# Bayesian Statistics Overview

Jim Berger

Duke University

*SAMSI-SAVI Workshop on Astrostatistics*
*SAMSI, RTP*
*September 19-21, 2012*

(first SAMSI workshop: September 21-23, 2002)

# Outline

- Introductory example: Ockham's razor
- Model selection and exoplanets
- Multiplicity adjustments
- Bayesian calibration of $p$-values
- Bayes/frequentist duality

# I. Ockham's Razor

- Preferring the simpler of two hypotheses to the more complex when both agree with the data is an old principle in science, attributed to the thirteen-century Franciscan monk William of Ockham (Occam in latin)

  *"Pluralitas non est ponenda sine necessitate."*

  (Plurality must never be posited without necessity.)

  *"Frustra fit per plura quod potest fieri per pauciora."*

  (It is vain to do with more what can be done with fewer.)

- Two aspects of simplicity:
  - Regard $H_0$ as *simpler* than $H_1$ if it makes *sharper predictions* about what data will be observed.
  - Models are more complex if they have extra adjustable parameters that can be tweaked to accommodate a wider variety of data.
    - "coin is fair" is a simpler model than "coin has unknown bias $\theta$."
    - $s = a + ut + \frac{1}{2}\,gt^2$ is simpler than $s = a + ut + \frac{1}{2}\,gt^2 + ct^3$.

**Example:** *Perihelion of Mercury* (with Bill Jefferys)

In the 19th century it was known that there was an unexplained residual motion of Mercury's perihelion (the point in its orbit where the planet was closest to the Sun) in the amount of approximately 43 seconds of arc per century; denote the true residual motion by $\theta$.

Various hypotheses:

- A planet 'Vulcan' close to the sun.
- A ring of matter around the sun.
- Oblateness of the sun.
- Law of gravity is not inverse square but inverse $(2 + \epsilon)$.
- General relativity

All these hypotheses, except general relativity, had a parameter that could be adjusted to deal with the observed residual motion.

**Data (in about 1920):** $X = 41.6$ where $X \sim N(x \mid \theta, 2^2)$.

**Prior distribution for $\theta$ under gravity model $M_G$:** $\pi_G(\theta) = N(\theta \mid 0, \tau^2)$.

- Symmetric about 0 (which corresponds to inverse square law).
- Decreasing away from zero; normality is convenient.
- Initially, $\tau = 50$, because a gravity effect which would yield $\theta > 100$ would have had other observed effects.

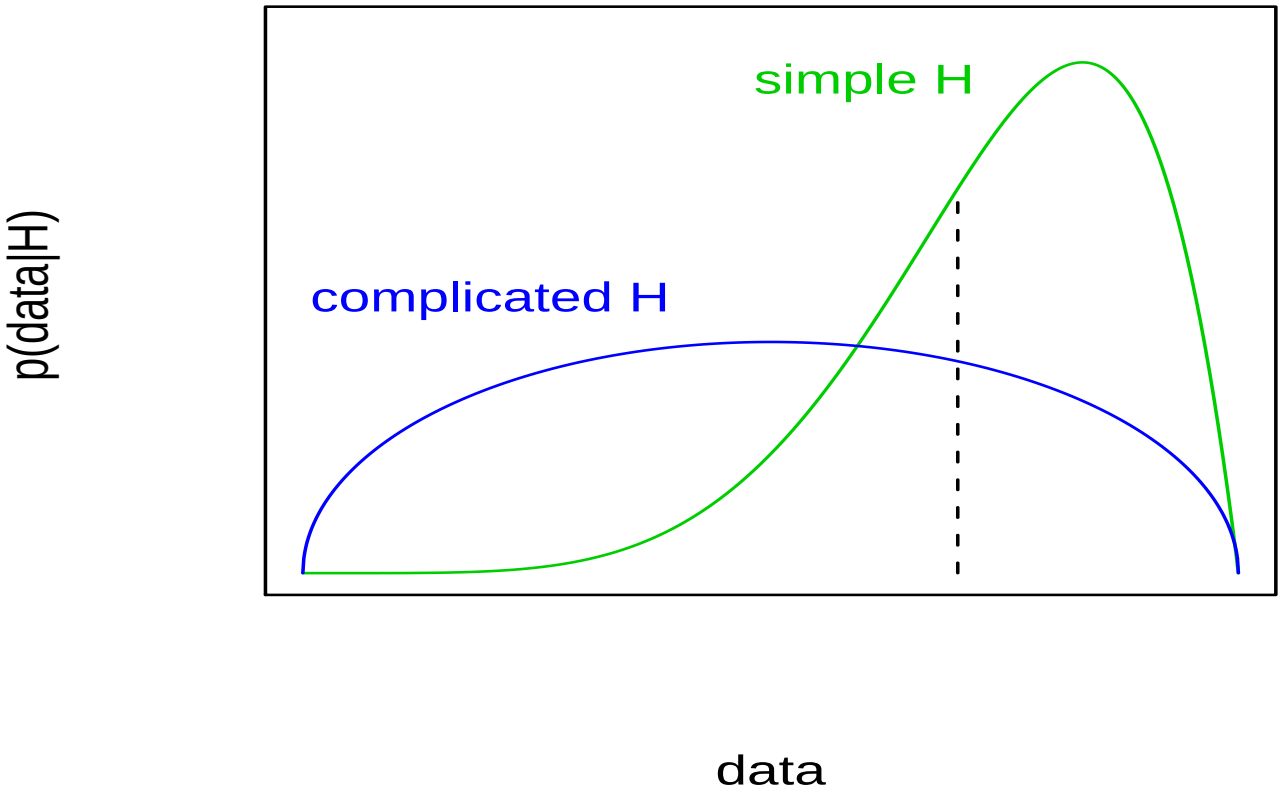
**General relativity model $M_E$:** Predicted $\theta_E = 42.9$.

**Bayes factor:**

$$
\begin{aligned}
B_{EG} &= \text{odds of obtaining the observed data under } M_E \text{ compared to } M_G \\
&= \frac{N(41.6 \mid \theta_E, 2^2)}{\int N(41.6 \mid \theta, 2^2)\pi_G(\theta)\, d\theta} \\
&= \frac{\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6 - 42.9)^2\right)}{\int \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6 - \theta)^2\right) \frac{1}{50\sqrt{2\pi}} \exp\left(-\frac{1}{2 \cdot 50^2}\theta^2\right)\, d\theta} \\
&= \frac{\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6 - 42.9)^2\right)}{\frac{1}{\sqrt{2 \cdot 2504\pi}} \exp\left(-\frac{1}{2 \cdot 2504}(41.6 - 0)^2\right)} = 28.6
\end{aligned}
$$

