

Lecture 4: Multiplicity Control and Model Prior Probabilities

Jim Berger
Duke University

CBMS Conference on Model Uncertainty and Multiplicity
July 23-28, 2012

Outline

- Introduction to Bayesian multiplicity control
- Types of multiplicities and their Bayesian
- Variable selection
- Subgroup analysis
- Multiplicities in high-energy physics
- Comparison of Bayesian and frequentist multiplicity control

I. Introduction to Bayesian Multiplicity Control

Example of the multiplicity control issue:

In a recent talk about the drug discovery process, the following numbers were given in illustration.

- 10,000 relevant compounds were screened for biological activity.
- 500 passed the initial screen and were studied in vitro.
- 25 passed this screening and were studied in Phase I animal trials.
- 1 passed this screening and was studied in a Phase II human trial.

This could be nothing but noise, if screening was done based on ‘significance at the 0.05 level.’

If no compound had any effect,

- about $10,000 \times 0.05 = 500$ would initially be significant at the 0.05 level;
- about $500 \times 0.05 = 25$ of those would next be significant at the 0.05 level;
- about $25 \times 0.05 = 1.25$ of those would next be significant at the 0.05 level
- the 1 that went to Phase II would fail with probability 0.95.

Viewpoints on Bayesian multiplicity control

Subjective Bayesian viewpoint: Accurately model what you do not know and learn from the data; multiplicity then automatically takes care of itself.

Objective Bayesian viewpoint: Model what you do not know objectively and learn from the data; will multiplicity then take care of itself?

Statistical control for society: Prevent noise from becoming false knowledge/policy.

Fundamental question: How can we accommodate these viewpoints? More precisely, which subjective Bayesian and which objective Bayesian methods can society reliably use for multiplicity control.

Key Fact: Bayesian analysis deals with multiplicity adjustment solely through the assignment of prior probabilities to models or hypotheses

(see, e.g., Jeffreys 1961; Waller and Duncan 1969; Meng and Dempster 1987; Berry 1988; Westfall, Johnson and Utts 1997; Carlin and Louis 2000).

Introductory example where all is clear: multiple testing of exclusive hypotheses

Suppose one is testing mutually exclusive hypotheses H_i , $i = 1, \dots, m$, so that exactly one and only one of the H_i is true.

Bayesian analysis: If the hypotheses are viewed as exchangeable, choose $P(H_i) = 1/m$ and analyze the data \mathbf{x} .

- Let $m_i(\mathbf{x})$ denote the marginal density of the data under H_i . (The data density integrated over the prior density for unknown parameters under H_i .) This is often called the *likelihood* of H_i .
- The posterior probability of H_i is

$$Pr(H_i | \mathbf{x}) = \frac{m_i(\mathbf{x})}{\sum_{j=1}^m m_j(\mathbf{x})}.$$

- Thus the likelihood $m_i(\mathbf{x})$ for H_i is ‘penalized’ by a factor of $O(\frac{1}{m})$, resulting in multiplicity control.

Societal control: The above is often not fully adequate for society, as it assumes one of the H_i is true. If there is a possibility of ‘no effect’ one needs to augment the *prior* structure by **null control**, e.g.,

- $Pr(H_0) \equiv Pr(\text{no effect}) = 1/2$,
- $Pr(H_i) = 1/(2m)$.

Note: The assignment of $Pr(\text{no effect}) = 1/2$ is often generous.

Cute fact: If $Pr(H_i | \mathbf{x})$ is the posterior probability in a multiplicity scenario without null control, adding null control results in a new posterior probability

$$Pr^*(H_i | \mathbf{x}) \approx \frac{Pr(H_i | \mathbf{x})}{1 + Pr(H_0 | \mathbf{x})/Pr(H_0)}.$$

Subjectivity and societal control: If $Pr(\text{no effect}) = 1/2$ is assigned, society should allow **any** choice of $Pr(H_i)$. (But without null control, society should probably insist on exchangeable probability assignments.)

Example: 1000 energy channels are searched for the Higgs boson. In each, one observes $X_i \sim N(x_i \mid \mu_i, 1)$, and at most one of $H_i : \mu_i > 0$ is true.

Suppose $x_5 = 3$, and the other 999 of the X_i are standard normal variates.

- If testing in isolation $H_5^0 : \mu_5 = 0$ versus $H_5^1 : \mu_5 > 0$, with prior probabilities of $1/2$ each and a standard unit information Cauchy prior on μ_i under H_5^1 , then $Pr(H_5^1 \mid x_5 = 3) = \frac{m_5^1(3)}{m_5^1(3) + m_5^0(3)} = 0.96$.
- With multiplicity control, assigning $Pr(H_i) = 1/1000$, this becomes (on average over the 999 standard normal variates) $Pr(H_5 \mid \mathbf{x}) = \frac{m_5(\mathbf{x})}{\sum_{j=1}^{1000} m_j(\mathbf{x})} = 0.019$ (and 0.38 for $x_5 = 4$; and 0.94 for $x_5 = 5$)
- With null control in addition to multiplicity control, ($Pr(\text{no effect}) = 1/2$ and $Pr(H_i) = 1/(2000)$), this becomes $Pr(H_5 \mid \mathbf{x}) = 0.019$.
- If null control was employed but *a priori* the physicist decided to use all of the non-null mass on H_5 , the *societal* answer would have *legitimately* been $Pr(H_5 \mid \mathbf{x}) = 0.96$.

An aside: This is the Bayesian solution regardless of the structure of the data; in contrast, frequentist solutions depend on the structure of the data.

Example: For each channel, test $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i > 0$.

Data: $X_i, i = 1, \dots, m$, are $N(x_i | \mu_i, 1, \rho)$, ρ being the correlation.

If $\rho = 0$, one can just do individual tests at level α/m (Bonferroni) to obtain an overall error probability of α .

If $\rho > 0$, harder work is needed:

- Choose an overall decision rule, e.g., “declare channel i to have the signal if X_i is the largest value and $X_i > K$.”
- Compute the corresponding error probability, which can be shown to be

$$\alpha = \Pr(\max_i X_i > K \mid \mu_1 = \dots = \mu_m = 0) = E^Z \left[1 - \Phi \left(\frac{K - \sqrt{\rho}Z}{\sqrt{1 - \rho}} \right)^m \right],$$

where Φ is the standard normal cdf and Z is standard normal.

Note that this gives (essentially) the Bonferroni correction when $\rho = 0$, and converges to $1 - \Phi[K]$ as $\rho \rightarrow 1$ (the one-dimensional solution).

Interim summary

Two types of Bayesian control:

- Bayesian null control gives prior probability (at least $1/2$) to ‘nothing happening.’
- Bayesian multiplicity control arises from dividing up the remaining probability among the alternative explanations (models); the more models, the less weight each gets.

Three advantages of Bayesian control:

- It is *possible* for society to allow *apriori* subjective (scientific) choices of probability assignments to various alternative hypotheses/models, while maintaining societal control.
- The Bayesian approach to multiplicity/null control does not depend on the error structure of the data.
- Objective multiplicity/null control mechanisms can be applied retroactively.

Do objective Bayes probability assignments automatically provide multiplicity or null control?

- Suppose $x_i \sim N(x_i \mid \mu_i, 1)$, $i = 1, \dots, m$, are observed.
- It is desired to test $H_i^0 : \mu_i = 0$ versus $H_i^1 : \mu_i \neq 0$, $i = 1, \dots, m$, but any test could be true or false regardless of the others.
- The simplest objective probability assignment is $Pr(H_i^0) = Pr(H_i^1) = 0.5$, independently, for all i .
- This does *not* control for multiplicity; indeed, each test is then done completely independently of the others.
- This does *not* have overall null control, because $Pr(\mu_1 = \mu_2 = \dots = \mu_m = 0) = 2^{-m}$.

