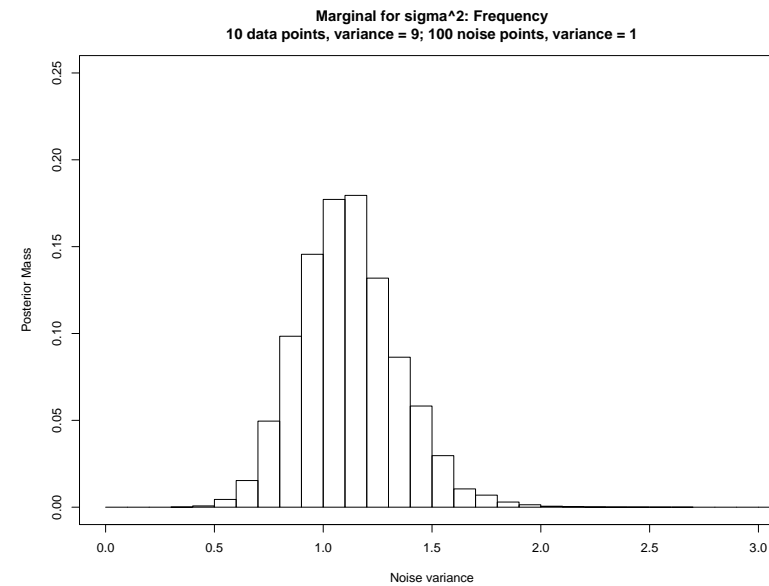
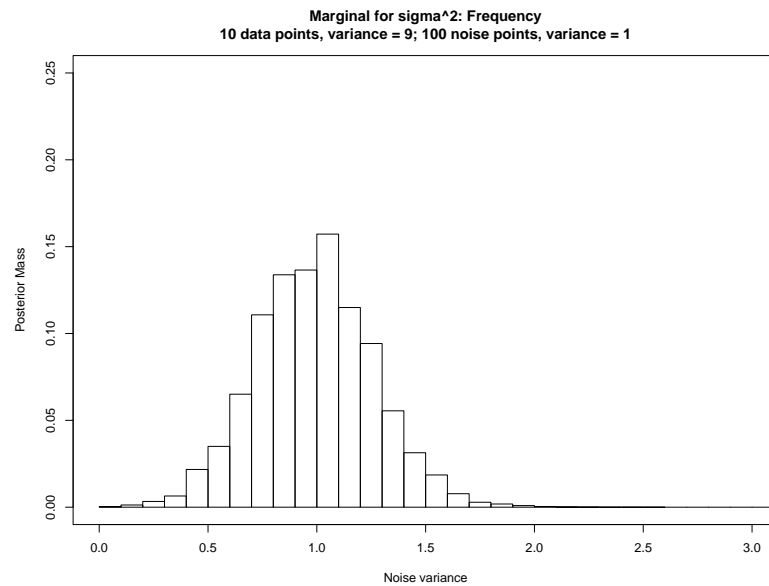
Marginal posteriors of p_0 for the case with 10 data points and 100 noise points. ($100/110 = .91$)

Left: $\pi(p_0) = 1$. Right: $\pi(p_0) = 11p_0^{10}$ (prior med. = 0.93)



