On the Choice of Priors in Bayesian (and Conditional Frequentist) Testing

Jim Berger

Duke University

Workshop on statistical issues related to discovery claims
Banff, Canada
July 14, 2010

Outline

- A testing example
- Why p-values can differ markedly from Bayesian answers
- Conditional frequentist testing
- Choice of priors for common nuisance parameters
- Choice of priors for non-common nuisance parameters
- LEE (Multiplicity) in testing: the choice of prior model probabilities

A Testing Example

San Jose Mercury News

mercurynews.com WEST VALLEY 102 Friday, September 25, 2009 THE NEWSPAPER OF SILICON VALLEY 75 cents

AIDS MILESTONE

New path for HIV vaccine

Some in study protected from infection, but trial raises more questions

By Karen Kaplan and Thomas H. Maugh II
Los Angeles Times

Hours after HIV researchers announced the achievement of a milestone that had eluded them for a quarter of a century, reality began to set in: Tangible progress could take another decade.

A Thai and American team announced early Thursday in Bangkok that they had found a combination of vaccines providing modest protection against infection with the virus that causes AIDS, unleashing excitement worldwide. The idea of a vaccine to prevent infection with the human immunodeficiency virus, HIV, had long been frustrating and fruitless.

But by Thursday afternoon, initial euphoria gave way to a more sober assessment. There is still a very long way to go before reaching the goal of producing a vaccine that reliably shields people from HIV.

Some researchers questioned whether the apparent 31 percent reduction in infections was a sta-

See **VACCINE**, *Page 14*

A researcher during the Thai phase III HIV Vaccine Trial, also known as RV 144, tests the "prime-boost" combination of two vaccines.

ASSOCIATED PRESS

Hypotheses, Data, and Classical Test:

- Alvac had shown no effect
- Aidsvax had shown no effect

Question: Would Alvac as a primer and Aidsvax as a booster work?

The Study: Conducted in Thailand with 16,395 individuals from the general (not high-risk) population:

- 74 HIV cases reported in the 8198 individuals receiving placebos
- 51 HIV cases reported in the 8197 individuals receiving the treatment

Model: $X_1 \sim \text{Binomial}(x_1 \mid p_1, 8198)$ and $X_2 \sim \text{Binomial}(x_2 \mid p_2, 8197)$, respectively, so that p_1 and p_2 denote the probability of HIV in the placebo and treatment populations, respectively.

Classical test of $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$ yielded a p-value of 0.04.

Bayesian Analysis: Reparameterize to p_1 and $V = 100\left(1 - \frac{p_2}{p_1}\right)$, so that we are testing

$H_0 : V = 0$, p_1 arbitrary

$H_1 : V \neq 0$, p_1 arbitrary.

Prior distribution:

- $Pr(H_i)$ = prior probability that H_i is true, $i = 0, 1$,
- Let $\pi_0(p_1) = \pi_1(p_1)$, and choose them to be either
  - uniform on (0,1)
  - subjective (scientific?) priors based on knowledge of HIV infection rates

  Note: the answers are essentially the same for either choice.
- For V under H_1, consider the priors
  - uniform on (-20, 60) (apriori subjective – scientific? – beliefs)
  - uniform on $(-100c/3, 100c)$ for $0 < c < 1$, to study sensitivity

  (constrained also to $V > 100(1 - \frac{1}{p_1})$.