Note: The same holds in any model selection problem such as variable selection: use of equal probabilities for all models does not induce any multiplicity adjustment or overall null control.

Inducing null control:

- Reformulate as a model selection problem, defining models \mathcal{M}_γ , where $\gamma = (\gamma_1, \dots, \gamma_m)$, with γ_i being 0 or 1 as $H_i^0 : \mu_i = 0$ or $H_i^1 : \mu_i \neq 0$ is true. (*An aside:* this is usually also the best way to compute.)
- Set $Pr(\mathcal{M}_\mathbf{0}) = Pr(\mu_1 = \mu_2 = \dots = \mu_m = 0) = 1/2$, and $Pr(\mathcal{M}_\gamma) = \frac{1}{2(2^m - 1)}$ otherwise.
- The previous unadjusted $Pr(H_i^1 | x_i)$ then becomes (using the cute fact)

$$Pr^*(H_i^1 | \mathbf{x}) \approx \frac{Pr(H_i^1 | x_i)}{1 + Pr(\mathcal{M}_\mathbf{0} | \mathbf{x})/Pr(\mathcal{M}_\mathbf{0})}.$$

- *Case 1. $\mathcal{M}_\mathbf{0}$ is true:* Then $Pr(\mathcal{M}_\mathbf{0} | \mathbf{x})/Pr(\mathcal{M}_\mathbf{0}) = O(e^{cm})$ for some c , so there is very strong null (and multiplicity) control.
- *Case 2. All of the null hypotheses are true, except one, which has a very large x_j :* Then $Pr(\mathcal{M}_\mathbf{0} | \mathbf{x})/Pr(\mathcal{M}_\mathbf{0}) \approx 0$, and there is no multiplicity control.

Inducing only multiplicity control (Scott and Berger, 2006 JSPI; other, more sophisticated full Bayesian analyses are in Gönen et. al. (03), Do, Müller, and Tang (02), Newton et al. (01), Newton and Kendzioriski (03), Müller et al. (03), Guindani, M., Zhang, S. and Mueller, P.M. (2007), ...; many empirical Bayes such as Storey, J.D., Dai, J.Y and Leek, J.T. (2007))

- Suppose $x_i \sim N(x_i | \mu_i, \sigma^2)$, $i = 1, \dots, m$, are observed, σ^2 known, and test $H_i^0 : \mu_i = 0$ versus $H_i^1 : \mu_i \neq 0$.
- If the hypotheses are viewed as exchangeable, let p denote the common prior probability of H_i^1 , and *assume p is unknown* with a uniform prior distribution. *This does provide multiplicity control.*
- Complete the prior specification, e.g.
 - Assume that the nonzero μ_i follow a $N(0, V)$ distribution, with V unknown.
 - Assign V the objective (proper) prior density $\pi(V) = \sigma^2 / (\sigma^2 + V)^2$.

- Then the posterior probability that $\mu_i \neq 0$ is

$$p_i = 1 - \frac{\int_0^1 \int_0^1 p \prod_{j \neq i} \left(p + (1-p)\sqrt{1-w} e^{wx_j^2/(2\sigma^2)} \right) dpdw}{\int_0^1 \int_0^1 \prod_{j=1}^m \left(p + (1-p)\sqrt{1-w} e^{wx_j^2/(2\sigma^2)} \right) dpdw}.$$

- (p_1, p_2, \dots, p_m) can be computed numerically; for large m , it is most efficient to use importance sampling, with a common importance sample for all p_i .

Example: Consider the following ten ‘signal’ observations:

-8.48, -5.43, -4.81, -2.64, -2.40, 3.32, 4.07, 4.81, 5.81, 6.24

- Generate $n = 10, 50, 500,$ and 5000 $N(0, 1)$ noise observations.
- Mix them together and try to identify the signals.

n	The ten 'signal' observations										#noise
	-8.5	-5.4	-4.8	-2.6	-2.4	3.3	4.1	4.8	5.8	6.2	$p_i > .6$
10	1	1	1	.94	.89	.99	1	1	1	1	1
50	1	1	1	.71	.59	.94	1	1	1	1	0
500	1	1	1	.26	.17	.67	.96	1	1	1	2
5000	1	1.0	.98	.03	.02	.16	.67	.98	1	1	1

Table 1: The posterior probabilities of being nonzero for the ten 'signal' means.

Note 1: The penalty for multiple comparisons is automatic.

Note 2: Theorem: $E[\#i : p_i > .6 \mid \text{all } \mu_j = 0] = O(1)$ as $m \rightarrow \infty$, so the Bayesian procedure exerts medium-strong control over false positives. (In comparison, $E[\#i : \text{Bonferroni rejects} \mid \text{all } \mu_j = 0] = \alpha$.)