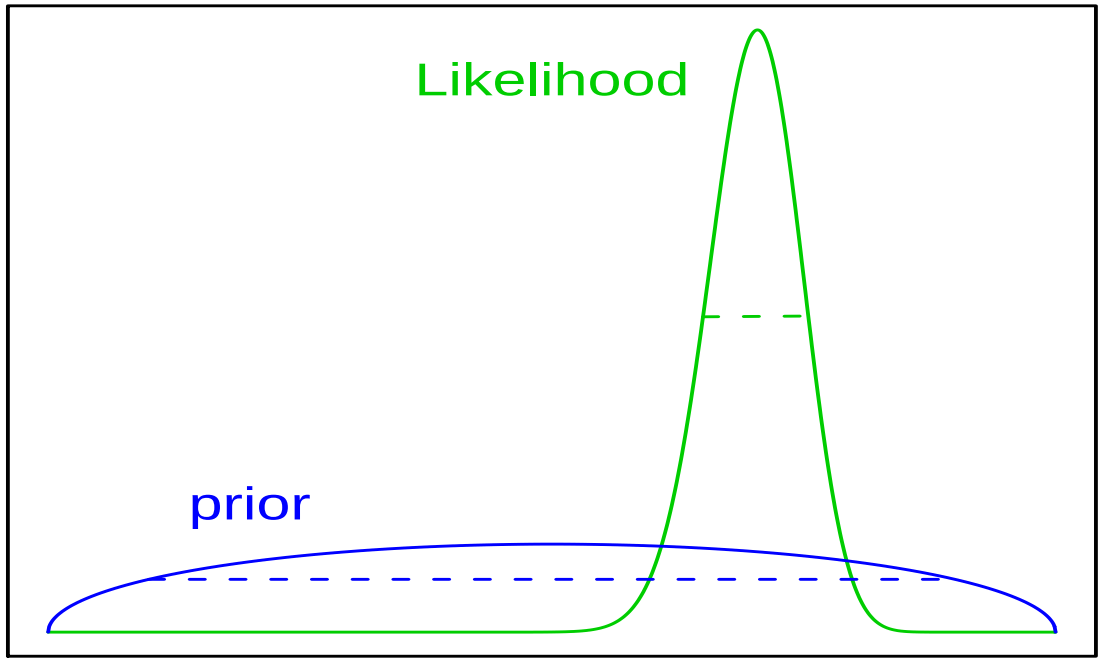
(Note, the lower bound on the Bayes factor over all $\tau^2$ is 27.76. The lower bound over all symmetric nonincreasing priors for $\theta$ is 15.04.)

## Marginal probabilities prefer simple models that fit the data




Note: Einstein had several theories of general relativity, at least one of which he rejected because of lack of fit to the Mercury data. So assign a *multiple hypothesis penalty*, e.g., give each of four theories prior probability 1/4.

# The Occam Factor



For $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ip})$, $\ell(\boldsymbol{\theta}_i) = f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)$,

$$\begin{aligned} m(\boldsymbol{x} \mid H_i) &= \int \pi(\boldsymbol{\theta}_i \mid H_i)\ell(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \\ &\cong \pi(\hat{\boldsymbol{\theta}}_i \mid H_i) \int_{|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i| < \frac{c}{\sqrt{n}}} \ell(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \\ &\cong \ell(\hat{\boldsymbol{\theta}}_i) \times \pi(\hat{\boldsymbol{\theta}}_i \mid H_i)\frac{K_i}{n^{p/2}} \\ &= \text{maximum likelihood } \times \text{ Occam factor} \end{aligned}$$

From the Laplace approximation, $K_i = (2\pi)^{p/2}|I_i(\hat{\boldsymbol{\theta}}_i)|^{-1/2}$, where $I_i(\cdot)$ is the Fisher information for one observation, so

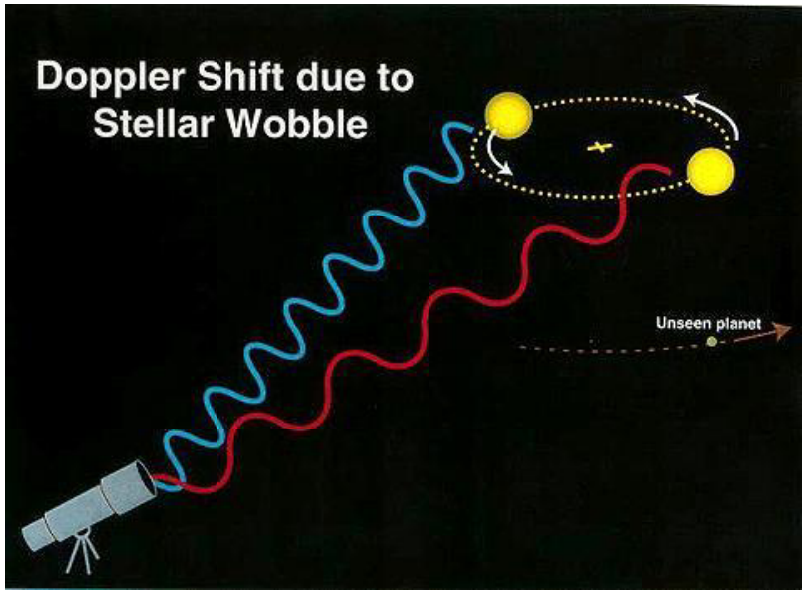
$$B_{01} \approx \frac{\ell(\hat{\boldsymbol{\theta}}_0)}{\ell(\hat{\boldsymbol{\theta}}_1)} \times \frac{\pi(\hat{\boldsymbol{\theta}}_0 \mid H_0)}{\pi(\hat{\boldsymbol{\theta}}_1 \mid H_1)} \cdot \left(\frac{n}{2\pi}\right)^{(p_1 - p_0)/2} \cdot \frac{|I_0(\hat{\boldsymbol{\theta}}_0)|^{-1/2}}{|I_1(\hat{\boldsymbol{\theta}}_1)|^{-1/2}} \,.$$

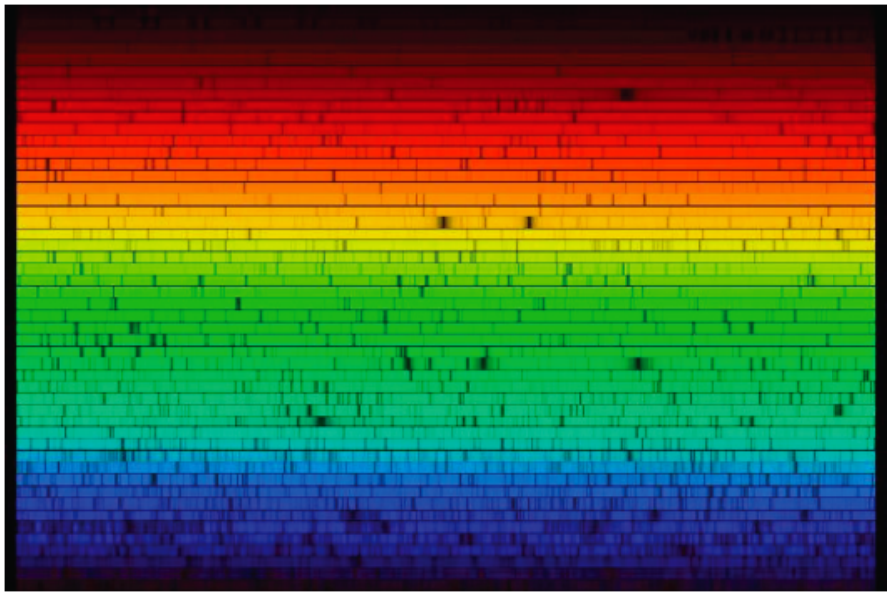
# II. Model selection via the exoplanet illustration

