Lecture 10: Odds and Ends and Applications

Jim Berger

Duke University

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

Case Studies and Discussion

- Ockham's Razor
- Selecting a single model for prediction
- More on p-values and their calibration
- Application: Cepheid variable stars
- Application: HIV vaccine
- Application: Psychokinesis

I. Ockham's Razor

Ockham's Razor

- attributed to 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.)

- preferring the simpler of two hypothesis to the more complex when both agree with data is an old principle in science
- 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 allow them to be tweaked to accommodate a wider variety of data
  - "coin is fair" is a simpler model than "coin has unknown bias θ"
  - $s = a + ut + \frac{1}{2}\,gt^2$ 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.

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 $(2 + \epsilon)$.

All these hypotheses had a parameter that could be adjusted to deal with whatever data on the motion of Mercury existed.

Data: $X = 41.6$ where $X \sim N(\theta \mid 2^2)$, θ being the perihelion advance of Mercury; the standard deviation of the measurement was 4.

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

- Symmetric about 0 (corresponding 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.

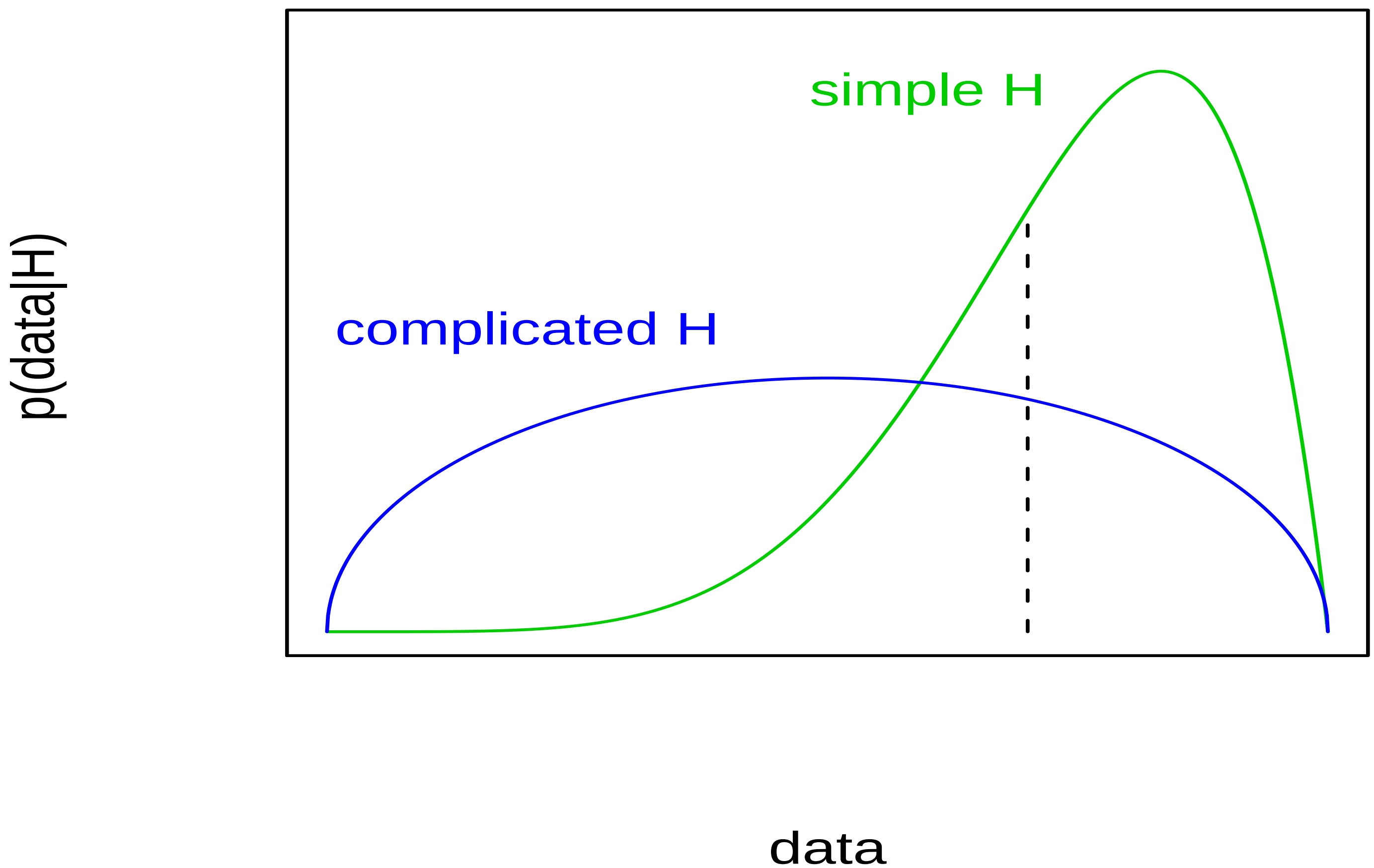
General Relativity (1915) model M_E: Predicted $\theta_E = 42.9$.

Bayes factor:

$$
\begin{aligned}
B_{EG} &= \frac{f_E(41.6)}{\int f_G(x \mid \theta)\pi_G(\theta)}\ d\theta \\
&= \frac{\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6-\theta_E)^2\right)}{\int \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6-\theta)^2\right) \frac{1}{50\sqrt{2\pi}} \exp\left(-\frac{1}{2\cdot 50^2}\theta^2\right)\ d\theta} \\
&= \frac{\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{8}(41.6-42.9)^2\right)}{\frac{1}{\sqrt{2\cdot 2504\pi}} \exp\left(-\frac{1}{2\cdot 2504}(41.6-0)^2\right)} = 28.6
\end{aligned}
$$

(Note, the lower bound on the Bayes factor over all τ^2 is 27.76. The lower bound over all symmetric nonincreasing priors for θ is 15.04.)

predictive probabilities prefer simple models



II. Selecting a Single Model for Prediction

(work with **M. Barbieri**)

Context: Prediction with Normal linear models

- Observe the $n \times 1$ vector

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,;$$

where $\mathbf{X}$ is the $n \times k$ design matrix, $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown coefficients, and $\boldsymbol{\epsilon}$ is $\mathcal{N}(0, \sigma^2 I)$.

- Choose from among submodels

$$M_{\mathbf{l}} : \boldsymbol{y} = \mathbf{X}_{\mathbf{l}}\,\boldsymbol{\beta}_{\mathbf{l}} + \boldsymbol{\epsilon}\,,$$

where $\mathbf{l} = (l_1, l_2, \ldots, l_k)$ is the model index, l_i being either 1 or 0 as covariate x_i is in or out of the model.

Basics of Bayesian prediction

- The goal is to predict a future $y^* = \mathbf{x}^* \boldsymbol{\beta} + \epsilon$, using squared error loss $(y^* - \widehat{y}^*)^2$.
- Combining the data and prior yields, for all $\mathbf{l}$,
  - $P(M_{\mathbf{l}} \mid \boldsymbol{y})$, the posterior probability of model $M_{\mathbf{l}}$;
  - $\pi^*_{\mathbf{l}}(\boldsymbol{\beta}_{\mathbf{l}}, \sigma \mid \boldsymbol{y})$, the posterior distribution of $(\boldsymbol{\beta}_{\mathbf{l}}, \sigma)$.
- The best predictor of y^* is, via *model averaging*,

$$\bar{\boldsymbol{y}}^* = \mathbf{x}^* \, \bar{\boldsymbol{\beta}} \equiv \mathbf{x}^* \sum_{\mathbf{l}} P(M_{\mathbf{l}} \mid \boldsymbol{y}) \, \widetilde{\boldsymbol{\beta}}_{\mathbf{l}} \, ,$$

where $\widetilde{\boldsymbol{\beta}}_{\mathbf{l}}$ is the posterior mean for $\boldsymbol{\beta}$ under $M_{\mathbf{l}}$.