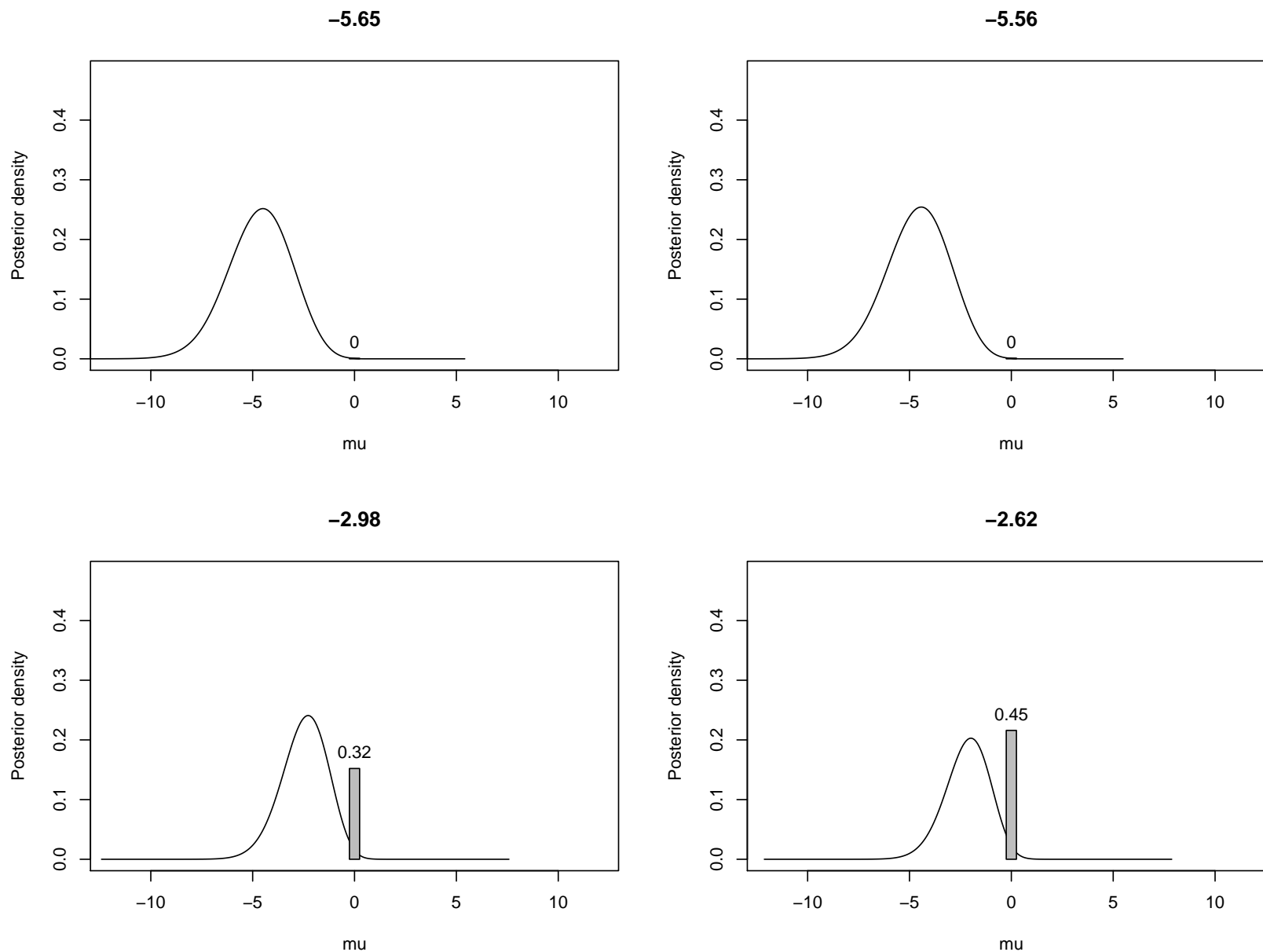


Figure 1: For four of the observations, $1 - p_i = \Pr(\mu_i = 0 | \mathbf{y})$ (the vertical bar), and the posterior densities for $\mu_i \neq 0$.

An Aside: Use for Discoveries

- p_i gives the probability that i is a discovery.
- The posterior density for $\mu_i \neq 0$ gives the magnitude of the effect of the possible discovery.
- If claiming J discoveries, with probabilities p_i ; the probability that *all* are discoveries can be computed from the posterior. (If approximate independence, $\prod_i p_i$.)
- If a discovery is claimed if $p_i > c$, the expected false discovery rate (Bayesian) is

$$\frac{\sum_{\{i:p_i > c\}} (1 - p_i)}{\#\{i : p_i > c\}} < 1 - c.$$

Use for Screening (Duncan, 65; Waller and Duncan, 1969)

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

$$L(\text{reject null}, \mu_i) = \begin{cases} 1 & \text{if } \mu_i = 0 \\ 0 & \text{if } \mu_i \neq 0, \end{cases}$$

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

where c reflects the relative costs of each type of error.

- Posterior expected loss is minimized by rejecting H_{0i} when

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

Null control and subjective multiplicity control:

- An easier formulation of null control is to say p is 0 with probability $1/2$ and has a Uniform prior with probability $1/2$.
 - Adding null control eliminates the $O(1)$ ‘false positives’ in the Bayesian analysis.
 - If there are true $H_i^1 : \mu_i \neq 0$, adding null control will typically not affect the answers (because of the cute fact).
- If scientific knowledge is available about p , one can encode it in a $Beta(a, b)$ prior. Which priors allow societal control?
 - Any prior more concentrated near 0 than the uniform is clearly allowed.
 - If all agree on exchangeability and null control is present, one can argue that society should allow any prior for p .
- If the $H_i^1 : \mu_i \neq 0$ are not apriori exchangeable, much more complicated assignments of prior probabilities can be considered.

Interim summary 2

- Objective Bayesian probability assignments can fail to provide either multiplicity control or null control. In particular, assigning all models equal prior probability fails.
- A key technique for multiplicity control is to specify *unknown* inclusion probabilities for hypotheses or variables, and assign them a prior distribution.
 - Giving the inclusion probability probability $1/2$ of being zero, also provides null control with little cost in power.
 - Some subjective priors for inclusion probabilities satisfy societal control.

II. Types of Multiplicities

- **Class 1:** Multiplicities not affecting the likelihood function
 - Consideration of multiple (test) statistics or multiple priors
 - Interim or sequential analysis
 - Some issues with multiple endpoints (not discussed)
- **Class 2:** Multiplicities affecting the likelihood function
 - Choice of transformation/model
 - Multiple testing
 - Variable selection
 - Subgroup analysis
- **Class 3:** Issues arising from multiple studies of the same situation: meta-analysis, replication, ... (not discussed).

Class 1: Multiplicities Not Affecting the Likelihood

- Consideration of multiple test statistics
 - *Example:* Doing a test of fit, and trying both a Kolmogorov-Smirnov test and a Chi-squared test.
 - Frequentists should either report all tests, or adjust; e.g., if p_i is the p-value of test i , base testing on the *statistic* $p_{min} = \min p_i$.
- Consideration of multiple priors: a Bayesian must either
 - have principled reasons for settling on a particular prior, or
 - implement a hierarchical or robustness analysis over the priors.
- Interim or sequential analysis
 - Bayesians do not adjust, as the posterior is unaffected.
 - Frequentists often adjust: ‘spend α ’ for interim options.
(But conditional frequentists do not necessarily adjust.)
- Some issues with multiple endpoints (not discussed).

Class 2: Multiplicities Affecting the Likelihood

- Choice of transformation/model
- Multiple testing
- Variable selection (later section)
- Subgroup analysis (later section)

Choice of transformation/model

- Frequentist solutions:
 - Develop the model on part of the data; perform inference on the other part (or do a new experiment).
 - Formal solutions: confidence set after testing, bootstrap
 - *This is often ignored, leading to overconfident inference.*
- Bayesian solution: model averaging.
 - Assign each model/transformation a prior probability.
 - Compute model/transformation posterior probabilities.
 - Perform inference with weighted averages over the models/transformations. (An overwhelmingly supported model/transformation will receive weight near one.)
 - *This is often ignored, leading to overconfident inference.*

Bayesian Solution: Model Averaging

- Assign probabilities $P(M_i)$ to models; the more models (multiplicities being considered), the less prior probability each model receives.
- Compute the posterior model probabilities $P(M_i | data)$
- If, say, inference concerning ξ is desired, it would be based on

$$\pi(\xi | data) = \sum_{i=1}^q P(M_i | data) \pi(\xi | data, M_i).$$

Note: ξ must have the same meaning across models, as in prediction.

Example: From i.i.d. vehicle emission data $\mathbf{X} = (X_1, \dots, X_n)$, one desires to determine the probability that the vehicle type will meet regulatory standards.

Traditional models for this type of data are Weibull and lognormal distributions given, respectively, by

$$M_1 : f_W(x; \beta, \gamma) = \frac{\gamma}{\beta} \left(\frac{x}{\beta}\right)^{\gamma-1} \exp \left[-\left(\frac{x}{\beta}\right)^\gamma \right]$$
$$M_2 : f_L(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(\log x - \mu)^2}{2\sigma^2} \right].$$

Note that both distributions are in the location-scale family (the Weibull being so after a log transformation).

Model Averaging Analysis:

- Assign each model prior probability $1/2$.
- Because of the common location-scale invariance structures, assign the right-Haar prior densities $\pi_W(\beta, \gamma) = 1/(\beta\gamma)$ and $\pi_L(\mu, \sigma) = 1/(\sigma)$, respectively (Berger, Pericchi and Varshavsky, 1998 Sankhyā).
- The posterior probabilities (and conditional frequentist error probabilities) of the two models are then

$$P(M_1 | \mathbf{x}) = 1 - P(M_2 | \mathbf{x}) = \frac{B(\mathbf{x})}{1 + B(\mathbf{x})},$$

where $z_i = \log x_i$, $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$, $s_z^2 = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2$, and

$$B(\mathbf{x}) = \frac{\Gamma(n)n^n \pi^{(n-1)/2}}{\Gamma(n-1/2)} \int_0^\infty \left[\frac{y}{n} \sum_{i=1}^n \exp\left(\frac{z_i - \bar{z}}{s_z y}\right) \right]^{-n} dy.$$

- For the studied data set, $P(M_1 | \mathbf{x}) = .712$. Hence,

$$\begin{aligned} P(\text{meeting standard}) &= .712 P(\text{meeting standard} | M_1) \\ &\quad + .288 P(\text{meeting standard} | M_2). \end{aligned}$$

Multiple testing

- Multiple hypothesis testing (earlier Bayesian analysis)
- Multiple multiple testing
 - e.g., plasma samples are sent to separate genomic, protein, and metabolic labs for ‘discovery’.
- Serial studies
 - the first three HIV vaccine trials failed
 - all 16 large Phase III Alzheimer’s trials have failed

Multiple multiple testing:

Example: Plasma samples are sent to the following labs in a pharmaceutical company:

- a metabolic lab, where an association is sought with any of 200 metabolites;
- a proteomic lab, where an association is sought with any of 2000 proteins;
- a genomics lab, where an association is sought with any of 2,000,000 genes.