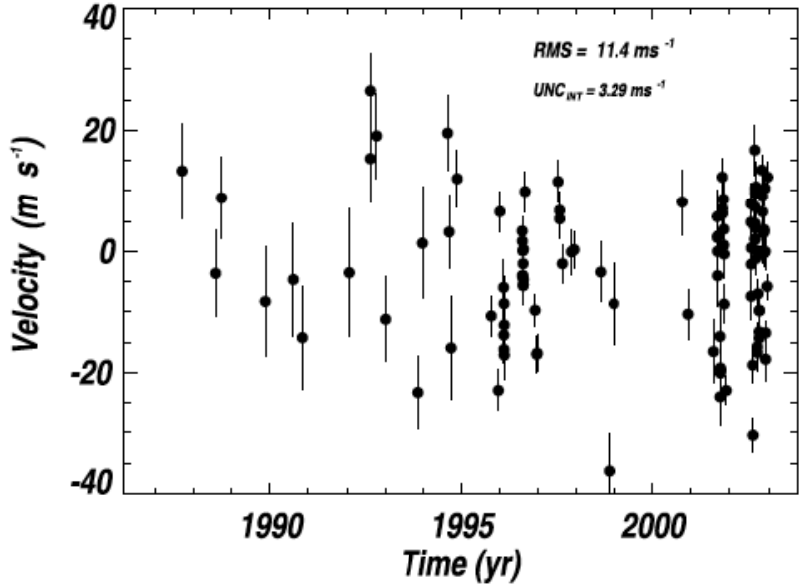
**With Tom Loredo and David Chernoff (Cornell Astronomy)**

**Bin Liu and Merlise Clyde (Duke Statistics)**

# One Detection Method: Use of Radial Velocity





# Keplerian Radial Velocity Model

## Parameters for single planet

- $\tau$ = orbital period (days)
- $e$ = orbital eccentricity
- $K$ = velocity amplitude (m/s)
- Argument of pericenter $\omega$
- Mean anomaly at $t = 0$, $M_0$
- System center-of-mass velocity $v_0$
- Stellar jitter $\sigma_J^2$

**Keplerian reflex velocity vs. time:** Letting $\boldsymbol{\theta} = \{K, \tau, e, M_0, \omega\}$,

$$v(t) = v_0 + g(t; \boldsymbol{\theta}) \equiv v_0 + K \left( e \cos\omega + \cos[\omega + \upsilon(t)] \right) ,$$

where the true anomaly $\upsilon(t)$, from Kepler's equation for eccentric anomaly, is

$$E(t) - e \sin E(t) = \frac{2\pi t}{\tau} - M_0; \qquad \tan\frac{\upsilon}{2} = \left( \frac{1+e}{1-e} \right)^{1/2} \tan\frac{E}{2} .$$

# The Likelihood Function

Keplerian observational velocity model:

Observations: $x_i = v_0 + g(t_i; \boldsymbol{\theta}) + \varepsilon_i + \eta_i$,

where $\varepsilon_i \sim N(\varepsilon_i \mid 0, \sigma_i^2)$ with the $\sigma_i^2$ known are observational errors and $\eta_i \sim N(\eta_i \mid 0, \sigma_J^2)$. (Perhaps better is $N(\varepsilon_i \mid 0, \lambda\sigma_i^2)$, with $\lambda$ unknown.)

Likelihood for data $\boldsymbol{x} = (x_1, \ldots, x_n)$ is

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}, v_0, \sigma_J^2) = \prod_{i=1}^{N} \frac{1}{2\pi\sqrt{\sigma_i^2 + \sigma_J^2}} \exp\left[-\frac{1}{2}\frac{[x_i - v_0 - g(t_i; \boldsymbol{\theta})]^2}{\sigma_i^2 + \sigma_J^2}\right]$$

$$\propto \left[\prod_i \frac{1}{2\pi\sqrt{\sigma_i^2 + \sigma_J^2}}\right] \exp\left[-\frac{1}{2}\chi^2(\boldsymbol{\theta})\right] ,$$

$$\text{where} \quad \chi^2(\boldsymbol{\theta}, \sigma_J^2) \equiv \sum_i \frac{[x_i - v_0 - g(t_i; \boldsymbol{\theta})]^2}{\sigma_i^2 + \sigma_J^2} .$$

This likelihood has extreme multimodality in $\tau$; challenging multimodality in $M_0$; and is smooth (but often vague) in $e$.

# Bayesian Approach to Exoplanet Detection

- Let $M_i$ denote the model that there are $i$ planets ($2 + 5i$ parameters).
- Determine prior distributions $\pi(\boldsymbol{\theta}_i, v_0, \sigma_J^2)$ for the parameters (semi-standard, as the result of a SAMSI program, except for $\sigma_J^2$).
- Compute the marginal likelihood of model $M_i$,

$$m_i(\boldsymbol{x}) = \int f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i, v_0, \sigma_J^2)\pi(\boldsymbol{\theta}_i, v_0, \sigma_J^2)\, d\boldsymbol{\theta}_i\, dv_0\, d\sigma_J^2\,.$$

  We have been working on an adaptive importance sampling algorithm for carrying out the computation.

- Typically look at Bayes factors (the ratio of marginal likelihoods) to determine the number of planets.

# Marginal likelihood computation: importance sampling

**Goal:** Computation of $m(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$.

**Importance sampling:** Choose a proper distribution $q(\boldsymbol{\theta})$, easy to generate from, and such that $q(\boldsymbol{\theta})$ is roughly proportional to $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Then

$$\begin{aligned} m(\boldsymbol{x}) &= \int \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\, q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &\approx \frac{1}{L}\sum_{i=1}^{L} \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \quad \text{with} \quad \boldsymbol{\theta}^{(i)} \sim q(\boldsymbol{\theta})\,. \end{aligned}$$

$q(\boldsymbol{\theta})$ is called the importance function.

**Choice of $q$ is crucial:** It should

- be easy to generate from;
- be roughly proportional to $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$;
- have tails that are somewhat heavier than those of $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

The reason $q(\boldsymbol{\theta})$ can't have too sharp tails is that the variance of the estimate is $V/L$,

$$V = \left( \int \frac{[f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})]^2}{q(\boldsymbol{\theta})} \, d\boldsymbol{\theta} - m^2(\boldsymbol{x}) \right) ,$$

and this may not exist if $q(\boldsymbol{\theta})$ has tails that are too light, which can result in an extremely slow converging algorithm.

**Note:** Assuming this variance is finite, it can be estimated by $\hat{V}/L$,

$$\hat{V} = \left( \frac{1}{L} \sum_{i=1}^{L} \left[ \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \right]^2 - \left[ \frac{1}{L} \sum_{i=1}^{L} \frac{f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})}{q(\boldsymbol{\theta}^{(i)})} \right]^2 \right) .$$

- One of the great advantages of importance sampling is the ease with which one can estimate accuracy.
- Some care is needed: monitor $\hat{V}$ as $L$ increases to make sure it is not increasing.

## Common choices of the importance function:

- If there is little data, choosing $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ is okay, and then $m(\boldsymbol{x}) \approx \frac{1}{L} \sum_{i=1}^{L} f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})$.
- If there is a lot of data and $\pi(\boldsymbol{\theta})$ does not have sharp tails (e.g. is Cauchy) choosing $q(\boldsymbol{\theta}) \propto f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is okay, if the likelihood is easy to generate from (e.g., is normal).
- A common choice is $q(\boldsymbol{\theta}) = N(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{I}})$ where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{I}}$ are the mle and observed Fisher information matrix for $\boldsymbol{\theta}$.
  - But the normal distribution has too sharp tails, so a much better choice is a $t$-distribution with, say, four degrees of freedom.
  - Better yet is $q(\boldsymbol{\theta}) = T_4(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}, c\hat{\boldsymbol{I}})$, and try different $c > 1$ until convergence is fast.
  - Better yet is $q(\boldsymbol{\theta}) = T_4(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}_\pi, c\hat{\boldsymbol{I}}_\pi)$, where $\hat{\boldsymbol{\theta}}_\pi$ and $\hat{\boldsymbol{I}}_\pi$ are the maximizer and Hessian of $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

# The Adaptive Importance Sampler

- It has an overall annealing layer, wherein one, as temperature $t \to 0$, is attempting to target $[f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})]^{1-t}$. (This is to try to find the modes of the integrand.)
- The importance function $q_t(\boldsymbol{\theta})$ tries to mimic the target with a mixture of $T_4$ densities, $q_t(\boldsymbol{\theta}) = \sum_{j=1}^{k} w_j T_4(\boldsymbol{\theta} \mid \boldsymbol{\mu}_j, \Sigma_j)$.
- Samples $\boldsymbol{\theta}^{(i)}$ are drawn from $q_t(\boldsymbol{\theta})$, and examined for high ratios of

$$\frac{[f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)})]^{1-t}}{q_t(\boldsymbol{\theta}^{(i)})};$$

  new components of the mixture may be added at those points.
- If a weight of a component in the mixture becomes too small, the component is dropped.
- The weights of the mixture and the mean and covariance matrices are chosen by an analytic fit to the annealing target, using K-L divergence estimated from the previous draws from $q_t(\boldsymbol{\theta})$.