Selecting a single model

- Often a single model is desired for prediction.
- A common misperception is that the best single model is that with the largest $P(M_{\mathbf{l}} \mid \boldsymbol{y})$;
  - however, this is true if there are only two models;
  - and it is true if $\mathbf{X}^{'}\mathbf{X}$ is diagonal, σ^2 is known, and suitable priors are used (Clyde and Parmigiani, 1996).
- The best single model will typically depend on $\mathbf{x}^*$.
- An important case is when the future covariates are like the past covariates, i.e., when $E[(\mathbf{x}^*)'\mathbf{x}^*] = \mathbf{X}^{'}\mathbf{X}$.

Posterior inclusion probabilities

The *posterior inclusion probability* for variable i is

$$p_i \equiv \sum_{\boldsymbol{l}\,:\,l_i=1} P(M_{\boldsymbol{l}} \mid \boldsymbol{y}),$$

i.e., the overall posterior probability that variable i is in the model.

These are of considerable independent interest

- as basic quantities of interest,
- as aids in searches of model space,
- in defining the median probability model.

The (posterior) median probability model

If it exists, the *median probability model*, $M_{\boldsymbol{l}^*}$, is defined to be the model consisting of those variables whose posterior inclusion probability is at least 1/2. Formally, $\boldsymbol{l}^*$ is defined, coordinatewise, by

$$l_i^* = \begin{cases} 1 & \text{if} \quad p_i \geq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Note: If computation is done by a model-jumping MCMC, the median probability model consists of those coordinates that were present in over half the iterations.

Existence of the median probability model

The median probability model exists when the models under consideration follow a *graphical model structure*, including

- when any subset of variables is allowed;
- the situation in which the allowed variables consist of main effects and interactions, but a higher order interaction is allowed only if lower order interactions are included;
- a sequence of nested models, such as arises in polynomial regression and autoregressive time series.

Example (Polynomial Regression): Model j is

$$y = \sum_{i=0}^{j} \beta_i \, x^i + \epsilon.$$

(Model) j	0	1	2	3	4	5	6
$P(M_j \mid \boldsymbol{y})$	~ 0	0.06	0.22	0.29	0.38	0.05	~ 0

(Covariate) i	0	1	2	3	4	5	6
$P(x^i$ is in model $\mid \boldsymbol{y})$	1	1	0.94	0.72	0.43	0.05	0

Thus M_3 is the median probability (optimal predictive) model, while M_4 is the maximum probability model.

Three optimality theorems

Theorem 1. If (i) the models under consideration have graphical structure; (ii) $\mathbf{X}'\mathbf{X}$ is diagonal, and (iii) the posterior mean of $\boldsymbol{\beta_l}$ is simply the relevant coordinates of $\widetilde{\boldsymbol{\beta}}$ (the posterior mean in the full model), then the best predictive model is the median probability model. Condition (iii) is satisfied under any mix of

- noninformative priors for the β_i;
- independent $\mathcal{N}(0, \sigma^2\lambda_i)$ priors for the β_i, with the λ_i given (objectively or subjectively specified, or estimated via empirical Bayes) and any prior for σ^2.

Corollary (Clyde and Parmigiani, 1996). If any submodel of the full model is allowed, $\mathbf{X}'\mathbf{X}$ is diagonal, $\mathcal{N}(0, \sigma^2\lambda_i)$ priors are used for the β_i, with the λ_i given and σ^2 known, and the prior probabilities of the models satisfy

$$P(M_{\boldsymbol{l}}) = \prod_{i=1}^{k} (p_i^0)^{l_i} (1 - p_i^0)^{(1-l_i)} ,$$

where p_i^0 is the prior probability that variable x_i is in the model, then the optimal predictive model is that with highest posterior probability (which is also the median probability model).

Theorem 2. Suppose a sequence of nested linear models is under consideration. If (i) prediction is desired at ‘future covariates like the past’ and (ii) the posterior mean under $M_{\boldsymbol{l}}$ satisfies $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{l}} = b\,\widehat{\boldsymbol{\beta}}_{\boldsymbol{l}}$, where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{l}}$ is the least squares estimate, then the best predictive model is the median probability model.

Condition (ii) is satisfied if we use either

- noninformative priors for model parameters; or
- g-type $\mathcal{N}_{k_{\boldsymbol{l}}}(\mathbf{0}, c\,\sigma^2\,(\mathbf{X}_{\boldsymbol{l}}^{\prime}\mathbf{X}_{\boldsymbol{l}})^{-1})$ priors, with the same constant $c > 0$ for each model, and any prior for σ^2.

Theorem 3. Theorems 1 and 2 essentially remain true even if there are non-orthogonal nuisance parameters (i.e., parameters common to all models) that are assigned the usual noninformative priors.

Nonparametric Regression

- $y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2).$
- Represent f via a (orthonormal) series expansion

$$f(x) = \sum_{j=1}^{\infty} \beta_j \, \phi_j(x).$$

- *Base* prior distribution: $\beta_i \sim N(0, v_i)$, with $v_i = \frac{c}{i^a}$, where c is unknown and a is specified.
- The model M_j, for $j = 1, 2, \ldots, n$, is given by:

$$y_i = \sum_{k=1}^{j} \beta_k \, \phi_k(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

- Choose equal prior probabilities for the models $M_j, \;\; j = 1, 2, \ldots n$. Within M_j, use the base prior to induce the prior distributions for $\boldsymbol{\beta}_j = (\beta_1, \ldots, \beta_j)$.

- For the data $\boldsymbol{y} = (y_1, \ldots, y_n)$, compute $P(M_j \mid \boldsymbol{y})$, the posterior probability of model M_j, for $j \leq n$.
- Within M_j, predict $\boldsymbol{\beta}_j$ by its posterior mean, $\tilde{\boldsymbol{\beta}}_j$.

Example 1. The Shibata Example

- $f(x) = -\log(1-x)$ for $-1 < x < 1$.
- Choose $\{\phi_1(x), \phi_2(x), \ldots\}$ to be the Chebyshev polynomials.
- Then $\beta_i = 2/i$, so the 'optimal' choice of the prior variances would be $v_i = 4/i^2$, i.e., $c = 4$ and $A = 2$.
- Measure the predictive capability of a model by expected squared error loss relative to the true function (here known) – thus we use a frequentist evaluation, as did Shibata.

	MaxPr	MedPr	ModAv	BIC	AIC
$a = 1$	0.99 [8]	0.89 [10]	0.84	1.14 [8]	1.09 [7]
$a = 2$	0.88 [10]	0.80 [16]	0.81	1.14 [8]	1.09 [7]
$a = 3$	0.88 [9]	0.84 [17]	0.85	1.14 [8]	1.09 [7]

Table 1: For $n = 30$ and $\sigma^2 = 1$, the expected loss and average model size for the maximum probability model (MaxPr), the Median Probability Model (MedPr), Model Averaging (ModAv), and BIC and AIC, in the Shibata example.