The company should do a joint multiplicity analysis.

A Bayesian analysis could give each lab $1/3$ of the prior probability of a discovery, with each third to be divided within the lab.

Serial testing: In the vaccine example, there were two previous failed trials. Should the third vaccine trial have a multiplicity adjustment? The (exchangeable) Bayesian solution:

- assign each trial common unknown probability p of success, with p having a uniform distribution, and compute the posterior probability that the current trial exhibits no efficacy

$$Pr(H_0 | x_1, x_2, x_3) = \left(1 + \frac{B_{01}(x_1)B_{01}(x_2) + B_{01}(x_1) + B_{01}(x_2) + 3}{3B_{01}(x_1)B_{01}(x_2) + B_{01}(x_1) + B_{01}(x_2) + 1} \times \frac{1}{B_{01}(x_3)} \right)^{-1}$$

where $B_{01}(x_i)$ is the Bayes factor of “no effect” to “effect” for trial i .

This changes the previous $Pr(H_0 | x_3) = 0.20$ to $Pr(H_0 | x_1, x_2, x_3) = 0.29$.

Example: There have been 16 large Phase III Alzheimer’s trials - all failing. (The probability of that is only 0.44.). One cannot do the Bayesian serial testing adjustment should the 17th trial succeed, without knowing the Bayes factors in each of the failed trials. But it could be as severe as

$$B_{01}^{adj} = 16 \times B_{01}(x_{17}).$$

III. Variable Selection

Example: a retrospective study of a data-base investigates the relationship between 200 foods and 25 health conditions. It is reported that eating broccoli reduces lung cancer (p -value=0.02).



- Not adjusting for multiplicity (5000 tests) in this type of situation is a leading cause of ‘junk science.’
- There are other contributing problems here, such as the use of p -values.

Frequentist solutions:

- Bonferonni could be used: to achieve an overall level of 0.05 with 5000 tests, one would need to use a per-test rejection level of $\alpha = 0.05/5000 = 0.00001$.
 - This is likely much too conservative because of the probably high dependence in the 5000 tests.
- Some type of bootstrap could be used, but this is difficult when faced, as here, with 2^{5000} models.

Bayesian solution:

- Assign prior variable inclusion probabilities.
- Implement Bayesian model averaging or variable selection.

- Options in choosing prior variable inclusion probabilities:
 - Objective Bayesian choices:
 - * *Option 1*: each variable has unknown common probability p_i of having no effect on health condition i .
 - * *Option 2*: variable j has common probability p_j of having no effect on each health condition.
 - * *Option 3*: some combination.
 - Main effects may have a common unknown prior inclusion probability p_1 ; second order interactions prior inclusion probability p_2 ; etc.
 - An oversight committee for a prospective study might judge that at most one effect might be found, and so could prescribe that a protocol be submitted in which
 - * prior probability $1/2$ be assigned to ‘no effect;’
 - * the remaining probability of $1/2$ could be divided among possible effects as desired pre-experimentally. (Bonferonni adjustments can also be unequally divided pre-experimentally.)

Example: Genome-wide Association Studies (GWAS)

- Early genomic epidemiological studies almost universally failed to replicate (estimates of the replication rate are as low as 1%), because they were doing multiple testing at ‘ordinary p-values’.
- A very influential paper in Nature (2007), by the Wellcome Trust Case Control Consortium, argues for a cutoff of $p < 5 \times 10^{-7}$ ($-e p \log(p) = 2.0 \times 10^{-5}$). (Derived from a Bayesian argument, with prior odds of an association set at 1/100,000.)
 - Found 21 genome/disease associations; all but one have been replicated.
 - Later studies in GWAS have recommended cutoffs as low as 5×10^{-8} ($-e p \log(p) = 2.3 \times 10^{-6}$).

Bayes argument for the cutoff:

- Let π_0 and $\pi_1 = 1 - \pi_0$ be the prior probabilities at a given location on the genome of not having or having an association, respectively.
- Let α and $(1 - \beta(\theta))$ be the Type I error and power for testing the null hypothesis of no association with a given rejection region \mathcal{R} .
- The pre-experimental probability of a false positive is then $\pi_0\alpha$.
- The pre-experimental probability of a true positive is then $\pi_1(1 - \bar{\beta})$, where $(1 - \bar{\beta}) = \int(1 - \beta(\theta))\pi(\theta)d\theta$ is average power wrt the prior $\pi(\theta)$.
- Pre-experimental ‘odds of true positive to false positive’ = $\frac{\pi_1}{\pi_0} \times \frac{(1 - \bar{\beta})}{\alpha}$.
- For the GWAS study,
 - they choose $\frac{\pi_1}{\pi_0} = \frac{1}{100,000}$ and $(1 - \bar{\beta}) = 0.5$, one achieves odds of 10 : 1 in favor of a true positive if $\alpha = 5 \times 10^{-7}$. (Could vary by a factor of 10.)
 - The article also argued that it is better to just compute the Bayes factors B_{10} , and the posterior odds = $\frac{\pi_1}{\pi_0} \times B_{10}$. These ranged between $\frac{1}{10}$ and 10^{68} for the 21 claimed associations.

Formal Bayesian Approach to Multiplicity Control in Variable Selection

Problem: Data \mathbf{X} arises from a normal linear regression model, with m possible regressors having associated unknown regression coefficients $\beta_i, i = 1, \dots, m$, and unknown variance σ^2 .

Models: Consider selection from among the submodels $\mathcal{M}_i, i = 1, \dots, 2^m$, having only k_i regressors with coefficients β_i (a subset of $(\beta_1, \dots, \beta_m)$) and resulting density $f_i(\mathbf{x} | \beta_i, \sigma^2)$.

Prior density under \mathcal{M}_i : Zellner-Siow priors $\pi_i(\beta_i, \sigma^2)$.

Marginal likelihood of \mathcal{M}_i : $m_i(\mathbf{x}) = \int f_i(\mathbf{x} | \beta_i, \sigma^2) \pi_i(\beta_i, \sigma^2) d\beta_i d\sigma^2$

Prior probability of \mathcal{M}_i : $P(\mathcal{M}_i)$

Posterior probability of \mathcal{M}_i :

$$P(\mathcal{M}_i | \mathbf{x}) = \frac{P(\mathcal{M}_i) m_i(\mathbf{x})}{\sum_j P(\mathcal{M}_j) m_j(\mathbf{x})}.$$

Common Choices of the $P(\mathcal{M}_i)$

Equal prior probabilities: $P(\mathcal{M}_i) = 2^{-m}$ does not control for multiplicity.