# Example Exoplanet Results

## I. HD73526, 18 observations

| | Marginal Likelihood | ESS/$N$ |
|---|---|---|
| $\mathcal{M}_0$ | $5.9013 \times 10^{-50} \pm 5.1325 \times 10^{-52}$ | 0.9320 |
| $\mathcal{M}_1$ | $4.4886 \times 10^{-41} \pm 3.2093 \times 10^{-42}$ | 0.5698 |
| $\mathcal{M}_2$ | $1.5511 \times 10^{-42} \pm 3.2878 \times 10^{-43}$ | 0.3458 |

| BayesFactor$\{\mathcal{M}_1 : \mathcal{M}_0\}$ | BayesFactor$\{\mathcal{M}_2 : \mathcal{M}_1\}$ |
|---|---|
| $7.606 \times 10^8$ | 0.03456 |

## II. HD73526, 30 observations

| | Marginal Likelihood | ESS/$N$ |
|---|---|---|
| $\mathcal{M}_0$ | $8.9566 \times 10^{-77} \pm 1.0852 \times 10^{-78}$ | 0.9510 |
| $\mathcal{M}_1$ | $5.8519 \times 10^{-70} \pm 1.7077 \times 10^{-71}$ | 0.6545 |
| $\mathcal{M}_2$ | $4.8122 \times 10^{-65} \pm 1.4284 \times 10^{-66}$ | 0.2034 |

| BayesFactor$\{\mathcal{M}_1 : \mathcal{M}_0\}$ | BayesFactor$\{\mathcal{M}_2 : \mathcal{M}_1\}$ |
|---|---|
| $6.534 \times 10^6$ | $8.233 \times 10^4$ |

## III. 47 Ursae Majoris

| | Marginal Likelihood | ESS/$N$ |
|---|---|---|
| $\mathcal{M}_0$ | $2.0198 \times 10^{-1004} \pm 9.2572 \times 10^{-1006}$ | 0.1002 |
| $\mathcal{M}_1$ | $3.4400 \times 10^{-896} \pm 3.10 \times 10^{-897}$ | 0.5643 |
| $\mathcal{M}_2$ | $1.3500 \times 10^{-816} \pm 1.77 \times 10^{-817}$ | 0.3324 |
| $\mathcal{M}_3$ | $2.8970 \times 10^{-825} \pm 9.1623 \times 10^{-825}$ | 0.0089 |

| BayesFactor$\{\mathcal{M}_1 : \mathcal{M}_0\}$ | BayesFactor$\{\mathcal{M}_2 : \mathcal{M}_1\}$ | BayesFactor$\{\mathcal{M}_3 : \mathcal{M}_2\}$ |
|---|---|---|
| $1.703 \times 10^{108}$ | $3.924 \times 10^{79}$ | ? |

# III. Introduction to Bayesian multiplicity control

(with James Scott)

**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.'

# Bayesian Multiplicity Control

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

**Example: multiple testing of exclusive hypotheses**

Suppose one is testing mutually exclusive hypotheses $H_i,\ i = 1, \ldots, m$, 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 $\boldsymbol{x}$.

- If $m_i(\boldsymbol{x}) = \int f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$ is the marginal density of the data under $H_i$, the posterior probability of $H_i$ is

$$Pr(H_i \mid \boldsymbol{x}) = \frac{\frac{1}{m}m_i(\boldsymbol{x})}{\sum_{j=1}^{m}\frac{1}{m}m_j(\boldsymbol{x})} = \frac{m_i(\boldsymbol{x})}{\sum_{j=1}^{m}m_j(\boldsymbol{x})}\,.$$

- Thus the likelihood $m_i(\boldsymbol{x})$ 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 $Pr(H_i \mid \boldsymbol{x})$ is the posterior probability in a multiplicity scenario without null control, adding null control results in a new posterior probability

$$Pr^*(H_i \mid \boldsymbol{x}) \approx \frac{Pr(H_i \mid \boldsymbol{x})}{1 + Pr(H_0 \mid \boldsymbol{x})/Pr(H_0)} \, .$$

*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 $\mu_i$ 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)
  $Pr(H_5^1 \mid \boldsymbol{x}) = \frac{m_5(\boldsymbol{x})}{\sum_{j=1}^{1000} m_j(\boldsymbol{x})} = 0.019$ (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 $Pr(H_5^1 \mid \boldsymbol{x}) = 0.019$.
- If null control was employed but *apriori* the physicist decided to use all of the non-null mass on $H_5$, the *societal* answer would have *legitimately* been $Pr(H_5^1 \mid \boldsymbol{x}) = 0.96$.

*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:* $X_i, i = 1, ..., m$, are $N(x_i \mid \mu_i, 1, \rho)$, $\rho$ being the correlation.

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

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

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

  where $\Phi$ 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 $\rho \to 1$ (the one-dimensional solution).

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

- Suppose $x_i \sim N(x_i \mid \mu_i, 1)$, $i = 1, \ldots, m$, are observed.
- It is desired to test $H_i^0 : \mu_i = 0$ versus $H_i^1 : \mu_i \neq 0$ , $i = 1, \ldots, m$, 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 $Pr(\mu_1 = \mu_2 = \ldots = \mu_m = 0) = 2^{-m}$.

*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 $\mathcal{M}_{\boldsymbol{\gamma}}$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$, with $\gamma_i$ 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 $Pr(\mathcal{M}_{\mathbf{0}}) = Pr(\mu_1 = \mu_2 = \ldots = \mu_m = 0) = 1/2$, and $Pr(\mathcal{M}_{\boldsymbol{\gamma}}) = \frac{1}{2(2^m - 1)}$ otherwise.
- The previous unadjusted $Pr(H_i^1 \mid x_i)$ then becomes (using the cute fact)

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

  - *Case 1.* $\mathcal{M}_{\mathbf{0}}$ *is true:* Then $Pr(\mathcal{M}_{\mathbf{0}} \mid \boldsymbol{x})/Pr(\mathcal{M}_{\mathbf{0}}) = \mathrm{O}(e^{cm})$ for some $c$, so there is very strong null (and multiplicity) control.
  - *Case 2. All of the null hypotheses are true, except one where there is a very large* $x_j$*:* Then $Pr(\mathcal{M}_{\mathbf{0}} \mid \boldsymbol{x})/Pr(\mathcal{M}_{\mathbf{0}}) \approx 0$, 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 all. (01), Newton and Kendziorski (03), Müller et al. (03), Guindani, M., Zhang, S. and Mueller, P.M. (2007), ...; many empirical Bayes such as Efron and Tibshirani (2002), Storey, J.D., Dai, J.Y and Leek, J.T. (2007), Efron (2010))

- Suppose $x_i \sim N(x_i \mid \mu_i, \sigma^2)$, $i = 1, \ldots, m$, are observed, $\sigma^2$ 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 $\mu_i$ 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

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

- $(p_1, p_2, \ldots, p_m)$ 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.

| | The ten 'signal' observations | | | | | | | | | | #noise |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | -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 $m \to \infty$, so the Bayesian procedure exerts medium-strong control over false positives. (In comparison, $E[\#i :$ Bonferroni rejects $\mid$ all $\mu_j = 0] = \alpha$.)