	MaxPr	MedPr	ModAv	BIC	AIC
$a = 1$	0.54 [14]	0.51 [19]	0.47	0.59 [11]	0.59 [13]
$a = 2$	0.47 [23]	0.43 [43]	0.44	0.59 [11]	0.59 [13]
$a = 3$	0.47 [22]	0.46 [45]	0.46	0.59 [11]	0.59 [13]

Table 2: For $n = 100$ and $\sigma^2 = 1$, the expected loss and average model size for the maximum probability model (MaxPr), the Median Probability Model (MedPr), Model Averaging (ModAv), and BIC and AIC, in the Shibata example.

	MaxPr	MedPr	ModAv	BIC	AIC
$a = 1$	0.34 [23]	0.33 [26]	0.30	0.41 [12]	0.38 [21]
$a = 2$	0.26 [42]	0.25 [51]	0.25	0.41 [12]	0.38 [21]
$a = 3$	0.29 [38]	0.29 [50]	0.29	0.41 [12]	0.38 [21]

Table 3: For $n = 2000$ and $\sigma^2 = 3$, the expected loss and average model size for the maximum probability model (MaxPr), the Median Probability Model (MedPr), Model Averaging (ModAv), and BIC and AIC, in the Shibata example.

Comments

- AIC is better than BIC (as Shibata showed), but the true Bayesian procedures are best.
- Model averaging is generally best (not obvious), followed closely by the median probability model. The maximum probability model can be considerably inferior.
- BIC is a very poor approximation to the Bayesian answers.
- The true Bayesian answers choose substantially *larger* models than AIC (and then shrink towards 0).

ANOVA

Many ANOVA problems, when written in linear model form, yield diagonal $\mathbf{X}'\mathbf{X}$ and any such problems will naturally fit under our theory. In particular, this is true for any balanced ANOVA in which each factor has only two levels. As an example, consider the full two-way ANOVA model with interactions:

$$y_{ijk} = \mu + a_i + b_j + ab_{ij} + \epsilon_{ijk}$$

with $i = 1, 2$, $j = 1, 2$, $k = 1, 2, \ldots, K$ and ϵ_{ijk} independent $N(0, \sigma^2)$, with σ^2 unknown. In linear model form, this leads to $\mathbf{X}'\mathbf{X} = 4K\mathbf{I}_4$.

Possible modeling scenarios

We use the simplified notation M_{1011} instead of $M_{(1,0,1,1)}$, representing the model having all parameters except a_1.

Scenario 1 - *All models with the constant μ:* Thus the set of models under consideration is $\{M_{1000}, M_{1100}, M_{1010}, M_{1001}, M_{1101}, M_{1011}, M_{1110}, M_{1111}\}$.

Scenario 2 - *Interactions present only with main effects, and μ included:* The set of models under consideration here is $\{M_{1000}, M_{1100}, M_{1010}, M_{1110}, M_{1111}\}$. Note that this set of models has graphical structure.

Scenario 3 - *An analogue of an unusual classical test:* In classical ANOVA testing, it is sometimes argued that one might be interested in testing for no interaction effect followed by testing for the main effects, even if the no-interaction test rejected. The four models that are under consideration in this process, including the constant μ in all, are $\{M_{1101}, M_{1011}, M_{1110}, M_{1111}\}$. This class of models does *not* have graphical model structure and yet the median probability model is guaranteed to be in the class.

Example 2. Montgomery (1991, pp.271–274) considers the effects of the concentration of a reactant and the amount of a catalyst on the yield in a chemical process. The reactant concentration is factor A and has two levels, 15% and 25%. The catalyst is factor B, with the two levels 'one bag' and 'two bags' of catalyst. The experiment was replicated three times and the data are

treatment combination	replicates		
	I	II	III
A low, B low	28	25	27
A high, B low	36	32	32
A low, B high	18	19	23
A high, B high	31	30	29

For each modeling scenario, two Bayesian analyses were carried out, both satisfying the earlier conditions so that the median probability model is known to be the optimal predictive model.

- I. The reference prior $\pi(\mu, \sigma) \propto \frac{1}{\sigma}$ was used for the common parameters, while the standard $N(0, \sigma^2)$ g-prior was used for a_1, b_1 and ab_{11}. In each scenario, the models under consideration were given equal prior probabilities of being true.

- II. The g-prior was also used for the common μ.

model	posterior probability	posterior expected loss
M_{1000}	0.0009	237.21
M_{1100}	0.0347	60.33
M_{1010}	0.0009	177.85
M_{1110}	0.6103	0.97
M_{1111}	0.3532	3.05

Table 4: Scenario 2 – graphical models, prior I. The posterior inclusion probabilities are $p_2 = 0.9982$, $p_3 = 0.9644$, and $p_4 = 0.3532$; thus M_{1110} is the median probability model.

model	posterior probability	posterior expected loss
M_{1011}	0.124	143.03
M_{1101}	0.286	36.78
M_{1110}	0.456	10.03
M_{1111}	0.134	9.41

Table 5: Scenario 3 – unusual classical models, prior II. The posterior inclusion probabilities are $p_2 = 0.876$, $p_3 = 0.714$, and $p_4 = 0.544$; thus M_{1111} is the median probability model.

When does the median probability model fail? (Merlise Clyde)

- Suppose
  - under consideration are the model with only a constant term, and the models with a constant term and a single covariate x_i, $i = 1, \ldots, k$, with $k \geq 3$;
  - the models have equal prior probability of $\frac{1}{k+1}$;
  - all covariates are nearly perfectly correlated, with each other and with y.
- Then
  - the posterior probability of the constant model will be near zero, and that of each of the other models will be approximately $1/k$;
  - thus the posterior inclusion probabilities will also be approximately $1/k < 1/2$;
  - so the median probability model is the constant model, which will have poor predictive performance compared to any other model.

A high correlation example where the theory does not apply

Example: Consider Hald's regression data (Draper and Smith, 1981), consisting of $n = 13$ observations on a dependent variable y, with four potential regressors: x_1, x_2, x_3, x_4. The full model is thus

$$y = c + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

with σ^2 unknown.

- All models that include the constant term are considered. This example does not formally satisfy the theory, since the models are not nested and the conditions of Theorem 3 do not apply.
- Least squares estimates are used for parameters.
- Default posterior model probabilities, $P(M_{\boldsymbol{l}}|\boldsymbol{y})$, are computed using the Encompassing Arithmetic Intrinsic Bayes Factor (Berger and Pericchi, 1996), together with equal prior model probabilities.
- Predictive risks, $R(M_{\boldsymbol{l}})$, are computed.