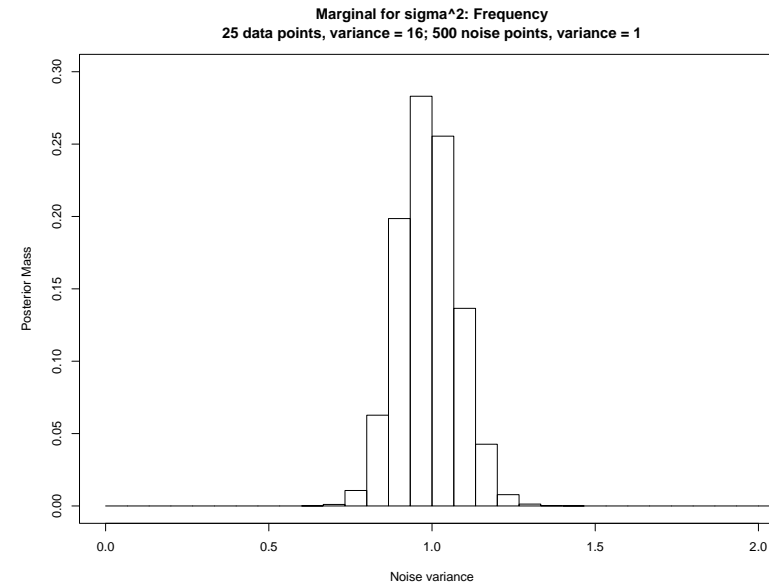
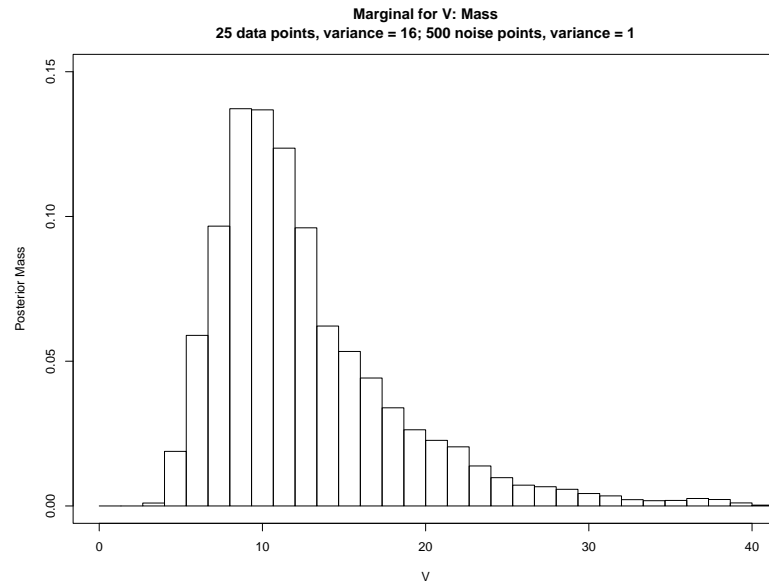
Marginal posteriors of τ^2 for the case with 10 data points and 100 noise points.

Left: $\pi(p_0) = 1$. Right: $\pi(p_0) = 11p_0^{10}$.



Marginal posteriors of σ^2 for the case with 10 data points and 100 noise points.

Left: $\pi(p_0) = 1$. Right: $\pi(p_0) = 11p_0^{10}$.



Marginal posteriors of τ^2 (left) and σ^2 (right) for the case with 25 signal means from the $N(0, 4^2)$ distribution and 500 noise points

Empirical Bayes (?) analyses

(Efron, Tibshirani, Genovese, Wasserman ...)

- Recall: posterior probability that the i -th null is true:

$$1 - \pi_i = \Pr(\gamma_i = 0 \mid x_i) = \frac{p_0 f_0(x_i)}{p_0 f_0(x_i) + (1 - p_0) f_1(x_i)}.$$

- Efron et al. (01) and Efron and Tibshirani (02) estimate f_0/f_1 nonparametrically.
- Efron et al. and G&W utilize a *conservative* 'estimate' of p_0 as the maximum value compatible with

$$f = p_0 f_0 + (1 - p_0) f_1 \geq p_0 f_0 \rightsquigarrow \hat{p}_0 = \min_x \frac{\hat{f}(x)}{\hat{f}_0(x)}.$$

(Do, Müller and Tang do a full nonparametric Bayes analysis.)

'Connections' (?) (empirical) Bayes – FDR

(Storey, Efron, Tibshirani, Genovese, Wasserman ...)

Storey argues that

- (i) the posterior probabilities of the hypotheses cannot be used directly, because they do not control for multiple comparisons (**false**, as we have seen)
- (ii) pFDR has a dual interpretation as a Bayesian and as a frequentist measure because

$$\text{pFDR}(C) = \Pr(\gamma_i = 0 \mid X_i \in C) = E[V]/E[R].$$

But this is the posterior probability given the data is in 'critical region' C , not given the data itself.

Bayesian FDR

Genovese and Wasserman (02); Newton et al. (04); Broët et al. (04)

- Recall that $\text{pFDR} = \Pr(H_0 \text{ true} \mid \text{reject } H_0)$
- For a 'Bayesian version' of pFDR, compute

$$1 - \pi_i = \Pr(H_0 \text{ true} \mid x_i)$$

and average over the rejection region

- to control pFDR at level α , reject if $\pi_i > p^*$ where

$$p^* = \inf \left\{ c : \frac{\sum_{i=1}^M I_{(\pi_i > c)} (1 - \pi_i)}{\sum_{i=1}^M I_{(\pi_i > c)}} \leq \alpha \right\}$$

- **But** FDR is taken as a priori being the quantity of interest. Is this reasonable from a Bayesian viewpoint?