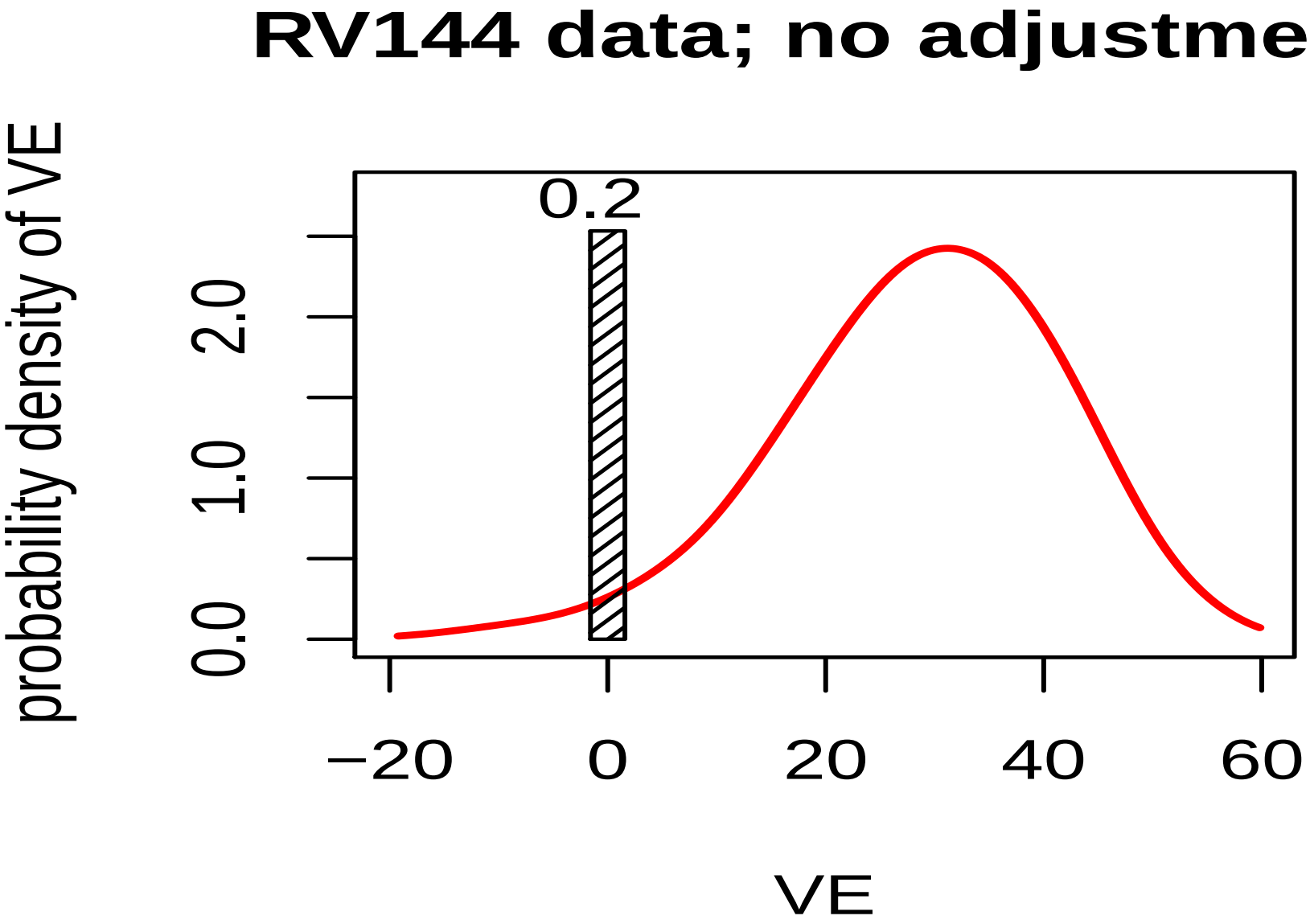
Posterior probability of the null hypothesis:

$$Pr(H_0 \mid \text{data}) = \frac{Pr(H_0)B_{01}}{Pr(H_0)B_{01} + Pr(H_1)} ,$$

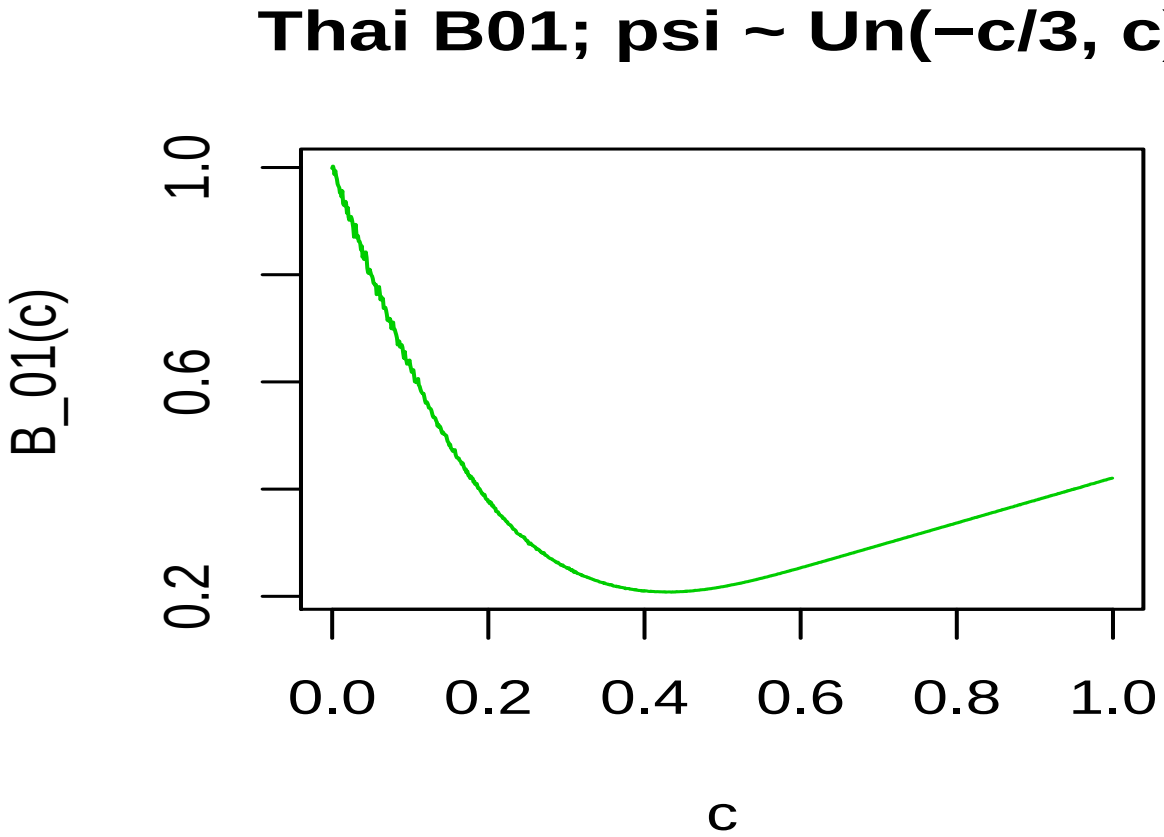
where the Bayes factor of H_0 to H_1 is

$$B_{01} = \frac{\int \text{Binomial}(74 \mid p_1, 8198)\, \text{Binomial}(51 \mid p_1, 8197)\pi_0(p_1)dp_1}{\int \text{Binomial}(74 \mid p_1, 8198)\, \text{Binomial}(51 \mid p_2, 8197)\pi_0(p_1)\pi_1(p_2 \mid p_1)dp_1 dp_2} .$$

- For the prior for V that is uniform on (-20, 60),
  $B_{01} \approx 1/4$ (recall, p-value $\approx .04$)
- If the prior probabilitites of the hypotheses are each 1/2, the overall posterior density of V has
  - a point mass of size 0.20 at $V = 0$,
  - a density (having total mass 0.80) on non-zero values of V:


RV144 data; no adjustment
probability density of VE
0.0
1.0
2.0
0.2
−20
0
20
40
60
VE

Robust Bayes: For the prior on V that is uniform on $(-100c/3, 100c)$, the Bayes factor is



Note: There is sensitivity to c, indeed $0.22 < B_{01}(c) < 1$, but why would this cause one to instead report $p = 0.04$, knowing it will be misinterpreted?

Note: Uniform priors are the extreme points of monotonic priors, and so such robustness curves are quite general.

Incorporating information from multiple tests: To adjust for the two previous similar failed trials, the (exchangeable) Bayesian solution

- assigns each trial common unknown probability p of success, with p having a uniform distribution;
- computes the resulting posterior probability that the current trial exhibits no efficacy

$$Pr(H_0 \mid \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = \left(1 + \frac{B_{01}(\boldsymbol{x}_1)B_{01}(\boldsymbol{x}_2) + B_{01}(\boldsymbol{x}_1) + B_{01}(\boldsymbol{x}_2) + 3}{3B_{01}(\boldsymbol{x}_1)B_{01}(\boldsymbol{x}_2) + B_{01}(\boldsymbol{x}_1) + B_{01}(\boldsymbol{x}_2) + 1} \times \frac{1}{B_{01}(\boldsymbol{x}_3)}\right)^{-1}$$

where $B_{01}(\boldsymbol{x}_i)$ is the Bayes factor of "no effect" to "effect" for trial i.