Model	$P(M_{\boldsymbol{l}} \mid \boldsymbol{y})$	$R(M_{\boldsymbol{l}})$
c	0.000003	2652.44
c,1	0.000012	1207.04
c,2	0.000026	854.85
c,3	0.000002	1864.41
c,4	0.000058	838.20
c,1,2	0.275484	8.19
c,1,3	0.000006	1174.14
c,1,4	0.107798	29.73

Model	$P(M_{\boldsymbol{l}} \mid \boldsymbol{y})$	$R(M_{\boldsymbol{l}})$
c,2,3	0.000229	353.72
c,2,4	0.000018	821.15
c,3,4	0.003785	118.59
c,1,2,3	0.170990	1.21
c,1,2,4	0.190720	0.18
c,1,3,4	0.159959	1.71
c,2,3,4	0.041323	20.42
c,1,2,3,4	0.049587	0.47

- The posterior inclusion probabilities are

$$p_1 = \sum_{\boldsymbol{l}:l_1=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.95, \;\; p_2 = \sum_{\boldsymbol{l}:l_2=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.73$$

$$p_3 = \sum_{\boldsymbol{l}:l_3=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.43, \;\; p_4 = \sum_{\boldsymbol{l}:l_4=1} P(M_{\boldsymbol{l}}|\boldsymbol{y}) = 0.55.$$

- Thus the median probability model is $\{c, x_1, x_2, x_4\}$ which clearly coincides with the optimal predictive model.
- Note that the risk of the maximum probability model $\{c, x_1, x_2\}$ is considerably higher than that of the median probability model.

Conclusions

- The (posterior) median probability model will typically be the optimal predictive model.
- The median probability model is typically easy to compute, requiring only rough estimates of the posterior inclusion probabilities.
- The posterior inclusion probabilities are themselves quantities of basic interest in model selection and searches of model space.

III. More on p-values and their Calibration

Background concerns with p-values

- Concerns with use of p-values trace back to at least Berkson (1937).
- Concerns are recurring in many scientific literatures:

  Environmental sciences: http://www.indiana.edu/∼stigtsts/

  Social sciences: http://acs.tamu.edu/∼bbt6147/

  Wildlife science: http://www.npwrc.usgs.gov/perm/hypotest/

  http://www.cnr.colostate.edu/∼anderson/null.html
- Numerous works specifically focus on comparing the Fisher and N-P approaches (e.g., Lehmann, 1993 JASA: The Fisher, N-P Theories of Testing Hypotheses: One Theory or Two?)
- Even the popular press has become involved: The Sunday Telegraph, September 13, 1998:


Review
Telegraph
My China girl
Josephine, six months after adoption, p4
Vicki Woods
Football, big business and little girls, p3
The great health hoax
Swedish studies pinpoint power line cancer link

Commonly aired issues with p-values

- Inappropriate fixation with $p = 0.05$.
- p-values do not measure the magnitude or importance of the effect being investigated.
- p-values are commonly misinterpreted as
  - the probability that H_0 is true, given the data;
  - the probability an error is made in rejecting H_0;
  - the probability that a 'replicating' experiment would reach the same conclusion.

Cohen (1994): The statistical significance test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!"

- Many (most?) null hypotheses are obviously false and will surely be rejected if one simply collects enough data.

  Thompson (1992): "Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical significance test to evaluate whether or not there were a lot of subjects, which the researchers already know because ... they're tired."

- Misuse of p-values is not likely to naturally go away.

  Campbell (1982): "p-values are like mosquitoes [that apparently] have an evolutionary niche somewhere and [unfortunately] no amount of scratching, swatting or spraying will dislodge them."

Possible solutions?

- Teach p-values for diagnostic purposes, but *not* for inference or decision.
  Hogben (1957): “We can already detect signs of such deterioration in the growing volume of published papers . . . recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration.”
  Rozeboom (1997): “Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students.”

- Only present p-values for precise hypotheses with an appropriate calibration; this will eliminate the worst excesses.

I. J. Good's efforts to calibrate p-values

The difference between p-values and Bayes factors fascinated Good, and he would return to study it throughout his career. As part of his Bayes/non-Bayes compromise, he wanted a simple formula relating the two. Here is a series of suggestions from Good.

- $B_{01} = 3$ or 4 times p: (0.08 or 0.12 in the vaccine example.)
- $10p/3 < B_{01} < 30p$: ($0.133 < B_{01} < 1.2$ in the vaccine example.)
- $B_{01} \approx p\sqrt{n}$: ($B_{01} = 2.5$ in the vaccine example.)
- $p\sqrt{2\pi n}/6 < B_{01} < 6p\sqrt{2\pi n}$: ($1.07 < B_{01} < 38.5$ in the vaccine example.)

Note that his later efforts at calibration all involved the sample size n.

Calibration of p-values from Vovk (1993 JRSSB) and Sellke, Bayarri and Berger (2001 Am. Stat.)

- 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

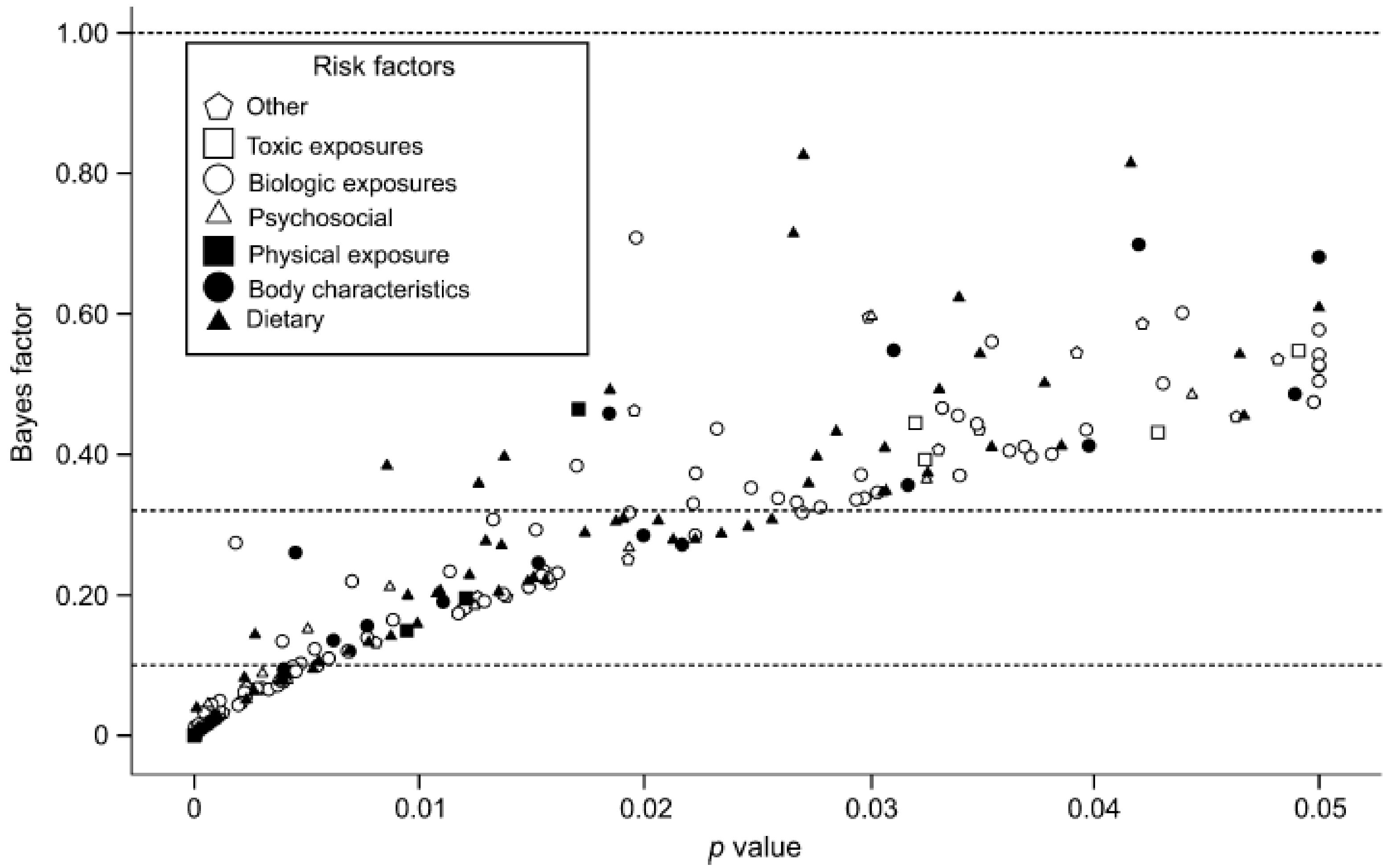


FIGURE 1. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each study. Shown are calculations assuming θ_A of 0.50 (relative risk = 1.65). The dashed lines correspond to threshold values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