Bayes exchangeable variable inclusion does control for multiplicity:

- Each variable, β_i , is independently in the model with unknown probability p (called the prior inclusion probability).
- p has a Beta($p \mid a, b$) distribution. (We use $a = b = 1$, the uniform distribution, as did Jeffreys 1961.)
- Then, since k_i is the number of variables in model \mathcal{M}_i ,

$$P(\mathcal{M}_i) = \int_0^1 p^{k_i} (1 - p)^{m - k_i} \text{Beta}(p \mid a, b) dp = \frac{\text{Beta}(a + k_i, b + m - k_i)}{\text{Beta}(a, b)}.$$

Note that this can be pre-computed; no uncertainty analysis (e.g. MCMC) in p is needed! (See Scott and Berger, 2008, for discussion.)

Empirical Bayes variable inclusion does control for multiplicities: Find the MLE \hat{p} by maximizing the marginal likelihood of p , $\sum_j p^{k_j} (1 - p)^{m - k_j} m_j(\mathbf{x})$, and use $P(\mathcal{M}_i) = \hat{p}^{k_i} (1 - \hat{p})^{m - k_i}$ as the prior model probabilities.

	Equal model probabilities				Bayes variable inclusion			
	Number of noise variables				Number of noise variables			
Signal	1	10	40	90	1	10	40	90
$\beta_1 : -1.08$.999	.999	.999	.999	.999	.999	.999	.999
$\beta_2 : -0.84$.999	.999	.999	.999	.999	.999	.999	.988
$\beta_3 : -0.74$.999	.999	.999	.999	.999	.999	.999	.998
$\beta_4 : -0.51$.977	.977	.999	.999	.991	.948	.710	.345
$\beta_5 : -0.30$.292	.289	.288	.127	.552	.248	.041	.008
$\beta_6 : +0.07$.259	.286	.055	.008	.519	.251	.039	.011
$\beta_7 : +0.18$.219	.248	.244	.275	.455	.216	.033	.009
$\beta_8 : +0.35$.773	.771	.994	.999	.896	.686	.307	.057
$\beta_9 : +0.41$.927	.912	.999	.999	.969	.861	.567	.222
$\beta_{10} : +0.63$.995	.995	.999	.999	.996	.990	.921	.734
False Positives	0	2	5	10	0	1	0	0

Table 2: Posterior inclusion probabilities for 10 real variables in a simulated data set.

Comparison of Bayes and Empirical Bayes Approaches

Theorem 1 *In the variable-selection problem, if the null model (or full model) has the largest marginal likelihood, $m(\mathbf{x})$, among all models, then the MLE of p is $\hat{p} = 0$ (or $\hat{p} = 1$.) (The naive EB approach, which assigns $P(\mathcal{M}_i) = \hat{p}^{k_i} (1 - \hat{p})^{m - k_i}$, concludes that the null (full) model has probability 1.)*

A simulation with 10,000 repetitions to gauge the severity of the problem:

- $m = 14$ covariates, orthogonal design matrix
- p drawn from $U(0, 1)$; regression coefficients are 0 with probability p and drawn from a Zellner-Siow prior with probability $(1 - p)$.
- $n = 16, 60,$ and 120 observations drawn from the given regression model.

Case	$\hat{p} = 0$	$\hat{p} = 1$
$n = 16$	820	781
$n = 60$	783	766
$n = 120$	723	747

Covariate	Fully Bayes	Emp. Bayes
East Asian Dummy	0.983	0.983
Fraction of Tropical Area	0.727	0.653
Life Expectancy in 1960	0.624	0.499
Population Density Coastal in 1960s	0.518	0.379
GDP in 1960 (log)	0.497	0.313
Outward Orientation	0.417	0.318
Fraction GDP in Mining	0.389	0.235
Land Area	0.317	0.121
Higher Education 1960	0.297	0.148
Investment Price	0.226	0.130
Fraction Confucian	0.216	0.145
Latin American Dummy	0.189	0.108
Ethnolinguistic Fractionalization	0.188	0.117
Political Rights	0.188	0.081
Primary Schooling in 1960	0.167	0.093
Hydrocarbon Deposits in 1993	0.165	0.093
Fraction Spent in War 1960–90	0.164	0.095
Defense Spending Share	0.156	0.085
Civil Liberties	0.154	0.075
Average Inflation 1960–90	0.150	0.064
Real Exchange Rate Distortions	0.146	0.071
Interior Density	0.139	0.067

Table 3: Exact variable inclusion probabilities for 22 variables in a linear model for GDP growth among a group of 30 countries.

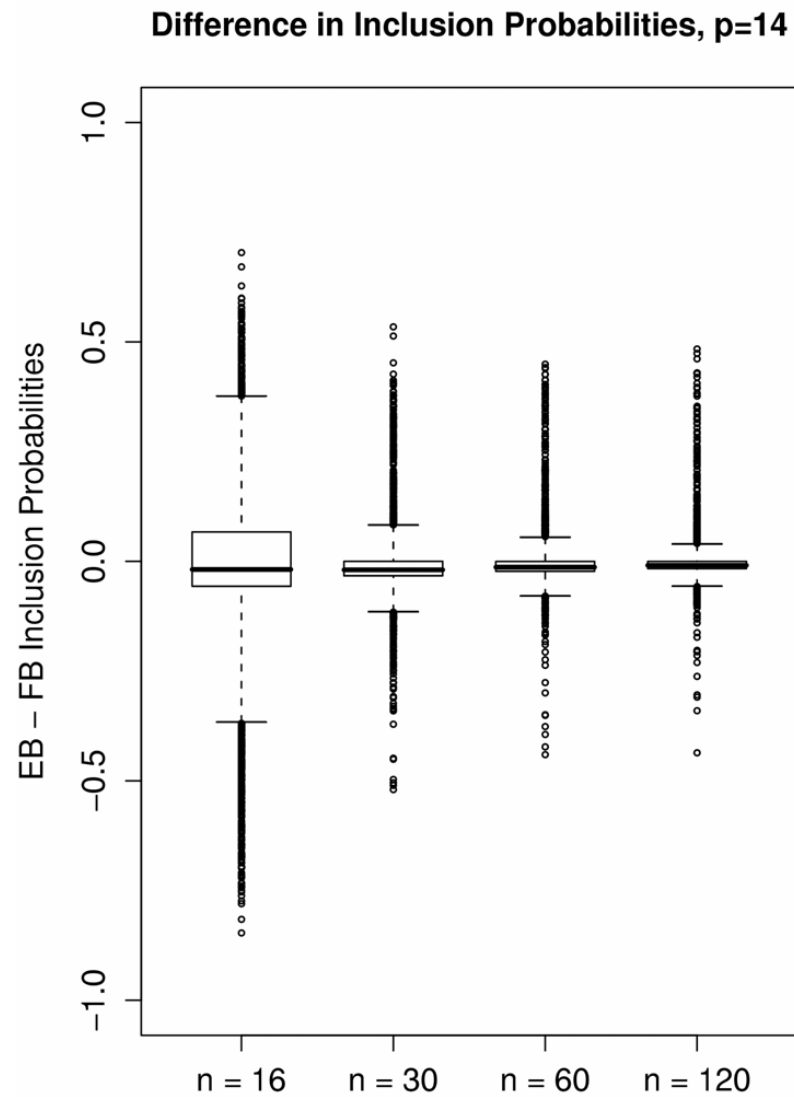


Figure 2: Empirical distribution of difference in inclusion probabilities between EB and FB, 10000 fake data sets with 14 possible covariates in each one, everything drawn from the prior.

Is empirical Bayes at least accurate asymptotically as $m \rightarrow \infty$?