The result is $Pr(H_0 \mid \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = 0.29$.

Sources of the difference between p-values and Bayes factors

1. A large part of the difference is due to a significant problem with p-values: they, in effect, replace the observed data x by the tail area $\{X : X \geq x\}$. (Imagine the different scientific insight in being told that $x = 5$, as opposed to being told that the measuring instrument indicates only that x is somewhere in the interval from 5 to ∞.)

2. Part of the difference will also typically arise from the Ockham's razor effect of Bayes factors, when the original model for H_1 contains more unknown parameters that the original model for H_0.

3. The third source of the difference is that the Bayes factor is comparing the null and alternative hypotheses, while the p-value is only computed based on the null model (although the choice of test statistic X will often, at least informally, depend on considerations of alternatives).

For some insight into this latter point, consider the situation where there is an unknown additive systematic bias b in the experiment, so that we really observe $X' = X + b$, which has density $f_i(x' - b)$ under H_i. We don't know this, however, so we end up testing the "statistical hypotheses" (as distinct from the original scientific hypotheses)

$H_0 : X' \sim f_0(x')$

$H_1 : X' \sim f_1(x')$.

The Bayes factor we compute is

$$B(x') = \frac{f_0(x')}{f_1(x')} = \frac{f_0(x+b)}{f_1(x+b)} = B(x+b)\,.$$

and, likewise, the p-value is

$$p(x') = \int_{x'}^{\infty} f_0(y)\; dy = p(x+b)\,.$$

Now suppose $p(x+b)$ is small because b is large, while $B(x+b)$ is not small.

Example: Jefferys (1990) analyzed a pyschokinesis experiment:

H_0 : the subjects did not exhibit psychokinesis

H_1 : the subjects did exhibit psychokinesis.

Results: p-value was 0.0003, while the Bayes factor varied between 0.01 and 12, depending on the prior utilized for the parameters under H_1 (with Bayes factors less than one arising only from priors that would have been very unreasonable apriori). Two interesting aspects:

1. The Bayes factor seems considerably more resistant to bias than does the p-value for the primary goal of determining if psychokinesis was exhibited, because it is comparing hypotheses, and the bias may have a similar effect on the evidence under each hypothesis.

2. The p-value is useful for the secondary goal of indicating that some bias is likely present; indeed, when the p-value is small and the Bayes factor is not, this would seem to be a strong indication of bias (or some other misspecification of the models).

Conditional Frequentist Testing

Often Bayesian tests are also conditional frequentist tests arising as follows (see Berger, Brown and Wolpert (1994), Berger, Boukai and Wang (1997a, 1997b), Dass (1998), and Dass and Berger (2001)):

- The Bayes test having equal prior probabilities of H_0 and H_1 yields posterior probability

$$Pr(H_0 \mid x) = \frac{\int f_0(x \mid \theta_0)\pi(\theta_0)d\theta_0}{\int f_0(x \mid \theta_0)\pi(\theta_0)d\theta_0 + \int f_1(x \mid \theta_1)\pi(\theta_1)d\theta_1} .$$

- The priors produce a conditioning statistic, $T(x) = \max\{p_0(x), p_1(x)\}$, defining parts of the sample space containing 'equivalent evidential strength', where p_i is the p-value for $H_i^* : X \sim \int f_i(x \mid \theta_i)\pi(\theta_i)d\theta_i$.
- If the rejection region is $R = \{x : p_0(x) < p_1(x)\}$,
  - the conditional Type I error probability $\alpha(x) = Pr(R \mid H_0^*, T)$ then equals $Pr(H_0 \mid x)$ *numerically*,
  - the conditional Type II error probability $\beta(x) = Pr(\bar{R} \mid H_1^*, T)$ then equals $Pr(H_1 \mid x)$ *numerically*.

- If the null model is simple or possesses a group invariance structure with nuisance parameters, $\alpha(x) = Pr(R \mid H_0, T)$, so this is a fully valid frequentist error probability for the original hypothesis.
- One computes the marginal likelihoods only once, whereas computation of the unconditional error probability would require simulation from the marginal distribution.
- Note there are many other conditional frequentist tests, so the range of frequentist answers is much larger than the range of Bayesian answers.
- The unconditional frequentist test (for a given test statistic) achieves uniqueness by being the uniformly worst procedure among the family of conditional frequentist tests.