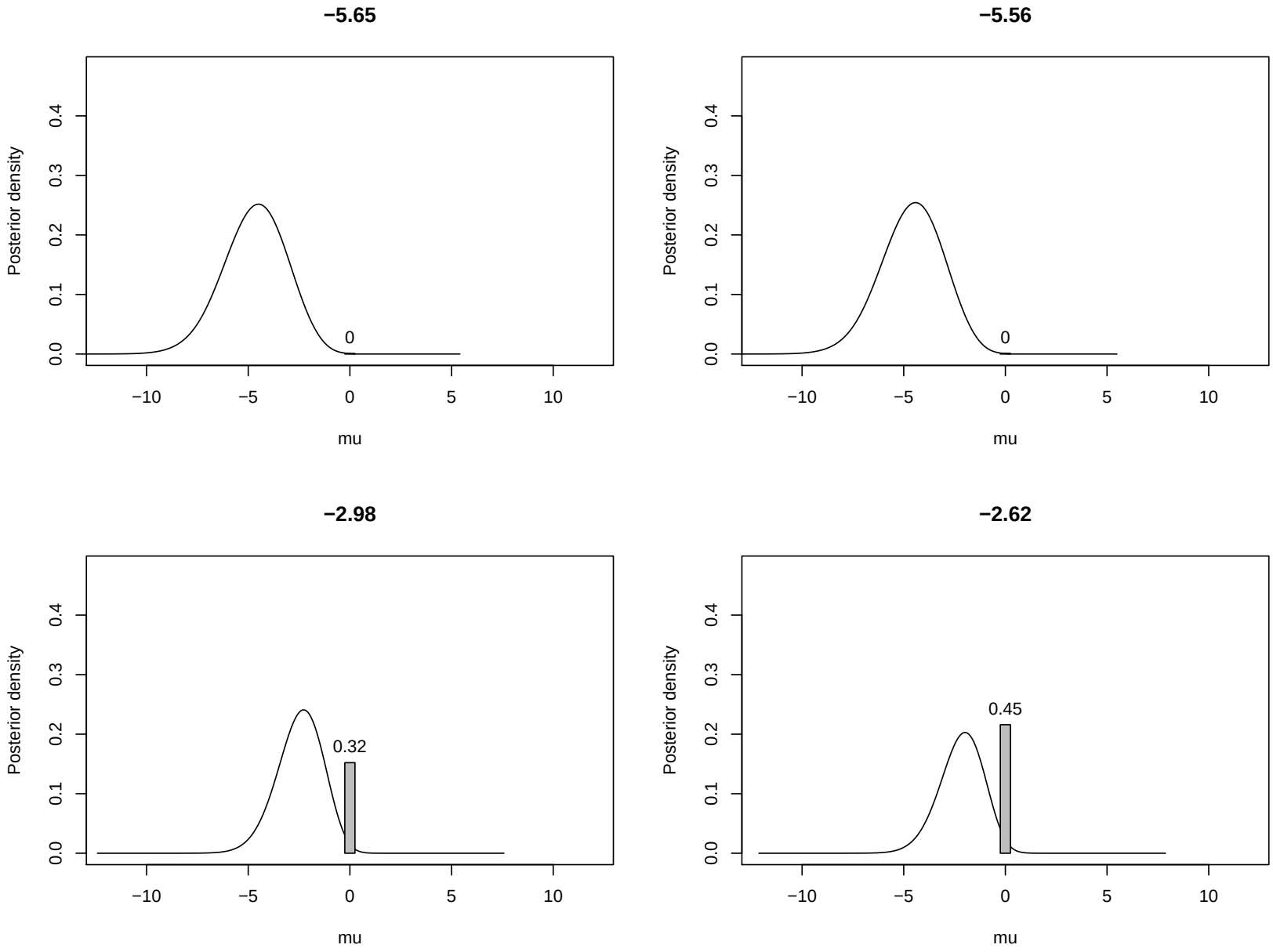



Figure 1: For four of the observations, $1 - p_i = \Pr(\mu_i = 0 \,|\, \boldsymbol{y})$ (the vertical bar), and the posterior densities for $\mu_i \neq 0$ .

## 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

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

# IV. Bayesian Calibration of $p$-values from Vovk (1993 JRSSB) and Sellke, Bayarri and Berger (2001 American Statistician)

- A *proper* $p$-value satisfies $H_0 : p(X) \sim \text{Uniform}(0, 1)$.
- Test versus $H_1 : p(X) \sim \text{Beta}(\xi, 1)$, $0 < \xi < 1$. Then, when $p < e^{-1}$,

$$B_{01}(p) = \frac{1}{\xi p^{(\xi-1)}} \geq -e\, p \log(p)\,.$$

  This bound also holds for any alternative $f(p)$, where $Y = -\log(p)$ has a decreasing failure rate (natural non-parametric alternatives).

- The corresponding bound on the conditional Type I frequentist error is

$$\alpha \geq (1 + [-e\, p \log(p)]^{-1})^{-1}\,.$$

| $p$ | .2 | .1 | .05 | .01 | .005 | .001 | .0001 | .00001 |
|---|---|---|---|---|---|---|---|---|
| $-ep\log(p)$ | .879 | .629 | .409 | .123 | .072 | .0189 | .0025 | .00031 |
| $\alpha(p)$ | .465 | .385 | .289 | .111 | .067 | .0184 | .0025 | .00031 |