Posterior model probabilities, given p :

$$P(\mathcal{M}_i | \mathbf{x}, p) = \frac{p^{k_i} (1-p)^{m-k_i} m_i(\mathbf{x})}{\sum_j p^{k_j} (1-p)^{m-k_j} m_j(\mathbf{x})}$$

Posterior distribution of p : $\pi(p | \mathbf{x}) = K \sum_j p^{k_j} (1-p)^{m-k_j} m_j(\mathbf{x})$

This *does* concentrate about the true p as $m \rightarrow \infty$, so one might expect that

$$P(\mathcal{M}_i | \mathbf{x}) = \int_0^1 P(\mathcal{M}_i | \mathbf{x}, p) \pi(p | \mathbf{x}) dp \approx P(\mathcal{M}_i | \mathbf{x}, \hat{p}) \propto m_i(\mathbf{x}) \hat{p}^{k_i} (1-\hat{p})^{m-k_i}.$$

This is not necessarily true; indeed

$$\begin{aligned} \int_0^1 P(\mathcal{M}_i | \mathbf{x}, p) \pi(p | \mathbf{x}) dp &= \int_0^1 \frac{p^{k_i} (1-p)^{m-k_i} m_i(\mathbf{x})}{\pi(p | \mathbf{x}) / K} \times \pi(p | \mathbf{x}) dp \\ &\propto m_i(\mathbf{x}) \int_0^1 p^{k_i} (1-p)^{m-k_i} dp \propto m_i(\mathbf{x}) P(\mathcal{M}_i). \end{aligned}$$

Caveat: Some EB techniques have been justified; see Efron and Tibshirani (2001), Johnstone and Silverman (2004), Cui and George (2006), and Bogdan et. al. (2008).

Theorem 2 *Suppose the true model size k_T satisfies $k_T/m \rightarrow p_T$ as $m \rightarrow \infty$, where $0 < p_T < 1$. Consider all models M_i such that $k_T - k_i = O(\sqrt{m})$, and consider the optimal situation for EB in which*

$$\hat{p} = p_T + O\left(\frac{1}{\sqrt{m}}\right) \quad \text{as } m \rightarrow \infty.$$

Then the ratio of the prior probabilities assigned to such models by the Bayes approach and the empirical Bayes approach satisfies

$$\frac{P_B(\mathcal{M}_i)}{P_{EB}(\mathcal{M}_i)} = \frac{\int_0^1 p^{k_i} (1-p)^{m-k_i} \pi(p) dp}{(\hat{p})^{k_i} (1-\hat{p})^{m-k_i}} = O\left(\frac{1}{\sqrt{m}}\right),$$

providing $\pi(\cdot)$ is continuous and nonzero.

IV. Subgroup Analysis

Subgroup Analysis (work in progress)

Our guiding principles:

- Null control and multiplicity control need to be present.
- To maximize power to detect real effects,
 - the subgroups and allowed population partitions need to be restricted to those that are scientifically plausible;
 - allowance for ‘scientifically favored’ subgroups should be made.
- Full Bayesian analysis is sought. In particular, any inferences should have an interpretation in terms of the actual population.

Allowable subgroups: Subgroups are specified by criteria, denoted by letters. For example, age is A , gender is B , and smoking status is C .

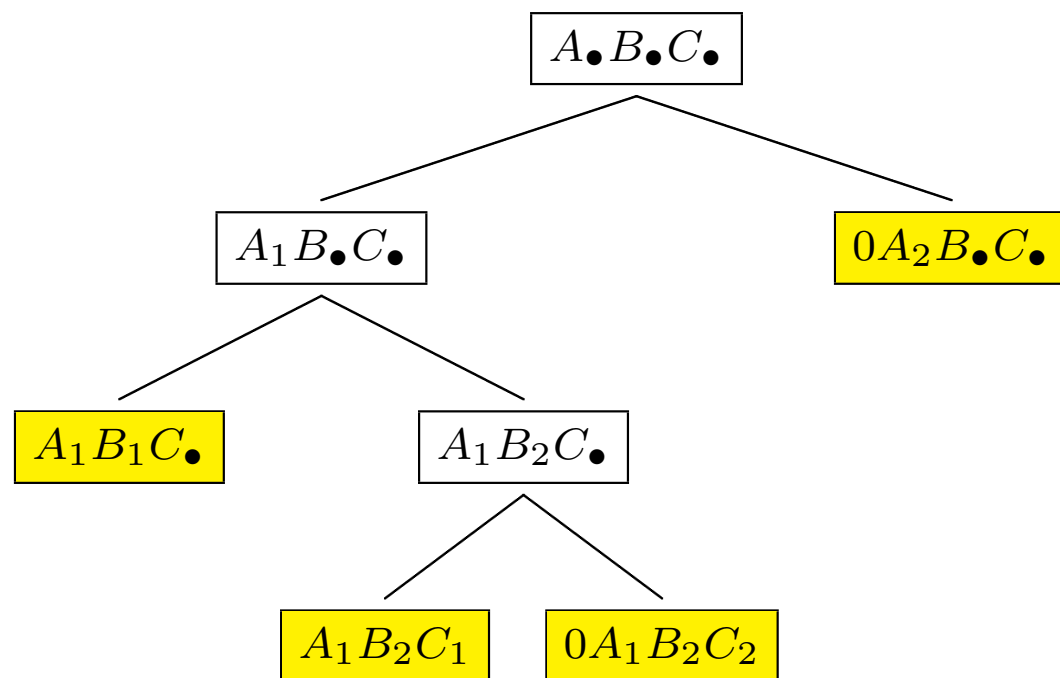
- Young is A_1 and old is A_2 . Male is B_1 and female is B_2 . Smoking is C_1 and non-smoking is C_2 . A_\bullet is any variant of the A factor.
- A subgroup is a concatenation of letters with numbered subscripts; $A_1B_1C_\bullet$ is young males, reflecting the fact that no split has been made according to smoking status. Young male smokers are $A_1B_1C_1$.

Key Fact: Allowing only subgroups of this form is a very strong restriction. For instance, $\{\{\text{male smokers}\} \cup \{\text{female nonsmokers}\}\}$ is not an allowed subgroup.

- For F factors, there are 3^F allowable subgroups (81 for $F=4$).
- For F factors, there are $(2^{2^F} - 1)$ possible subgroups (65,535 for $F=4$).

Allowed statistical models consist of partitions of the population into allowable subgroups arising from terminal nodes of trees based on factor splits, with possible zero effects, as follows:

- A factor ordering is selected (probabilistically), e.g. ABC , ACB .
- At each level of the tree, one zero-effect node is possibly assigned by
 - randomly determining if there is to be a zero-effect node at that level; if so, it will be denoted by a ‘0’ in front of the label.
 - then randomly choosing one of the nodes at that level to become the zero-effect node; it then becomes a terminal node.
- The non zero-effect nodes at a given level are possibly split by the factor corresponding to that level.
 - randomly deciding if the node is to be split; if not it becomes a terminal node;
 - if split, creating two new nodes at the next level of the tree.
- The statistical model (population partition) is the collection of terminal nodes (i.e., the last nodes in the branches of the tree).

Construction Stepsno 0; split on A 0 assigned, terminal; split on B no 0; no split, terminal; split on C

0 assigned

Thus the ensuing statistical model \mathcal{M} (population partition) consists of the four yellow nodes, two of which have zero treatment effect and two of which have non-zero and differing treatment effects.

Motivation for the choice of models (population partitions):

Why declare a node to be terminal after a failure to split? Suppose one split on A and then split on B for one branch and C for the other branch, declaring the resulting nodes to be terminal. The resulting partition is $\{\text{young male, young female, old smoker, old non-smoker}\}$, which does not seem to be a reasonable scientific model (i.e., that the population divides up into these four distinct subgroups, each having differing effects).