Choosing priors for "common parameters" in testing

If pure subjective (scientific?) choice is not possible, here are some guidelines:

- Gold standard: if there are parameters in each hypothesis that have the same group invariance structure, one can use the right-Haar priors for those parameters (even though improper) (Berger, Pericchi and Varshavsky, 1998)
- Silver standard: if there are parameters in each hypothesis that have the same scientific meaning, reasonable default priors (e.g. the constant prior 1) can be used (e.g., in the vaccine example, where p_1 meant the same in H_0 and H_1).
- Bronze standard: to try to obtain parameters that have the same scientific meaning (beware the "fallacy of Greek letters"), one strategy often employed is to orthogonalize the parameters, i.e., reparameterize so that the partial Fisher information for those parameters is zero.

Example: (location-scale)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d with density

$$f(x_i \mid \mu, \sigma) = \frac{1}{\sigma}\, g\left(\frac{x_i - \mu}{\sigma}\right)$$

Several models are entertained:

M_N: g is $N(0,1)$

M_U: g is Uniform $(0,1)$

M_C: g is Cauchy $(0,1)$

M_L: g is Left Exponential ($\frac{1}{\sigma}\, e^{x-\mu}$, $x \leq \mu$)

M_R: g is Right Exponential ($\frac{1}{\sigma}\, e^{-(x-\mu)}$, $x \geq \mu$)

All models have location-scale parameters (μ, σ), for which the right-Haar prior density is

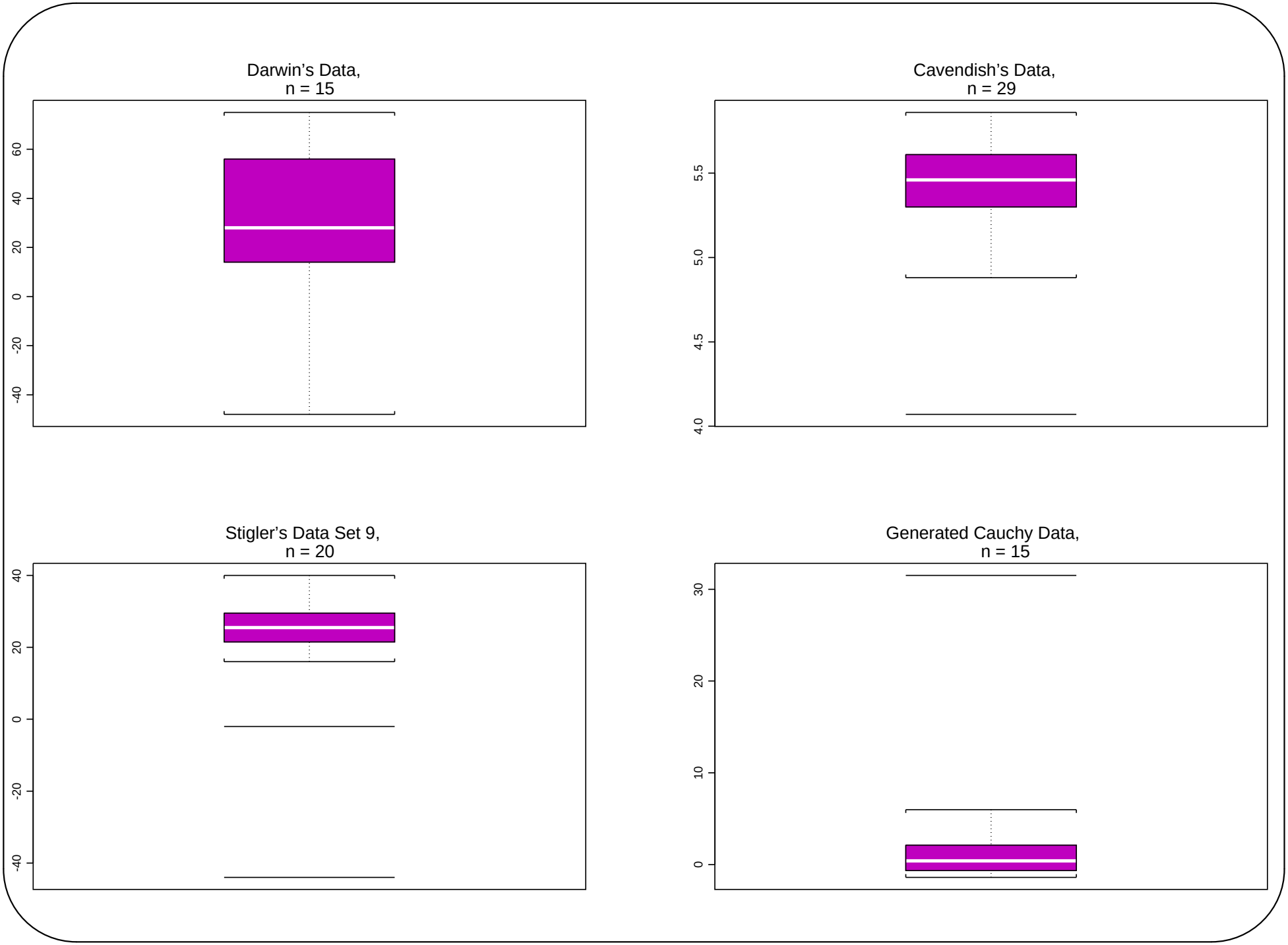
$$\pi(\mu, \sigma) = \frac{1}{\sigma}\,.$$