Figure 1: J.P. Ioannides: Am J Epidemiol 2008;168:374–383

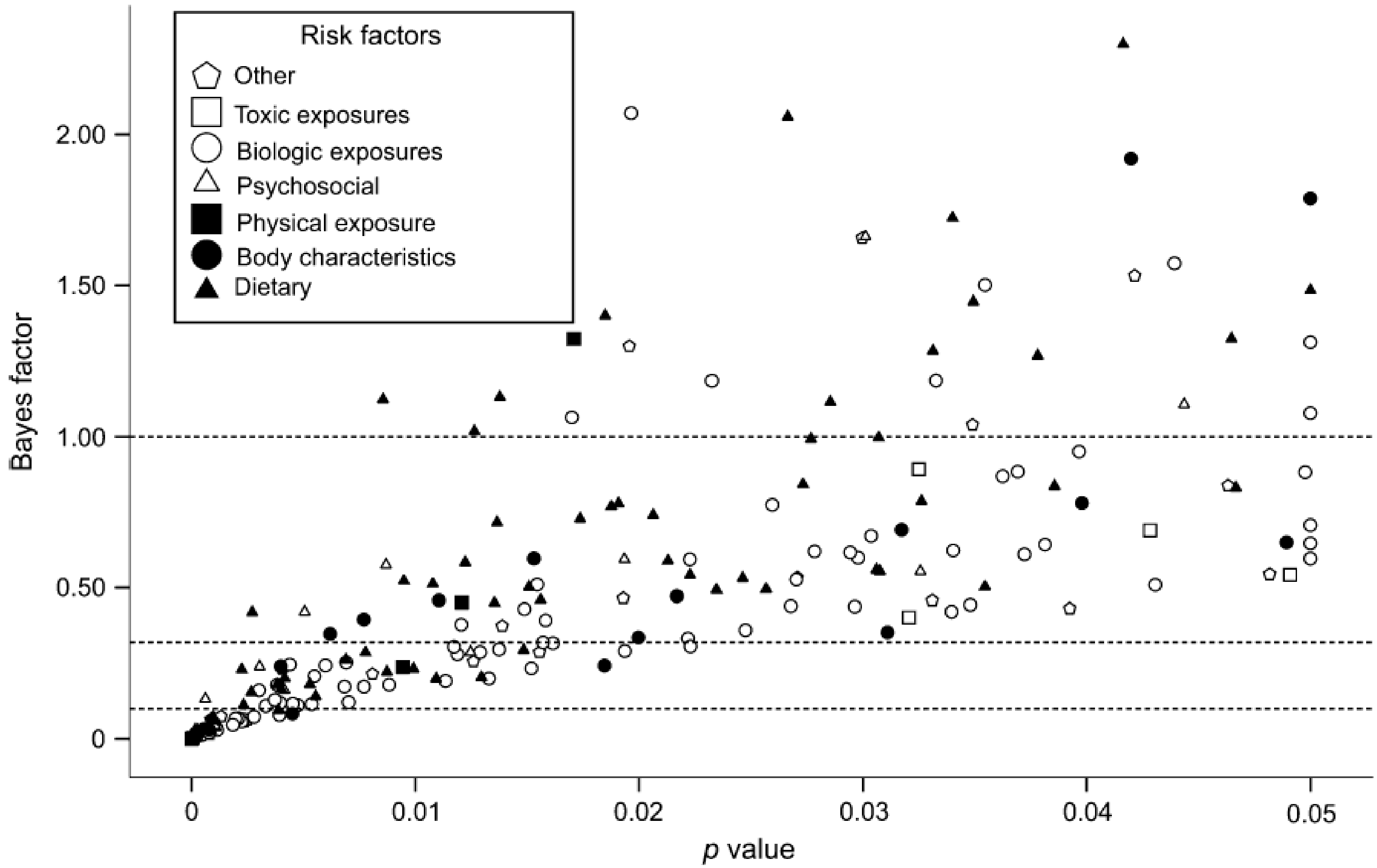


FIGURE 2. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed p value in each study. Shown are calculations assuming θ_A of 1.50 (relative risk = 4.48). The dashed lines correspond to threshold values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