Avoiding 'cheating' by Bayesian modeling

- Finner and Roters(01) remark that control of FDR allows 'cheating' "by adding additional hypotheses of no interest which are known to have p -values near 0" (Recall: the FDR critical value for $p_{(i)}$ is $i\alpha/M$).
- For instance, if one is interested in maximizing the chances of rejecting 8 hypotheses of interest while controlling FDR at $\alpha = 0.1$ one can add 100 'fake' hypothesis with p -values ≈ 0 , so that the 8 'interesting' p -values will have threshold $\geq 101\alpha/108 = .093$
- Bayesian modeling would expose any such 'cheating.'

Decision-theoretic Evaluations

	d_0	d_1	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

- Bickel (04) remarks that neither FDR nor pFDR have clear decision theoretical justifications (only asymptotically). He defends use of the proportion of false positives instead

$$\text{PFP} = \frac{E[V]}{E[R]}$$

and the corresponding ‘positive’ modification, which he calls *decisive* FDR.

- Several authors have considered decision-theoretic results and evaluations with different loss functions.

- Decision rules $\rightsquigarrow \mathbf{d} = \mathbf{d}(\mathbf{x}) = \{d_1, d_2, \dots, d_m\}$.
 $d_i = 1$ if i th-null is rejected, $d_i = 0$ otherwise.
- ‘Truth’ is represented by γ with $\gamma_i = 0$ indicating that the i th-null is true, and $\gamma_i = 1$ that it is false.
- Usual k_0, k_1 loss

	$d_i = 0$	$d_i = 1$
$\gamma_i = 0$	0	k_1
$\gamma_i = 1$	k_0	0

With the $k_0 - k_1$ loss,

- given \mathbf{x} , optimal Bayes decision rule has $d_i = 1$ iff

$$\Pr(\gamma_i = 1 \mid \mathbf{x}) = \pi_i > \frac{k_1}{k_0 + k_1}$$

	$d = 0$	$d = 1$	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

- for independent tests, the global loss is

$$L(\mathbf{d}, \gamma) = \sum_{i=1}^m L(d(x_i), \gamma_i) = k_1 V + k_0 T$$

- the posterior Bayes expected global loss:

$$E^{\gamma | \mathbf{x}} [L(\mathbf{d}, \gamma)]$$

is minimized for the previous individual thresholding

- the frequentist risk function is then

$$k_1 E^{\mathbf{X} | \gamma} [V] + k_0 E^{\mathbf{X} | \gamma} [T]$$

So it seems that only ER based on (functions of)

$E[V]$, $E[T]$ can have Decision theoretical justification.

Also bad behaviour:

	$d = 0$	$d = 1$	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

Müller, et. al. (2005) consider minimization of four global (posterior) expected losses

‘Univariate’ loss functions

- $L_N(\mathbf{d}) = cE^*[V] + E^*[T]$

Bayes rule with $c = k_1/k_0$

- $L_R(\mathbf{d}) = cE^*[FDR] + E^*[FNR]$

‘Bayes rule’ for loss function depending on the data
(suggested by Storey(03) and G&W(02))

	$d = 0$	$d = 1$	
H_0	U	V	M_0
H_1	T	S	M_1
	W	R	M

'Bivariate controlling' loss functions

- L_{2N} 'controls' ($E^*[T]$, $E^*[V]$)
minimize $E^*[T]$ subject to $E^*[V] \leq \alpha M$.
- L_{2R} 'controls' ($E^*[FNR]$, $E^*[FDR]$)
minimize $E^*[FNR]$ subject to $E^*[FDR] \leq \alpha$.