The marginal densities

$$m(\mathbf{x} \mid M) = \int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^n \left[\frac{1}{\sigma} g\left(\frac{x_i - \mu}{\sigma} \right) \right] \frac{1}{\sigma} \, d\mu \, d\sigma$$

for the five models are given by

1. Normal: $m(\mathbf{x} \mid M_N) = \dfrac{\Gamma((n-1)/2)}{(2\,\pi)^{(n-1)/2} \sqrt{n} \left(\sum_i (x_i - \bar{x})^2 \right)^{(n-1)/2}}$
2. Uniform: $m(\mathbf{x} \mid M_U) = \dfrac{1}{n\,(n-1)(x_{(n)} - x_{(1)})^{\,n-1}}$
3. Cauchy: $m(\mathbf{x} \mid M_C)$ is given in Spiegelhalter (1985).
4. Left Exponential: $m(\mathbf{x} \mid M_L) = \dfrac{(n-2)!}{n^n (x_{(n)} - \bar{x})^{\,n-1}}$
5. Right Exponential: $m(\mathbf{x} \mid M_R) = \dfrac{(n-2)!}{n^n (\bar{x} - x_{(1)})^{\,n-1}}$


Darwin's Data,
n = 15
60
40
20
0
-20
-40
Cavendish's Data,
n = 29
5.5
5.0
4.5
4.0
Stigler's Data Set 9,
n = 20
40
20
0
-20
-40
Generated Cauchy Data,
n = 15
30
20
10
0

The objective posterior probabilities of the five models, for each data set are as follows:

Data set	Models				
	Normal	Uniform	Cauchy	L. Exp.	R. Exp.
Darwin	.390	.056	.430	.124	.0001
Cavendish	.986	.010	.004	4×10^{-8}	.0006
Stigler 9	7×10^{-8}	4×10^{-5}	.994	.006	2×10^{-13}
Cauchy	5×10^{-13}	9×10^{-12}	.9999	7×10^{-18}	1×10^{-4}

Choosing priors for non-common parameters

If subjective (scientific?) choice is not possible, here are some guidelines:

- Vague proper priors are horrible (related to the Jeffreys-Lindley paradox): for instance, if $X \sim N(\mu, 1)$ and we test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ with a Uniform$(-c, c)$ prior for θ, the Bayes factor is

$$B_{01}(c) = \frac{f(x \mid 0)}{\int_{-c}^{c} f(x \mid \mu)(2c)^{-1} d\mu} \approx \frac{2c\, f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu) d\mu}$$

  for large c, which depends dramatically on the choice of c.

- Improper priors are problematical, because they are unnormalized; is

$$B_{01} = \frac{f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu)(1) d\mu} \quad \text{or} \quad B_{01} = \frac{f(x \mid 0)}{\int_{-\infty}^{\infty} f(x \mid \mu)(2) d\mu} \, ?$$

- Robust solution: if one can specify a plausible range $c_1 \leq c \leq c_2$, look at $B_{01}(c)$ over this range and hope the conclusion is robust. (Not obvious for higher dimensional parameters, but there is a literature.)