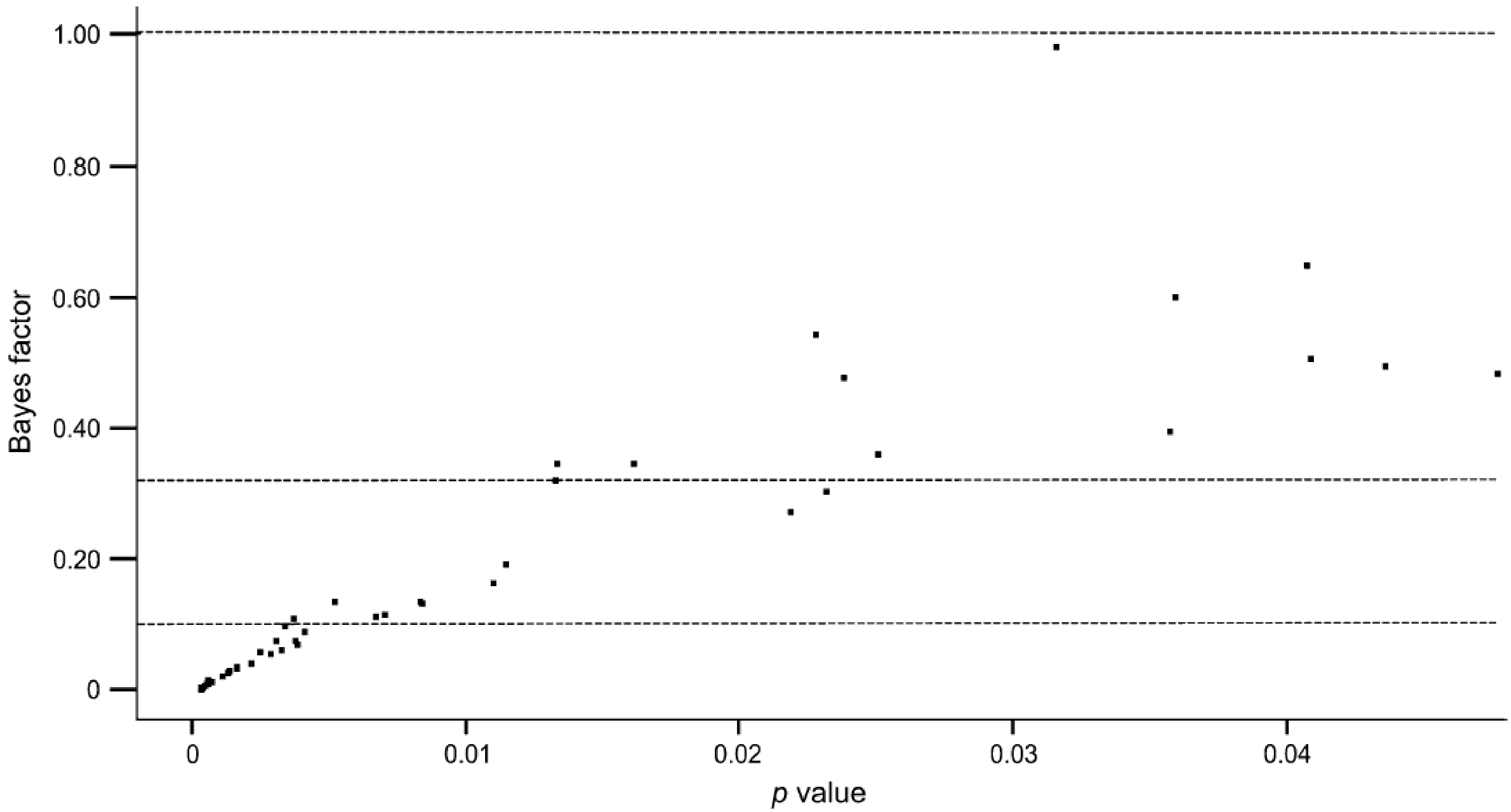


FIGURE 3. Estimated Bayes factors for 50 meta-analyses of genetic associations with formally statistically significant results. The Bayes factor is plotted against the observed p value in each meta-analysis. Calculations assume θ_A equal to the median relative risk observed in the 50 genetic associations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

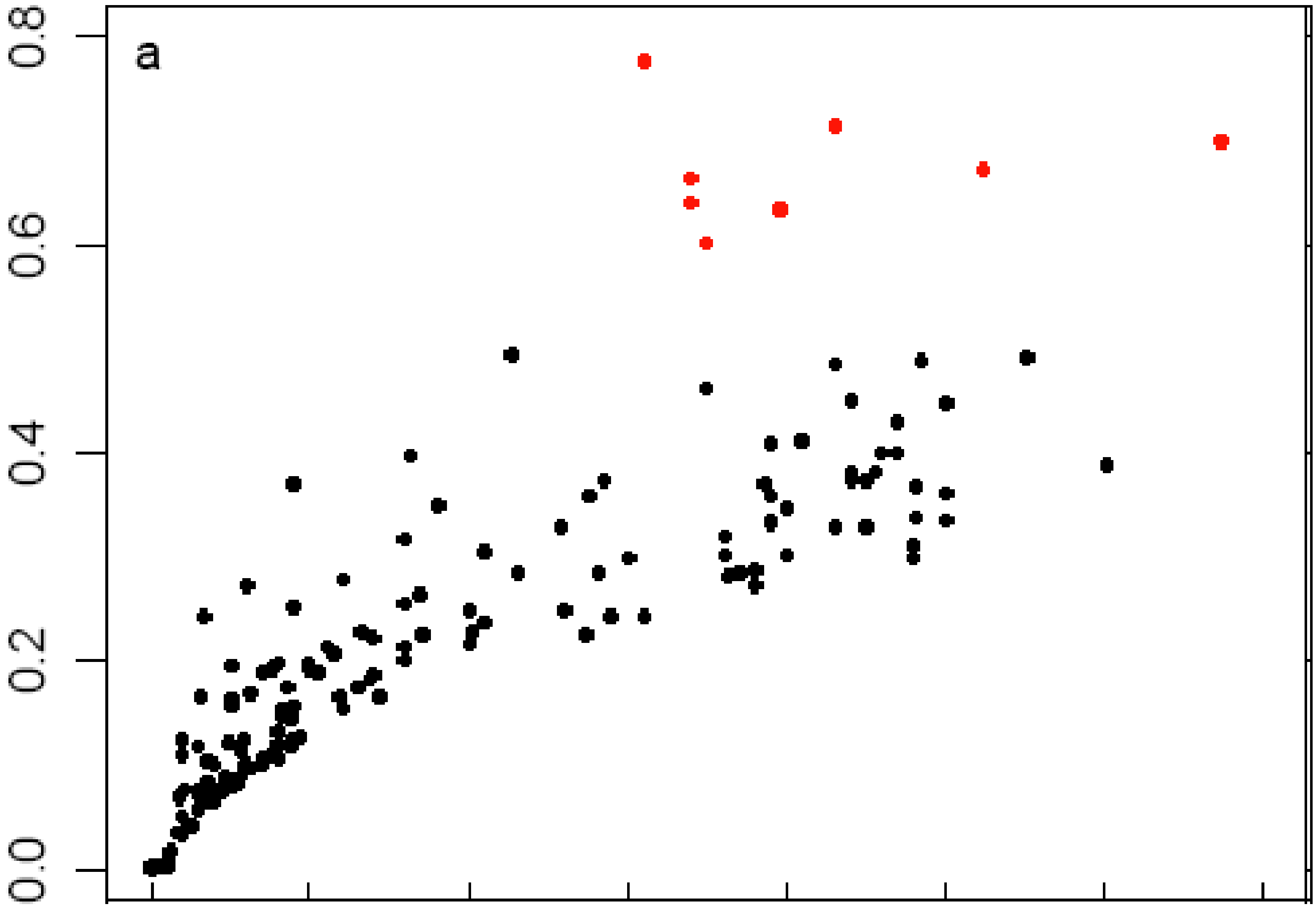


Figure 2: Elgersma and Green (2011): $\alpha(p)$ versus observed p-values for 314 articles in Ecology in 2009.

Three legitimate uses of p-values

- As an indication that something unusual has happened and one should investigate further.
  - Very useful at initial stages of the analysis:
    - If p is not small, don't proceed.
    - If it is small, perform the Bayesian analysis.
  - It would be better to measure 'unusual' by $-ep \log p$.
- For model criticism (though again calibrate by $-ep \log p$).
- As a statistic to measure 'strength of evidence', for use in conditional frequentist testing; indeed, it yields conditional frequentist tests that are exactly the same as objective Bayesian tests. (This was the story about how use of Fisher's p-value and conditioning ideas, when combined with Neyman's frequentist testing formulation, results in identical answers as Jeffreys objective Bayesian testing.)