(This is G&W's proposal; it is maybe the most popular)

Their findings:

- All optimal rules are thresholding rules for π_i , all of them data-dependent except for the genuinely Bayesian L_N
- Pathological behavior of L_{2R} : Since $E^*[\text{FDR}]$ is 'controlled' as M grows, to achieve the desired (fixed) $E^*[\text{FDR}]$, "we have to knowingly flag some genes as differentially expressed even when $\pi_i \approx 0$ ".
- L_{2N} has a similar pathological behaviour (but slower)
- For L_N , $E^*[\text{FDR}]$ vanishes as $M \rightarrow \infty$
- The loss L_R induces counterintuitive jumps in $E^*[\text{FDR}]$ and is not recommended either

More realistic loss functions

(Duncan, 65; Waller and Duncan, 1969; Scott and Berger, 06)

- Separately specify the cost of a false positive and the cost of missing a true signal. Scott and Berger (06) use:

$$L(d_i = 1, \mu_i) = \begin{cases} 1 & \text{if } \mu_i = 0 \\ 0 & \text{if } \mu_i \neq 0, \end{cases}$$
$$L(d_i = 0, \mu_i) = \begin{cases} 0 & \text{if } \mu_i = 0 \\ c|\mu_i| & \text{if } \mu_i \neq 0, \end{cases}$$

where c is an adjustable parameter reflecting the relative costs of each type of error.

- The posterior expected loss is minimized by rejecting H_{0i} ($d_i = 1$, i.e. calling μ_i a signal) whenever

$$\pi_i > 1 - \frac{c \cdot \int_{-\infty}^{\infty} |\mu_i| \cdot \pi(\mu_i \mid \gamma_i = 1, \mathbf{x}) d\mu_i}{1 + c \cdot \int_{-\infty}^{\infty} |\mu_i| \cdot \pi(\mu_i \mid \gamma_i = 1, \mathbf{x}) d\mu_i}.$$

- This is also a thresholding rule for π_i
- The larger $E[|\mu_i| \mid \gamma_i = 1, \mathbf{x}]$, the smaller the threshold
- Many thousands of noise points would drive down the π_i 's for all observations. But posterior expectation of $|\mu_i|$ for an extreme observation could be large enough to result in rejection of H_{0i} even though odds *against* x_i being a signal are large.

Scott and Berger (2006) simulations

- Consider several combinations of c , # signals, # added noise, τ^2 , and priors for p_0 .
- Consider the previous example with # signals = 10. The signal points are $(-5.65, -5.56, -2.62, -1.20, -1.01, -0.90, -0.15, 1.65, 1.94, 3.57)$.
- Note that various of the signals are very weak (Noise is $N(0,1)$), and that the loss function does not care much about 'discovering' low signals
- $\mu_i \sim N(0, 3^2)$, and for p_0 consider two priors: the Uniform (U) and the beta $11p_0^{10}$ (Be)

Case		Performance		
# noise	α	Non-zero μ_i 's Discovered	False Positives	FDR
25	U	6 of 10	3	.33
25	Be	4 of 10	1	.20
500	U	4 of 10	7	.64
500	Be	3 of 10	3	.50
5000	U	3 of 10	4	.57
5000	Be	3 of 10	1	.25

$$c = 1$$

Case		Performance		
# noise	α	Non-zero μ_i 's Discovered	False Positives	FDR
25	U	4 of 10	0	0
25	Be	4 of 10	0	0
500	U	4 of 10	11	.73
500	Be	3 of 10	1	.25
5000	U	2 of 10	0	0
5000	Be	3 of 10	5	.625

$$c = 3$$

Consider now

- more signals: # signals = 25
- stronger signals: $\mu_i \sim N(0, 4^2)$
- The signal points are: (-7.93, -7.04, -4.46, -4.28, -3.85, -2.01, -1.81, -1.66, -1.62, -1.28, -0.57, 0.32, 0.74, 1.11, 1.82, 1.94, 2.23, 2.7, 2.82, 3.06, 3.62, 4.79, 5.48, 5.62, 8.12).