Case 1: $\pi(\mu)$ is Uniform$(0, 10)$ (e.g., known upper limit on μ)

- Observe $x = 2$: $p = 0.025$, while $\Pr(H_0 \mid x = 2) = 0.54$
- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 1.3 \times 10^{-3}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 6.0 \times 10^{-8}$

Case 2: $\pi(\mu)$ is Normal$(4, 1)$ (arising from a previous experiment)

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 4.7 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 5.8 \times 10^{-8}$

Case 3: $\pi(\mu)$ is a point mass at 4 (the prediction of a new theory).

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 3.4 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 1.1 \times 10^{-7}$

Conservative conversion of p to Pr$(H_0 \mid x)$*:* $\Pr(H_0 \mid x) = (1 + (-ep \log p)^{-1})^{-1}$:

- Observe $x = 4$: $p = 3.1 \times 10^{-5}$, while $\Pr(H_0 \mid x = 4) = 8.8 \times 10^{-4}$
- Observe $x = 6$: $p = 1.0 \times 10^{-9}$, while $\Pr(H_0 \mid x = 6) = 5.7 \times 10^{-8}$

Various proposed default priors for non-common parameters

- Fractional priors (O'Hagan): use a fraction γ of the model likelihood (usually $\gamma =$'parameter dimension'/'sample size') as the prior, with $L(\theta)^{1-\gamma}$ as the likelihood.
- Intrinsic priors (Berger, Pericchi and others): generate priors from "training samples" (either actual subsets of the data, or imaginary data generated under the null model).
- Conventional priors that have at least some nice properties: e.g., Zellner-Siow priors for linear models are
  - invariant to scale changes in covariates
  - consistent (the true model will be selected as $n \to \infty$)
  - information-consistent (e.g., will reject as t or F statistics $\to \infty$)
  - coherent (roughly, are logically connected)
- Various efforts at 'predictive matching' priors.
- Approximations (such as BIC); these can capture part of the prior influence, but not all.

Random (and likely incoherent) thoughts about Bob's talk:

- If the parameter prior is part of the proposed theory (e.g., general relativity, standard Higgs prediction?), it should be subjected to Bayesian updating.
  - At the extreme, if the theory predicts mass in (5,10), and that range is excluded, the theory is wrong (though resurrection is possible).
- Bayes is really about how to process information from different sources and, in such a way, that all known uncertainties are accounted for:
  - The temporal view of Bayes as the way knowledge accumulates is silly, in that it assumes priors are perfect.
    - Priors do often get updated by data, but also often change by – oooppppss, I can't believe I forgot about that, or wow – that's a piece of knowledge I had never imagined.
  - All that really matters is that, when presenting the analysis, the prior(s) are defensible.

- If the theory says nothing about the parameter (e.g., axion theory), any prior will be solely an "investigative prior" proposed as a way to interrogate the data.
  - Use of Bayes factors helps a lot here, in that one leaves the current estimate of the prior probability unspecified; everyone can process the effect of exclusions informally.
  - The 'default' Bayesian choices of priors for testing are only based on the current situation, not on what happened before.

Bayesian Approach to LEE (Multiplicity)

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

Simple Example: Multiple Testing under Exclusivity

Suppose one is testing mutually exclusive hypotheses $H_i,\ i = 1, \ldots, m$, so each hypothesis is a separate model. If the hypotheses are viewed as exchangeable, choose $P(H_i) = 1/m$.

Example: 1000 energy channels are searched for a signal:

- if the signal is known to exist and occupy only one channel, but no channel is theoretically preferred, each channel can be assigned prior probability 0.001.
- if the signal is not known to exist (e.g., it is the prediction of a non-standard physics theory) prior probability 1/2 should be given to 'no signal,' and probability 0.0005 to each channel.

This is the Bayesian solution regardless of the structure of the data.

Example: In each channel, one tests $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i > 0$.

Data: $X_i, i = 1, ..., m$ are normally distributed with mean θ_i, variance 1, and correlation ρ.

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 μ_i to be the signal if X_i is the largest value and $X_i > K$."
- Compute the corresponding error probability:

$$\alpha = E^Z \left[1 - \Phi \left(\frac{K - \sqrt{\rho} Z}{\sqrt{1-\rho}} \right)^m \right],$$

  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 $\rho \to 1$.