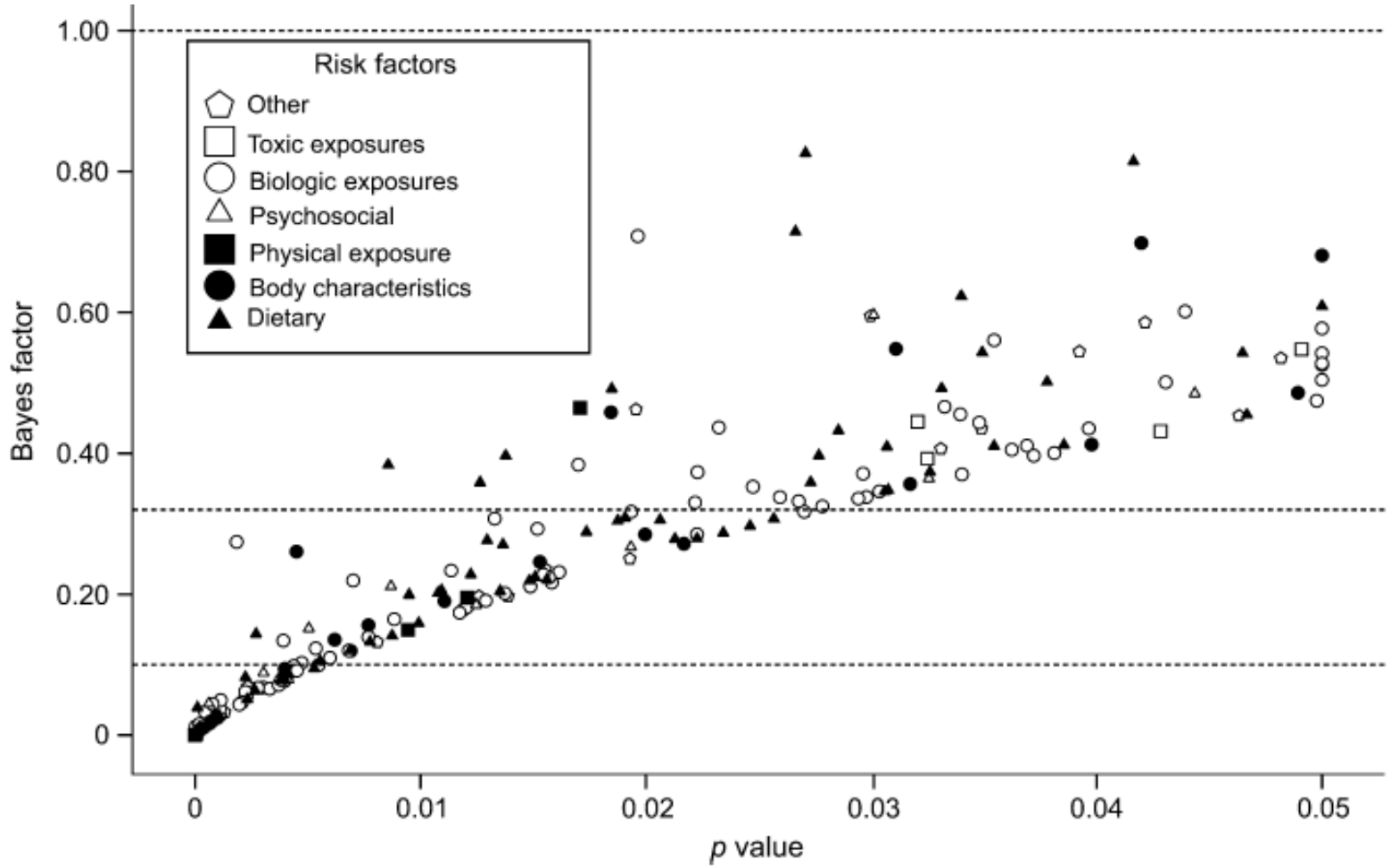



**FIGURE 1.** Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each study. Shown are calculations assuming $\theta_A$ of 0.50 (relative risk = 1.65). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

Figure 2: J.P. Ioannides: Am J Epidemiol 2008;168:374–383

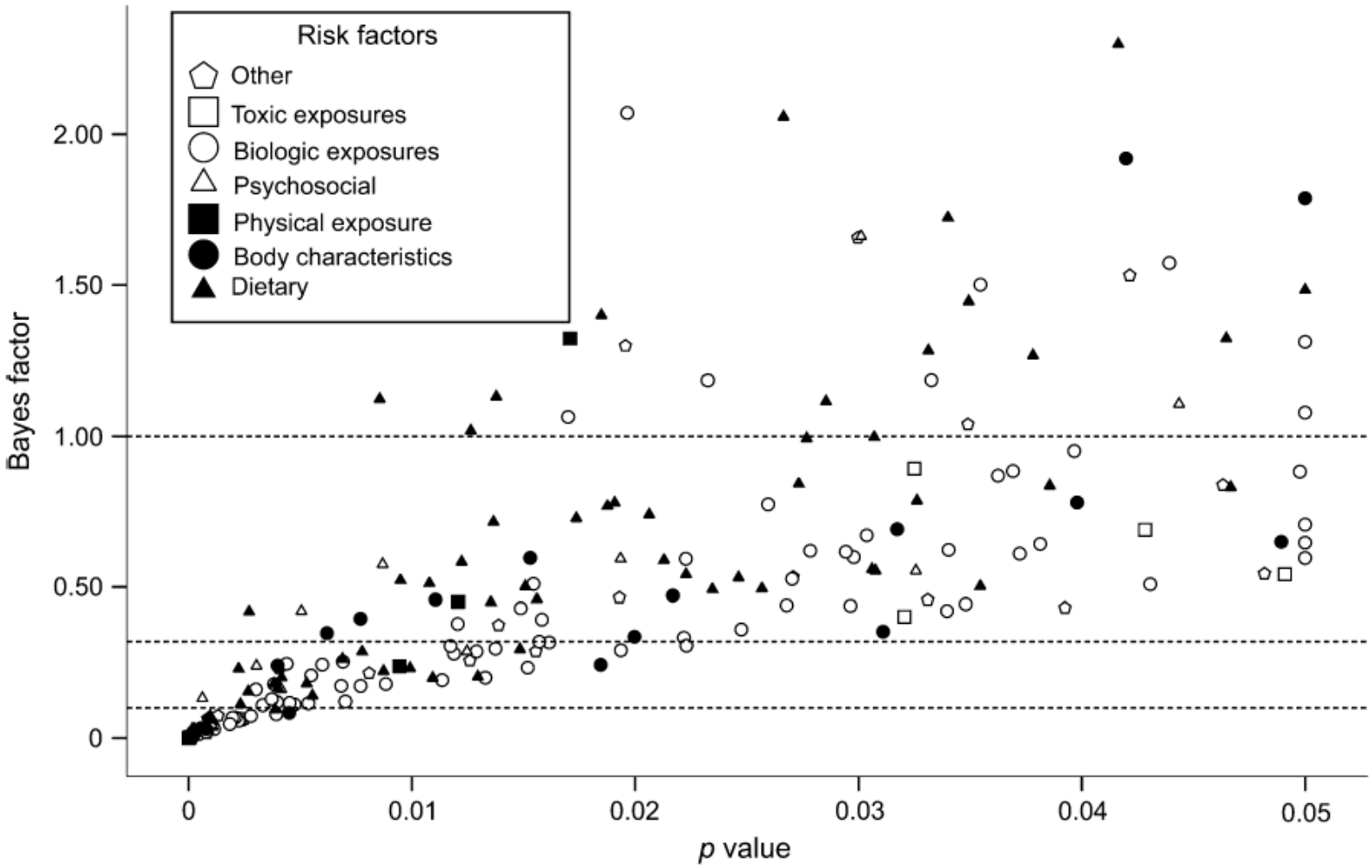



**FIGURE 2.** Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each study. Shown are calculations assuming $\theta_A$ of 1.50 (relative risk = 4.48). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