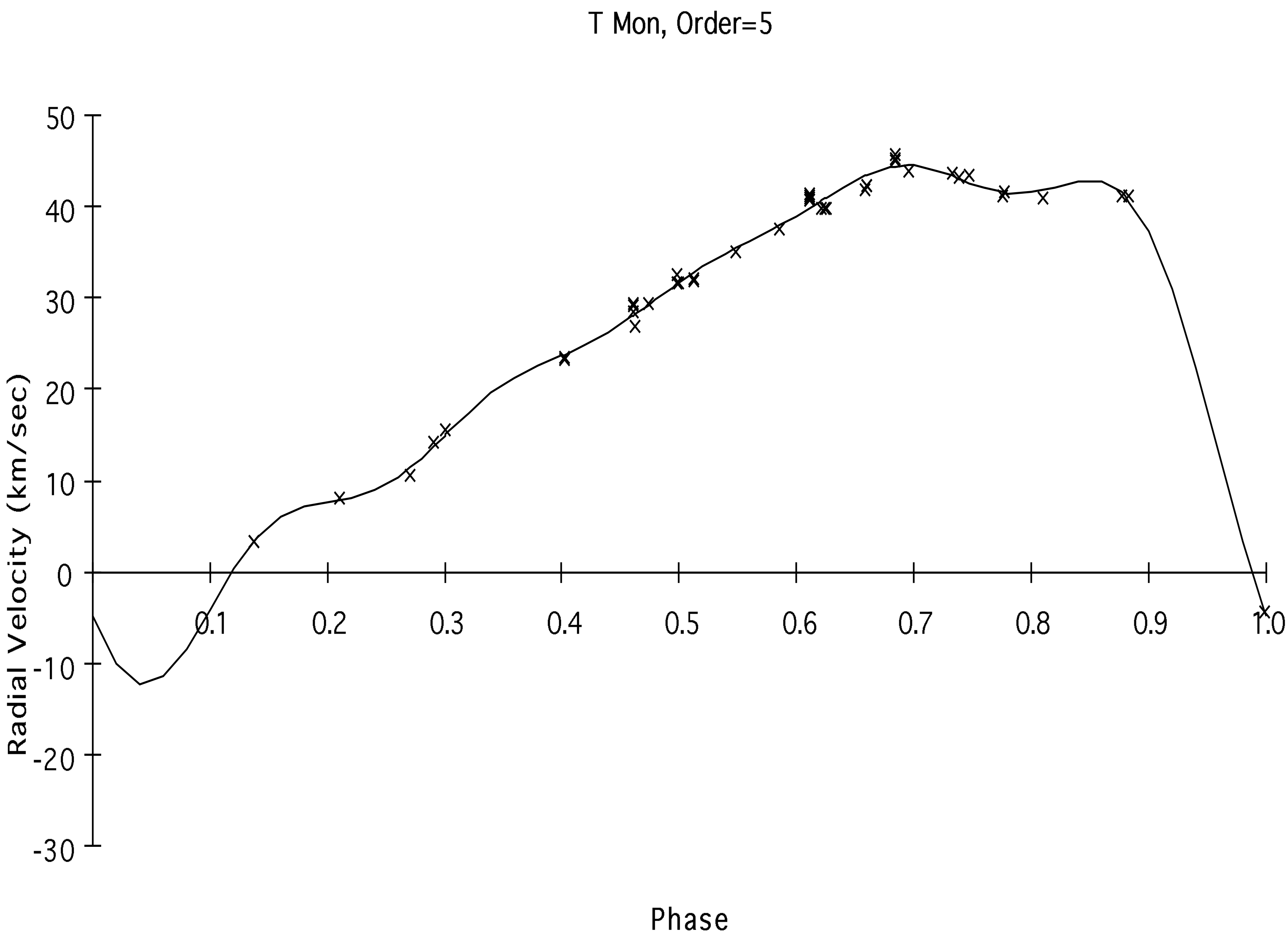
IV. Application: Cepheid Star Oscillations

(Berger, Jefferys, Müller, and Barnes, 2001)

- The astronomical problem
- Bayesian model selection
- The data and likelihood function
- Choice of prior distributions
- Computation and results

The Astronomical Problem

- A Cepheid star pulsates, regularly varying its luminosity (light output) and size.
- From the Doppler shift as the star surface pulsates, one can compute surface velocities at certain phases of the star's period, thus learning how the radius of the star changes.
- From the luminosity and 'color' of the star, one can learn about the angular radius of the star (the angle from Earth to opposite edges of the star).
- Combining, allows estimation of s, a star's distance.


T Mon, Order=5
50
40
30
20
10
0
-10
-20
-30
Radial Velocity (km/sec)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Phase

Curve Fitting

To determine the overall change in radius of the star over the star's period, the surface velocity must be estimated at phases other than those actually observed, leading to a curve fitting problem (also for luminosity).

Difficulties:

- Observations have measurement error.
- Phases at which observations are made are unequally spaced.
- 100 possible models (curve fits) are entertained.
- Resulting models have from 10 to 50 parameters.

Other Statistical Difficulties

- All uncertainties (measurement errors, model uncertainty and estimated curve inaccuracy) need to be accounted for.
- There are significant nonlinear features of the model.
- Prior information, such as understanding of the Lutz-Kelker bias, needs to be incorporated into the analysis.

Bayesian Analysis with Uncertain Models

- The data, $\boldsymbol{Y}$, arises from one of the models $M_1, \ldots, M_q$.
- Under M_i, the density of $\boldsymbol{Y}$ is $f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)$.
- $\boldsymbol{\theta}_i$ is an unknown vector of parameters under M_i.
- Specify prior probabilities of models (usually $1/q$).
- Prior distributions $\pi_i(\boldsymbol{\theta}_i)$ for the $\boldsymbol{\theta}_i$ are specified.
- Bayes theorem then gives the model posterior probabilities (and posterior distributions of the other unknowns).