General Approach to Bayesian Multiplicity Adjustment

1. Represent the problem as a *model uncertainty* problem: Models $\mathcal{M}_i$, with densities $f_i(\mathbf{x} \mid \boldsymbol{\theta}_i)$ for data $\mathbf{x}$, given unknown parameters $\boldsymbol{\theta}_i$; prior distributions $\pi_i(\boldsymbol{\theta}_i)$; and marginal likelihoods $m_i(\mathbf{x}) = \int f_i(\mathbf{x} \mid \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$.

2. Specify prior probabilities, $P(\mathcal{M}_i)$, of models to reflect the multiplicity issues; **Bayesian analysis controls multiplicity through** $P(\mathcal{M}_i)$ [a]

- *Subjective Bayesian Analysis:* If the $P(\mathcal{M}_i)$ are real subjective probabilities, that's it: multiplicity correction has been done.
- *Objective Bayesian Analysis:* One has to be careful to make choices of the $P(\mathcal{M}_i)$ that ensure multiplicity correction (e.g., specifying equal prior probabilities does *not* generally control multiplicity)!

3. Implement Bayesian model averaging (model selection?), based on

$$P(\mathcal{M}_i \mid \mathbf{x}) = \frac{P(\mathcal{M}_i)\ m_i(\mathbf{x})}{\sum_{j=1}^{k} P(\mathcal{M}_j)\ m_j(\mathbf{x})}\ .$$

[a] see, e.g., Jeffreys 1961; Waller and Duncan 1969; Meng and Demptster 1987; Berry 1988; Westfall, Johnson and Utts 1997; Carlin and Louis 2000.

Variable Selection

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

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

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

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

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

Posterior probability of $\mathcal{M}_i$:

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

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

Equal prior probabilities: $P(\mathcal{M}_i) = 2^{-m}$

Bayes exchangeable variable inclusion:

- 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, who also suggested alternative choices of the $P(\mathcal{M}_i)$. Probably $a = b = 1/2$ is better.)
- Then, since k_i is the number of variables in model $\mathcal{M}_i$,

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

Empirical Bayes exchangeable variable inclusion: 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.

Controlling for multiplicity in variable selection

Equal prior probabilities: $P(\mathcal{M}_i) = 2^{-m}$ does *not* control for multiplicity here (as it did in the simpler examples); it corresponds to fixed prior inclusion probability $p = 1/2$ for each variable.

Empirical Bayes exchangeable variable inclusion does control for multiplicity, in that $\hat{p}$ will be small if there are many β_i that are zero.

Bayes exchangeable variable inclusion also controls for multiplicity (see Scott and Berger, 2008), although the $P(\mathcal{M}_i)$ are fixed.

Note: The control of multiplicity by Bayes and EB variable inclusion usually reduces model complexity, but is *different* than the usual Bayeisan Ockham's razor effect that reduces model complexity.

- The Bayesian Ockham's razor operates through the effect of model priors $\pi_i(\boldsymbol{\beta}_i, \sigma^2)$ on $m_i(\mathbf{x})$, penalizing models with more parameters.
- Multiplicity correction occurs through the choice of the $P(\mathcal{M}_i)$.

	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 1: 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

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

Posterior model probabilities, given p:

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

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

This *does* concentrate about the true p as $m \to \infty$, so one might expect that $P(\mathcal{M}_i \mid \mathbf{x}) = \int_0^1 P(\mathcal{M}_i \mid \mathbf{x}, p)\pi(p \mid \mathbf{x})dp \approx P(\mathcal{M}_i \mid \mathbf{x}, \hat{p}) \propto m_i(\mathbf{x})\ \hat{p}^{k_i}(1-\hat{p})^{m-k_i}$.

This is not necessarily true; indeed

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

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 $k_T/m \to p_T$ as $m \to \infty$, 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*

$$\hat{p} = p_T + O(\frac{1}{\sqrt{m}}) \quad as \quad m \to \infty \,.$$

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.

Summary

"I shall resist the temptation of saying more, because model selection is a can of worms for both objectivists and subjectivists and frequentists*."

*Bob Cousins, Tuesday.

Thanks!