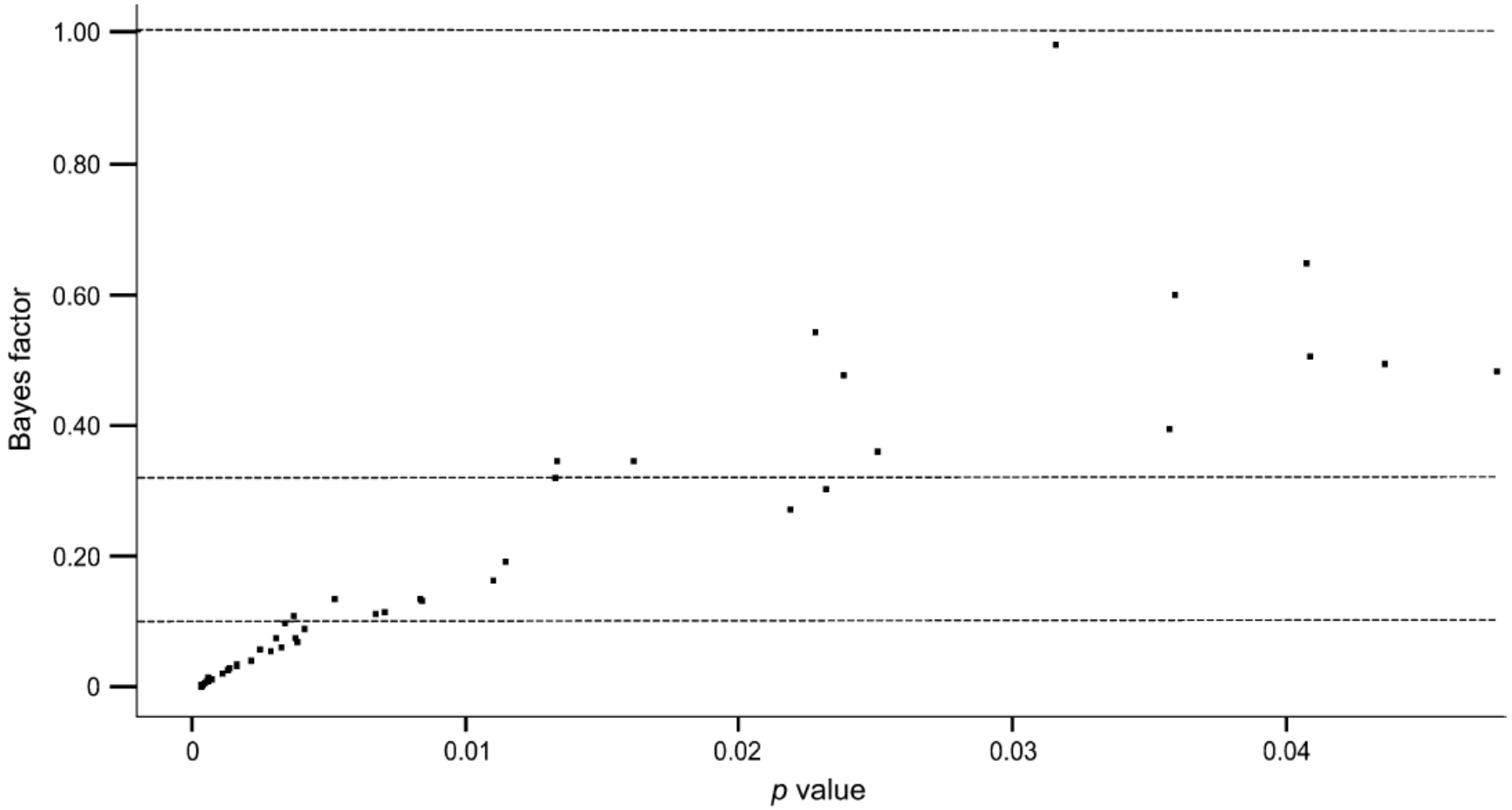



**FIGURE 3.** Estimated Bayes factors for 50 meta-analyses of genetic associations with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each meta-analysis. Calculations assume $\theta_A$ equal to the median relative risk observed in the 50 genetic associations (relative risk = 1.44). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

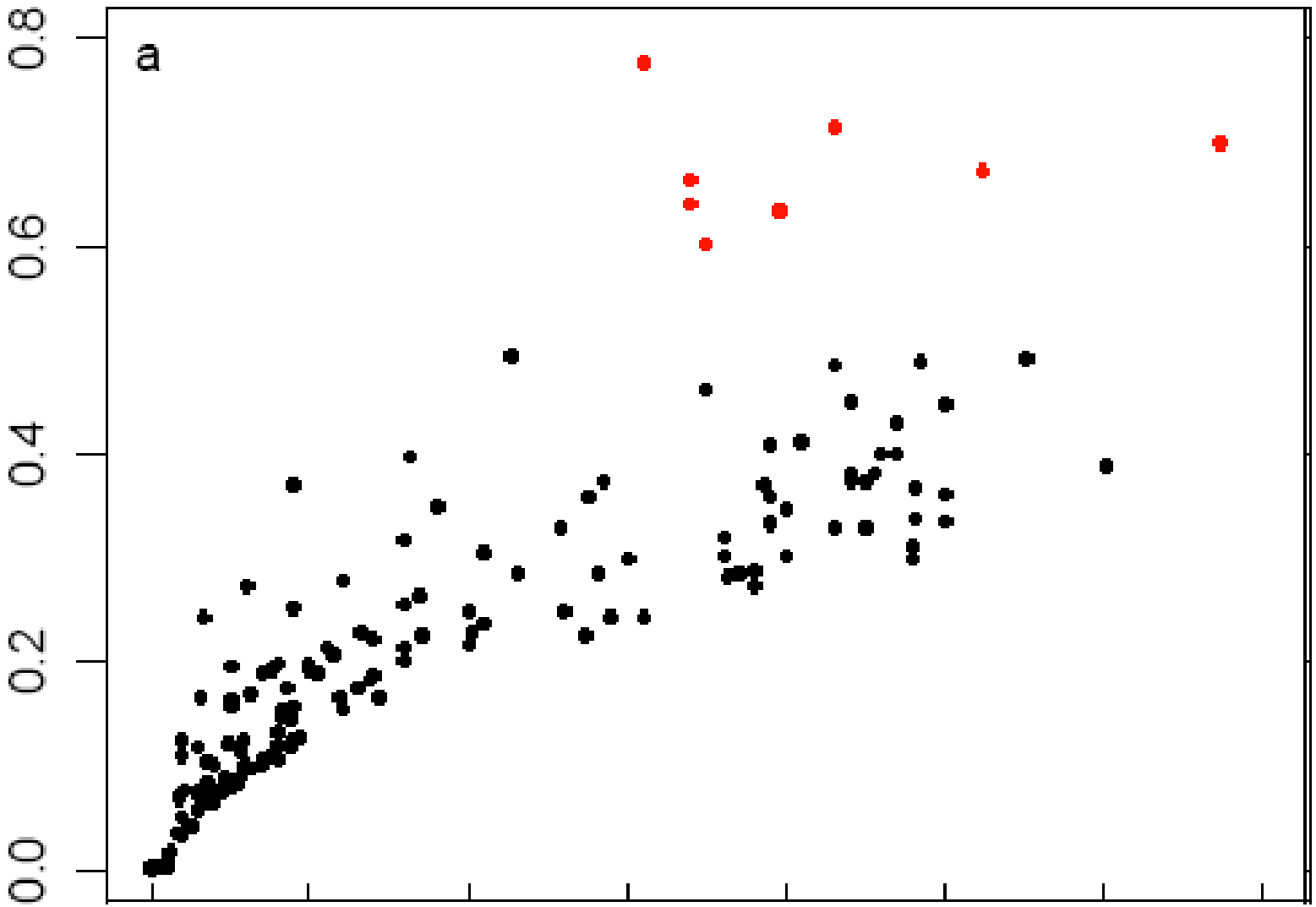



Figure 3: Elgersma and Green (2011): $\alpha(p)$ versus observed $p$-values for 314 articles in Ecology in 2009.

# V. Bayesian/frequentist duality

- Estimation with objective priors
- Frequentist and Bayesian odds

# Estimation with objective priors

*Bayesian estimation procedures based on objective priors (e.g.,* reference priors*) are excellent and often optimal frequentist procedures.*

**A psychiatry diagnosis example** (with Mossman, 2001):

- Within a population, $p_0 = Pr(\text{Disease } D)$.
- A diagnosic test results in either a Positive (P) or Negative (N) reading.
- $p_1 = Pr(P \,|\, \text{patient has } D)$.
- $p_2 = Pr(P \,|\, \text{patient does not have } D)$.
- It follows from Bayes theorem that

$$\theta \equiv Pr(D|P) = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2}.$$

**The Statistical Problem:** The $p_i$ are unknown. Based on (independent) data $X_i \sim \text{Binomial}(n_i, p_i)$ (arising from medical surveys of $n_i$ individuals), find a $100(1-\alpha)\%$ confidence set for $\theta$.