Why allow only one terminal node at each level to be a zero-effect node? Consider, for instance, the model $\{B_1C_1A_\bullet, 0B_1C_2A_\bullet, 0B_2C_1A_\bullet, B_2C_2A_\bullet\}$, which has two zero effect subgroups at level 2 of the tree. This model would be saying that non-smoking men and smoking women have zero effect, while the others have non-zero effect, which does not seem scientifically plausible.

There can be multiple zero-effect nodes at a given level that are plausible.
Yes, but then they will have occurred also at a higher level of the tree.

The Elicited Prior contains user specified features, such as

- k , the number of factor splits allowed;
- the prior probability of each factor having an effect;
- the prior probability of ‘zero effect’ for an individual (or this can be left unknown).

The Operational Prior consists of

- probabilities on the factor ordering;
- the tree splitting probabilities;
- the probabilities of assigning ‘zero effect’ to terminal nodes.

Challenges:

- Choosing the operational prior so that the resulting model probabilities match those from the elicited.
- Determining what is proper societal control for multiplicity and null control.

Full Bayesian inference:

- Perform a standard Bayesian model uncertainty analysis:
 - Pre-compute the prior model probabilities, including integrating out any unknown multiplicity parameters.
 - Utilize standard objective model parameter priors (e.g., Zellner-Siow).
 - Utilize appropriate stochastic search or other computational strategies if the model space is huge.
- Of primary interest is the posterior probability of an effect for an individual with characteristics X :
 - found by summing the posterior probabilities of all models in which individuals with those characteristics were in a subgroup that exhibited an effect;
 - equivalent to the posterior probability of an effect for a last level subgroup.
 - It is hard to make sense of the posterior probability of an effect for a higher level subgroup.

Illustration: Analysis of data from the Step trial of MRKAd5 vaccine

Overall Population: data provided little evidence of any effect, beneficial or harmful

Uncircumcised men:

6 HIV cases reported in the 394 individuals receiving placebos

22 HIV cases reported in the 394 individuals receiving the treatment

Two-sided p-value: 0.0021.

This was so small that there seemed to be conclusive evidence of harm from the vaccine, and all testing with this vaccine and other variants was stopped.

Bayesian Analysis: Objective Prior

Let $\theta = [\text{P}(HIV \text{ under placebo}) - \text{P}(HIV \text{ under vaccine})]$ in the subgroup.

Null and alternative hypotheses:

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0,$$

Objective prior (to the scientists):

- Choose $Pr(\theta = 0) = 0.5$, or deal with it by a sensitivity study.
- Give equal weight, to say, the vaccine doubling or halving the infection rate.
- Do this up to an upper (lower) limit of a five-fold increase or five-fold decrease in the infection rate.

Bayesian answer: $Pr(\theta = 0 \mid data) = 0.04$ and $Pr(\theta < 0 \mid data) = 0.96$.

Although this is much larger than the p -value of 0.0021, it still seems to be strong evidence that the vaccine was harmful to this subgroup.

Multiple Testing Adjustment for Subgroups:

Suppose we had initially considered five subgroups:

- all men
- circumcised men
- uncircumcised men
- men with negative Cd5 count
- men with positive Cd5 count

The Bayesian multiple testing adjustment would convert the earlier $Pr(\theta = 0 \mid data)$ from 0.041 to 0.27.

In reality, there were 18 subgroups considered apriori – just among the males – so the adjustment for multiple testing should be even larger.

Note that the Bayesian adjustment can be done posthoc, with information concerning the subgroups considered through the design.

V. Multiplicities in High-Energy Physics

A Bayesian Formulation of the Basic HEP Problem

The statistical model (following Richard Lockhart's Banff II writeup):

- N is the observed Poisson number of events.
- The events are independent and each has characteristics ('marks' in the Poisson process world) $X_i, i = 1, \dots, N$.
- Under H_0 : *background only*,
 - the mean of N is b ,
 - the density of the X_i is $f_b(x) > 0$.
- There may be a signal Poisson process with mean s and density $f_s(x)$.
- Under H_1 : *background + signal*,
 - the mean of N is $b + s$,
 - the density of the X_i is $(\gamma f_b(x) + (1 - \gamma) f_s(x))$, where $\gamma = \frac{b}{(b+s)}$.
- Consider the case where $f_b(x)$ and $f_s(x)$ are known but b and s are unknown.

Bayes factor of H_1 to H_0 for priors $\pi_0(b)$ and $\pi_1(b, s) = \pi_0(b)\pi_1(s | b)$:

$$\begin{aligned} B_{10} &= \frac{\int_0^\infty \int_0^\infty (b+s)^N e^{-(b+s)} \prod_{i=1}^N [\gamma f_b(x_i) + (1-\gamma) f_s(x_i)] \pi_1(b, s) ds db}{\int_0^\infty b^N e^{-b} \prod_{i=1}^N [f_b(x_i)] \pi_0(b) db} \\ &= \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^N \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)} \right] \pi_0(b) \pi_1(s | b) ds db}{\int_0^\infty b^N e^{-b} \pi_0(b) db}. \end{aligned}$$

Note that, if b is known, this becomes

$$B_{10} = \int_0^\infty e^{-s} \prod_{i=1}^N \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)} \right] \pi_1(s | b) ds.$$

Priors: Intrinsic priors are $\pi_0^I(b) = b^{-1/2}$ (note that it is improper) and $\pi_1^I(s | b) = b(s+b)^{-2}$ (note that it is proper).

Note: Ignoring the densities f_s and f_b and basing the answer solely on N is equivalent to assuming that $f_s \equiv f_b$.

Multiplicity (look-elsewhere) concerns are automatically handled:

Suppose N_j of the X_i are in bin B_j , $j = 1, \dots, M$, and that we assume we have only densities $f_s(B_j)$ and $f_b(B_j)$. Then

$$B_{10} = \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^M \left[1 + \frac{s f_s(B_j)}{b f_b(B_j)} \right]^{N_j} \pi_0(b) \pi_1(s | b) ds db}{\int_0^\infty b^N e^{-b} \pi_0(b) db}.$$

Suppose $f_s(B_j)$ gives probability one to some unknown bin B (the signal could occur in only one bin), with each bin being equally likely. Then

$$\begin{aligned} B_{10} &= \frac{E^B \left[\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^M \left[1 + \frac{s f_s(B)}{b f_b(B_j)} \right]^{N_j} \pi_0(b) \pi_1(s | b) ds db \right]}{\int_0^\infty b^N e^{-b} \pi_0(b) db} \\ &= \frac{1}{M} \sum_{j=1}^M \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \left[1 + \frac{s}{b f_b(B_j)} \right]^{N_j} \pi_0(b) \pi_1(s | b) ds db}{\int_0^\infty b^N e^{-b} \pi_0(b) db}, \end{aligned}$$

so that the results from each H_j : *signal in B_j* are downweighted by $1/M$.

VI. Comparison of Bayesian and Frequentist Approaches to Multiplicity

Frequentist Approaches: Per-Comparison, Family-wise and FDR error-rates

For M tests, $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i \neq 0$,

	accept H_0	Reject H_0	
H_0 true	U	V	M_0
H_0 false	T	S	M_1
(observed \rightarrow)	W	R	M

R = total number of rejections (discoveries)

V = # false discoveries

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

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

Per-Comparison (PCER). Controls the proportion of false discoveries $\frac{E[V]}{M}$ at level α by testing each H_{0i} at level α

‘Ignores the multiplicity problem’ (too liberal)

Family-wise (FWER). Classical Bonferroni: Controls $\Pr(V \geq 1)$ at level $\leq \alpha$ by testing each H_{0i} at level $\frac{\alpha}{M}$.