Case		Performance		
# noise	α	Non-zero μ_i 's Discovered	False Positives	FDR
50	U	13 of 25	1	.07
50	Be	10 of 25	0	0
500	U	11 of 25	1	.08
500	Be	13 of 25	1	.07
5000	U	9 of 25	1	.10
5000	Be	10 of 25	2	.17
10000	U	10 of 25	3	.23
10000	Be	8 of 25	0	0

$$c = 1$$

Case		Performance		
# noise	α	Non-zero μ_i 's Discovered	False Positives	FDR
50	U	16 of 25	3	.16
50	Be	14 of 25	1	.07
500	U	14 of 25	14	.50
500	Be	11 of 25	2	.15
5000	U	10 of 25	4	.29
5000	Be	10 of 25	4	.29
10000	U	10 of 25	7	.41
10000	Be	10 of 25	10	.50

$$c = 3$$

Conclusions

- Most likely, FDR and their relatives are here to stay, but they have to be introduced as a 'primitive': can not be justified on Bayesian reasoning nor on decision theoretical grounds.
- The probability that the nulls are true is a crucial ingredient in both Bayesian and FDR analyses:
 - in Bayesian analyses to provide the solution to the multiple comparisons problem;
 - in FDR analyses to increase 'power.'
- The Bayesian analyses are sensitive to the choice of prior distributions; determination of suitable objective-Bayes priors for the problem is still an issue.

References

- 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* **85**, 289–300.
- Benjamini, Y. and Hochberg, Y. (2000). The adaptive control of the false discovery rate in multiple hypothesis testing with independent statistics. *J. Educ. Behav. Statist.* **25**, 60–83
- Bickel, D.R. (2004). Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression? *Statistical Applications in Genetics and Molecular Biology* 3 (1) 8,
<http://www.bepress.com/sagmb/vol3/iss1/art8> (2004).
- Black, M.A. (2004). A Note on the adaptive control of the false discovery rates. *J.R. Statist. Soc. B* **66**, 297–304

- Broët, P., Lewin, A., Richardson, S., Dalmasso, C. and Magdelenat, H. (2004). A mixture model based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* **20** (16), 2562-2571.
- Do, K-A., Müller, P., and Tang, F. (2005). A Bayesian Mixture Model for Differential Gene Expression. *Applied Statistics* **54** (3), 627–644.
- Dudoit, S., Shaffer, J.P., and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.
- Duncan, D.B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7**, 171-222.
- Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96** 1151–1160
- Finner, H., and Roters, M. (2001). On the false discovery rate and expected Type I errors. *Biometrical Journal* **43**, 895–1005

Genovese, C.R. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society* **64**, 499–518.

Genovese, C.R. and Wasserman, L. (2003). Bayesian and frequentist multiple testing. In *Bayesian Statistics 7* (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West, eds.), 145–162. Oxford University Press.

Gönen, M., Westfall, P.H., Johnson, W.O. (2003). Bayesian multiple testing for two-sample multivariate endpoints. *Biometrics* **59**, 76–82.

House, L.L., Clyde, M.A., and Huang, Y.-C.T. (2006). Bayesian Analysis. Posted online July 29, 2005.
<http://ba.stat.cmu.edu/journal/forthcoming/house.pdf>

Langaas, M., Lindqvist, B.H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses with application to DNA microarray data. *J.R. Statist. Soc. B* **67**, 555–572.

- Müller, P., Parmigiani, G., Robert, C., and Rouseau, J. (2004). Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays. *Journal of the American Statistical Association* **99**, 990–1001.
- Newton, M.A., and Kendziorski, C.M. (2003). Parametric empirical Bayes methods for microarrays. In *The Analysis of Gene Expression Data: Methods and Software* (G. Parmigiani, , E.S. Garret, R.A. Irizarri, and S.L. Zeger, ed.), 254–271. Springer.
- Newton, M.A., Kendziorski, C.M., Richmon, C.S., Blattner, F.R., and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.
- Newton, M.A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5**, 155-176.

Scott, G. and Berger, J.O. (2006). An exploration of Bayesian multiple testing. *Journal of Statistical Planning and Inference* (in press).

Shaffer, J.P. (1995). Multiple hypothesis testing: a review. *Annual Review of Psychology* **46**, 561–584. Also *Technical Report # 23*, National Institute of Statistical Sciences.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

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

Storey J.D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035.

Stoto, M.A., Schonlau, M., and Mariano, L.T. (2004). Syndromic Surveillance: Is it worth the effort? *Chance* **17**, 19–24.

Waller, R.A. and Duncan, D.B. (1969). A Bayes rule for the symmetric multiple comparison problem. *Journal of the American Statistical Association* **64**, 1484--1503.

Yang, Y., and Rempala, G.A., (2004). A note on multiple tests for gene expression data. *Unpublished manuscript*