**Suggested Solution:** Assign $p_i$ the Jeffreys-rule objective prior

$$\pi(p_i) \propto p_i^{-1/2}(1-p_i)^{-1/2}\,.$$

By Bayes theorem, the posterior distribution of $p_i$ given the data, $x_i$, is

$$\pi(p_i \mid x_i) = \frac{p_i^{-1/2}(1-p_i)^{-1/2} \times n x_i p_i^{x_i}(1-p_i)^{n_i - x_i}}{\int p_i^{-1/2}(1-p_i)^{-1/2} \times n x_i p_i^{x_i}(1-p_i)^{n_i - x_i}\, dp_i},$$

which is the Beta$(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ distribution.

Finally, compute the desired confidence set (formally, the $100(1-\alpha)\%$ equal-tailed posterior credible set) through Monte Carlo simulation from the posterior distribution by

- drawing random $p_i$ from the Beta$(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ posterior distributions, $i = 0, 1, 2$;
- computing the associated $\theta = p_0 p_1 / [p_0 p_1 + (1 - p_0) p_2]$;
- repeating this process $10,000$ times, yielding $\theta_1, \theta_2, \ldots, \theta_{10,000}$;
- using the $\frac{\alpha}{2}\%$ upper and lower percentiles of these generated $\theta$ to form the desired confidence limits.

| $n_0 = n_1 = n_2$ | $(x_0, x_1, x_2)$ | 95% confidence interval |
|---|---|---|
| 20 | (2,18,2) | (0.107, 0.872) |
| 20 | (10,18,0) | (0.857, 1.000) |
| 80 | (20,60,20) | (0.346, 0.658) |
| 80 | (40,72,8) | (0.808, 0.952) |

Table 2: The 95% equal-tailed posterior credible interval for $\theta = p_0 p_1 / [p_0 p_1 + (1 - p_0) p_2]$, for various values of the $n_i$ and $x_i$.

**Unconditional frequentist performance of the objective Bayes procedure:** The goal was to find confidence sets for

$$\theta = Pr(D \mid P) = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2} \, .$$

Consider the frequentist percentage of the time that the 95% Bayesian sets miss on the left and on the right (ideal would be 2.5%) for the indicated parameter values when $n_0 = n_1 = n_2 = 20$.

| $(p_0, p_1, p_2)$ | O-Bayes | Log Odds | Gart-Nam | Delta |
|---|---|---|---|---|
| $(\frac{1}{4}, \frac{3}{4}, \frac{1}{4})$ | 2.86,2.71 | 1.53,1.55 | 2.77,2.57 | 2.68,2.45 |
| $(\frac{1}{10}, \frac{9}{10}, \frac{1}{10})$ | 2.23,2.47 | 0.17,0.03 | 1.58,2.14 | 0.83,0.41 |
| $(\frac{1}{2}, \frac{9}{10}, \frac{1}{10})$ | 2.81,2.40 | 0.04,4.40 | 2.40,2.12 | 1.25,1.91 |

*Conclusion:* By construction, reasonable 'Bayesian credibility' is guaranteed; unconditional frequentist performance is clearly fine (and the expected lengths of the Bayesian intervals were smallest).

# Frequentist and Bayesian Odds:

Let $\alpha$ and $(1-\beta(\theta))$ be the Type I error and power for testing $H_0$ versus $H_1$ with rejection region $\mathcal{R}$. Then

$$\begin{aligned} O &= \text{Pre-experimental odds of } \textit{correct rejection} \text{ to } \textit{incorrect rejection} \\ &= [\text{prior odds of } H_1 \text{ to } H_0] \times \frac{(1-\bar{\beta})}{\alpha}\,, \end{aligned}$$

where $(1-\bar{\beta}) = \int (1-\beta(\theta))\pi(\theta)d\theta$ is average power wrt the prior $\pi(\theta)$. Then

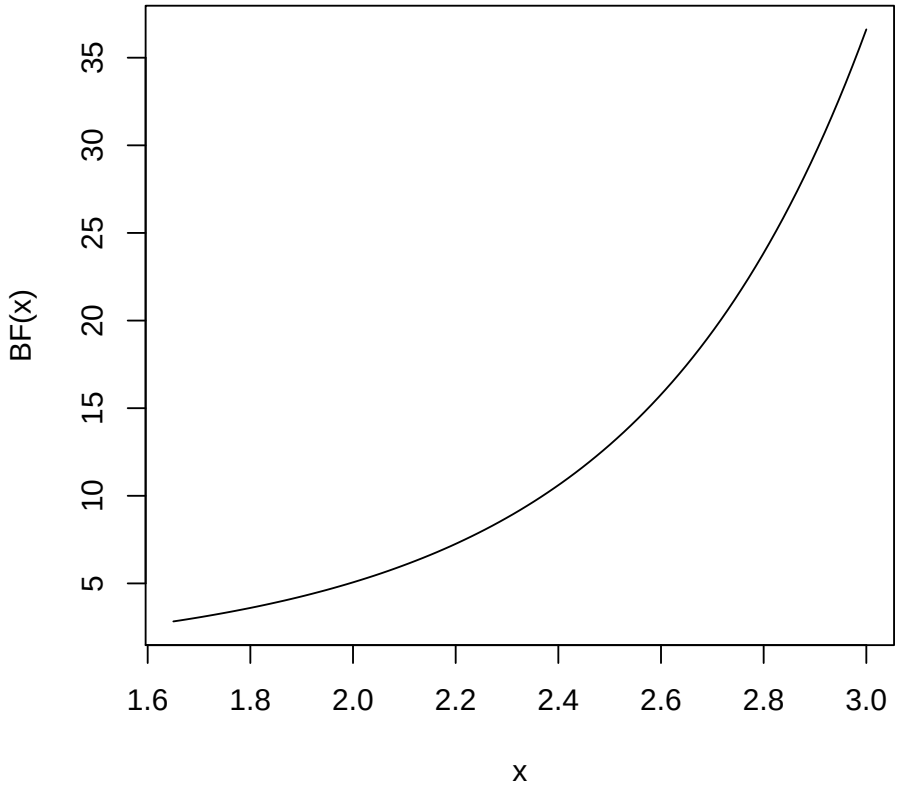
$$\frac{(1-\bar{\beta})}{\alpha} = \frac{\text{average power}}{\text{type 1 error}}$$

can be viewed as the *experimental odds* of correct rejection to incorrect rejection.

| average power | 0.05 | 0.25 | 0.50 | 0.75 | 1.0 | 0.01 | 0.25 | 0.50 | 0.75 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| type I error | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| correct/incorrect | 1 | 5 | 10 | 15 | 20 | 1 | 25 | 50 | 75 | 100 |

But that is pre-experimental; better is to report the actual data-based odds of correct rejection to incorrect rejection, namely the Bayes factor $B_{10}(z)$.

For an example in which $\frac{(1-\bar{\beta})}{\alpha} = 9$, $B_{10}(z)$ was




**The Bayes/frequentist duality:** For simple nulls (or nulls that are simple for the test statistic) $E[B_{10}(Z) \mid H_0, \mathcal{R}] = \frac{(1-\bar{\beta})}{\alpha}$, so reporting $B_{10}(z)$ is a valid conditional frequentist procedure. (Kiefer, 1977 JASA; Brown, 1978 AOS)

Thanks!