Results in *very conservative* tests.

... something in between ...

False Discovery rate (FDR)

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

- focus on $\frac{V}{R}$ instead \rightsquigarrow % of false discoveries (erroneous rejections) among the rejected hypotheses
- Not defined for $R = 0$ (all M nulls accepted), so (Benjamini and Hochberg, 95) propose to control:

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

with Simes (86) α -level multiple comparisons test

- Closely related: **Positive FDR**: $\text{pFDR} = E \left[\frac{V}{R} \mid R > 0 \right]$

Properties and Comments:

	accept H_0	reject H_0	
H_0 true	U	V	M_0
H_1 true	T	S	M_1
	W	R	M

- Simes(86) shows control of FWER under null.
- B&H(95) show control of FDR always
- asymptotically $FDR \approx pFDR \approx PFP = \frac{E[V]}{E[R]}$
- Part of the attractiveness of FDR seems to be that, instead of using $\alpha = 0.05$, people use (e.g.) $FDR = 0.15$.
- Genovese and Wasserman(02, 03) observe that FDR is an expected value, and the realized proportion of discoveries, V/R can vary greatly from this expected value; arguing that this variability should be taken into account.

- B&H algorithm controls FDR at level $p\alpha$ (Finner and Roters, 04) \leadsto use and estimate of p to increase power (keeping control at level α); (Benjamini and Hochberg, 2000; Black 04, Storey, Storey et. al., Genovese, Langaas et al. (2005), Cayon, Rice)
- Finner and Roters(01) observe that control of FDR allows ‘cheating’ “by adding additional hypotheses of no interest which are known to have p -values near 0” (the FDR critical value for ranked p -values $p_{[i]}$ is $i\alpha/m$); for instance, to maximize the chance of rejecting 8 hypotheses of interest while controlling FDR at α one can add 100 ‘uninteresting’ hypothesis with p -values ≈ 0 , so that the 8 ‘interesting’ p -values will have threshold $\geq 101\alpha/108$

Connections between FDR and Bayes

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

- Frequentists need to estimate p to obtain good power for FDR; this is also key for Bayesians. This hints that there should be some type of exchangeability of hypotheses to apply FDR; this would also address the Finner-Rotens objection.
- Genovese and Wasserman (02) have a more Bayesian definition of “Bayesian FDR” \rightsquigarrow focus on *realized* FDR, namely V/R , and its posterior distribution, so that uncertainty in V/R can be studied. (They still study frequentist properties.)

- Frequentist analyses that estimate (instead of control) errors, as pFDR, often have same models as Bayesians

$$f(x_i) = pf_0(x_i) + (1 - p)f_1(x_i)$$

- f_1 unknown (and often also f_0) \rightsquigarrow often frequentist nonparametric estimates (EB needs only estimate the ratio) (Efron&al 01, Efron&Tibshirani 02, Genovese&Wasserman)
- often p is not estimated, but a lower bound used instead

$$f = pf_0 + (1 - p)f_1 \geq pf_0 \rightsquigarrow \hat{p} = \min_x \frac{\hat{f}(x)}{\hat{f}_0(x)}.$$

- Also full nonparametric Bayes analysis. (Do, Müller, Tang)

- Storey suggests that

- pFDR has a dual interpretation as a Bayesian and as a frequentist measure because

$$\text{pFDR}(C) = \Pr(H_{0i} \text{ true} \mid X_i \in C) = \frac{E[V]}{E[R]}.$$

But this is the posterior probability given the data is in ‘critical region’ C , not given the data itself;

- proposes the q -value, defined as

$$q\text{-value}(X_i) = \inf_{\alpha} \Pr(H_{0i} \text{ true} \mid X_i \in C_{\alpha})$$

and calls it a “posterior Bayesian p -value,” but, again, it depends on a tail region of data, not 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

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

and average over the rejection region

- to ‘control Bayesian pFDR’ at level α , reject the i -th null if $p_i > c^*$ where c^* gives the largest rejection region for which the above average $\leq \alpha$
- **But** FDR is taken a priori being the quantity of interest. Is this reasonable from a Bayesian viewpoint?

Decision-theoretic Evaluations of FDR

- FDR seems most useful for screening but, in depending only on p -values, it seems like it would not reflect typical decision-theoretic screening losses, which depend in part on the magnitude of the effects.
- FDR does not seem good for ‘discovery’ which corresponds to “0-1” loss; indeed, one cannot derive FDR from this loss. ($E[V]$ and $E[T]$ arise, but not versions of the ratio V/R .)
- One could argue that it is a ‘global loss’ but there are difficulties in interpretation.
- Müller, et. al. (2002) study a variety of losses that include FDR (or variants) *as primitives*, but find problems with doing so.

Difficulties in interpreting FDR as a loss function

- It could be argued that the Loss for taking decisions d when the truth is H could be directly defined as a linear function of V/R ; the risk would then be a function of the (expected) FDR.
- Let $d_i = 0$ if i -th null is accepted, and $H_i = 0$ if i -th null is true. The problem: such a loss function does not depend on $\sum_i H_i$, only on $\sum_i (1 - d_i)H_i$ (and on $\sum_i d_i$)
- This is difficult to justify intuitively (at least for 'scientific' purposes)

- Assume we have $S = 18$ ‘true’ signals detected and that $FDR = 2/20$ or 1%. We have identical loss if:
 - We only had $M_0 = 2$ noise (*all* noise is declared to be signal)
 - We have $M_0 = 20000$ and hence the procedure is superb in sorting out noise
- In the same situation as before ($R = 20, V = 2$) we have the same loss if
 - In reality there are $M_1 = 18$ signals, so they are all discovered
 - The truth is that there are $M_1 = 1800$ signals so we only ‘discover’ 1/10000.
- Efforts have been made to define Loss Functions that also take into account the FNR, but this has also problems

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’ (*expected*) 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’ (*expected*) 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 p_i , all of them data-dependent except for the 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 $p_i \approx 0$ ".
- 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

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.
- Bickel, D.R. (2003). Error-rate and decision-theoretic methods of multiple testing. Alternatives to controlling conventional false discovery rates, with an application to microarrays. *Tech. Rep*, Office of Biostatistics and Bioinformatics, Medical College of Georgia.
- Do, K.A., Müller, P., and Tang, F. (2002). *Tech. Rep*, Department of Bioestatics, University of Texas.
- Dudoit, S., Shaffer, J.P., and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.
- DuMouchel, W.H. (1988). A Bayesian model and graphical elicitation model for multiple comparison. In *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.) 127–146. Oxford University Press.
- Duncan, D.B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7**, 171-222.
- Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23** 70–86.
- 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. (2002a). 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. (2002b). Bayesian and frequentist multiple testing. *Tech. Rep.* Department of Statistics, Carnegie-Mellon University.
- Morris, J.S., Baggerly, K.A., and Coombes, K.R. (2003). Bayesian shrinkage estimation of the relative abundance of mRNA transcripts using SAGE. *Biometrics*, **59**, 476–486.
- Müller, P., Parmigiani, G., Robert, C., and Rouseau, J. (2002), “Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays,” Tech. rep., University of Texas, M.D. Anderson Cancer Center.
- Newton, M.A., and Kendziorski, C.M. (2003). Parametric empirical Bayes methods for microarrays. In *The analysis of gene expression data: methods and software*, 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.
- Scott, G. and Berger, J.O. (2003). An exploration of Bayesian multiple testing. To appear in Volume in honour of Shanti Gupta.
- 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*
- Storey J.D., Taylor, J.E., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society B* **66** 187–205.
- Storey J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *PNAS* **100** 9440–9445.
- 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.