Model Averaging

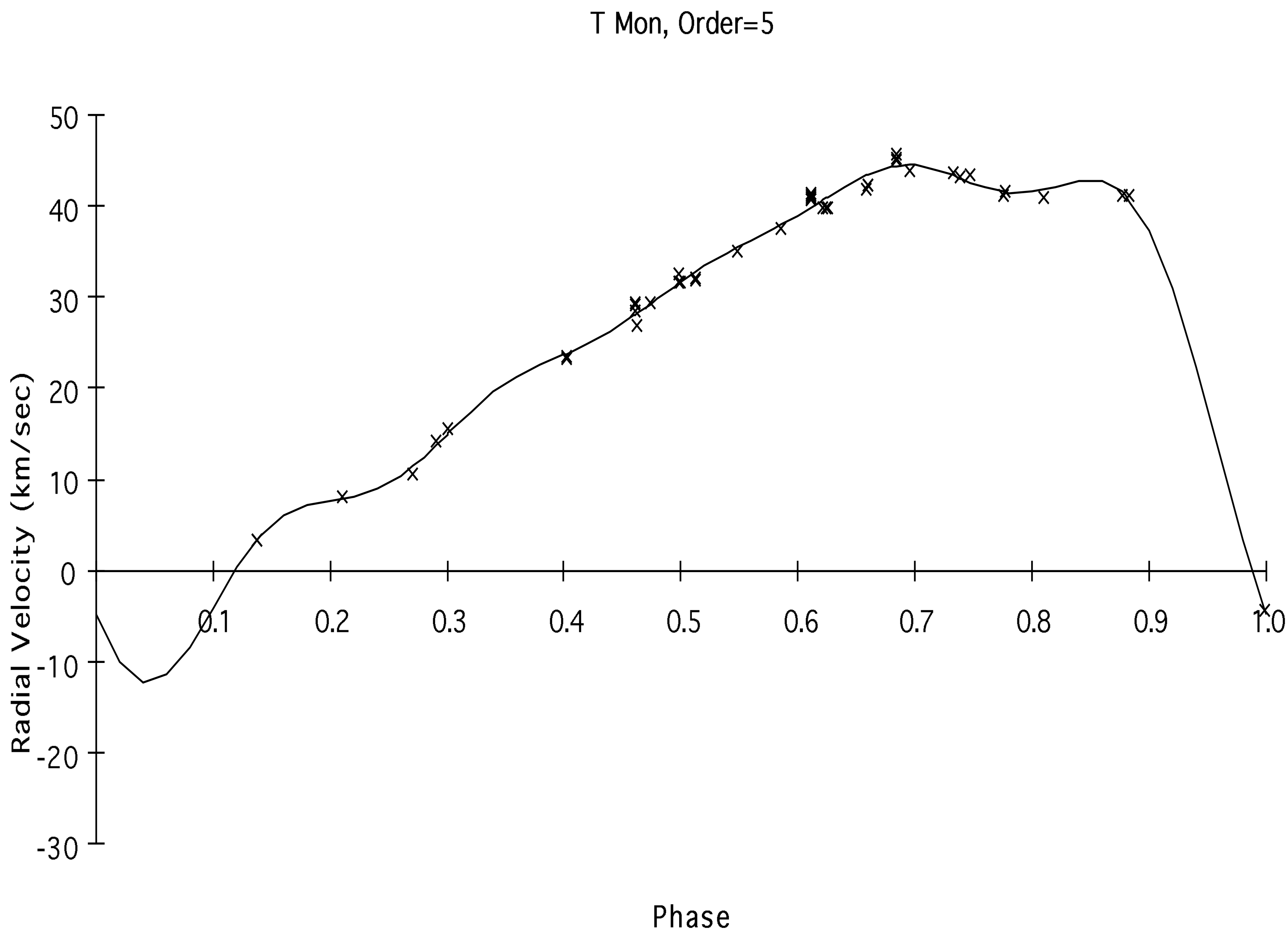
- Suppose Models 5, 6, and 7 have posterior probabilities 0.34, 0.56 and 0.10, respectively.
- Model uncertainty is then handled by *Bayesian model averaging:* if Models 5, 6, and 7 provided distance estimates of 750, 790, and 800 parsecs, the 'model-averaged' distance estimate would be $(0.34)\,750 + (0.56)\,790 + (0.10)\,800 = 777.4$ parsecs.
- The 'model-averaged' variance can also be computed; it can be several times that of a single model.

The Data and Statistical Model

- Data:
  - m observed radial velocities U_i, $i = 1, \ldots, m$.
  - n vectors of photometry data consisting of luminosity V_i, $i = 1, \ldots, n$, and color index C_i, $i = 1, \ldots, n$.
- Specified standard deviations σ_{U_i}, σ_{V_i}, and σ_{C_i}; unknown adjustment factors are inserted, leading to variances $\sigma^2_{U_i}/\tau_u$, $\sigma^2_{V_i}/\tau_v$, $\sigma^2_{C_i}/\tau_c$.
- The statistical model for measurement error:

$$
\begin{aligned}
U_i &\sim N(u_i, \sigma^2_{U_i}/\tau_u), \\
V_i &\sim N(v_i, \sigma^2_{V_i}/\tau_v), \\
C_i &\sim N(c_i, \sigma^2_{C_i}/\tau_c),
\end{aligned}
$$

  where u_i, v_i, and c_i denote the true unknown mean velocity, luminosity, and color index.
- Let $\boldsymbol{G}_u$ and $\boldsymbol{G}_v$ be the known diagonal matrices of $\sigma^2_{U_i}$ and $\sigma^2_{V_i}$.


T Mon, Order=5
Radial Velocity (km/sec)
50
40
30
20
10
0
-10
-20
-30
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Phase

Curve Fitting I: Fourier Analysis

- Model the periodic velocities, u, as trigonometric polynomials. For the velocity u at phase ϕ,

$$u = u_0 + \sum_{j=1}^{M} [\theta_{1j} \cos(j\phi) + \theta_{2j} \sin(j\phi)],$$

  where u_0 is the mean velocity and M is the (unknown) order of the trigonometric polynomial.

- There is a similar polynomial model for luminosity, v, having unknown order N.

Statistically, these trigonometric polynomial models can be written as the linear models

$$\boldsymbol{U} = u_0\mathbf{1} + \boldsymbol{X}_u\boldsymbol{\theta}_u + \boldsymbol{\varepsilon}_u \text{ and } \boldsymbol{V} = v_0\mathbf{1} + \boldsymbol{X}_v\boldsymbol{\theta}_v + \boldsymbol{\varepsilon}_v,$$

- u_0 and v_0 are the (unknown) mean velocity and luminosity and $\mathbf{1}$ is the column vector of ones.
- $\boldsymbol{X}_u$ and $\boldsymbol{X}_v$ are matrices of the trigonometric covariates (e.g., terms like $\sin(j\phi)$);
- $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$ are the unknown Fourier coefficients;
- $\boldsymbol{\varepsilon}_u$ and $\boldsymbol{\varepsilon}_v$ are independently multivariate normal errors: $\boldsymbol{N}(\mathbf{0}, \boldsymbol{G}_u/\tau_u)$ and $\boldsymbol{N}(\mathbf{0}, \boldsymbol{G}_v/\tau_v)$.

Color need not be modeled separately because it is related to luminosity (v) and velocity (or change in radius) by

$$c_i = a[0.1v_i - b + 0.5\log(\phi_0 + \Delta r_i/s)],$$

where a and b are known constants, ϕ_0 and s are the angular size and distance of the star, and Δr, the change in radius corresponding to phase ϕ, is given by

$$\Delta r = -g\sum_{j=1}^{M}\frac{1}{j}[\theta_{1j}\sin(j(\phi-\Delta\phi)) - \theta_{2j}\cos(j(\phi-\Delta\phi))],$$

with 'phase shift' $\Delta\phi$ and g a known constant.

Objective Priors with Uncertain Models

- For estimation within one model, objective prior distributions are readily available (Jeffreys-rule priors, maximum entropy priors, reference priors).
  Example: If θ is an unknown normal mean, use $\pi(\theta) = 1$.
- These are typically improper (integrate to infinity); this is not a problem for estimation, but for testing and model selection is usually inappropriate.
- Various guidelines for choice of priors in model selection have been given: here are three we need.

Common Model Parameters

- When all models have 'common' parameters, they can be assigned the usual improper objective prior.
  Example: All Cepheid radial velocity models have common mean level u_0; it is okay to assign the usual objective (improper) prior $\pi(u_0) = 1$.
- This is generally true if models have the same 'location parameters' (as above) or 'scale parameters' (as in a normal variance).
- For discussion of other types of 'common parameters', see Berger and Pericchi (2001).

Priors for General Linear Models:

$$\boldsymbol{Y} = \theta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ are observations (e.g., radial velocities in the Cepheid problem);
- $\boldsymbol{X}$ is the matrix of covariates (e.g., terms like $\sin(j\phi)$ from trigonometric polynomials);
- $\boldsymbol{\theta}$ is an unknown vector (e.g., Fourier coefficients);
- $\mathbf{1}$ is a vector of ones and θ_0 the unknown mean level;
- $\boldsymbol{\varepsilon}$ is $\boldsymbol{N}(\mathbf{0}, \sigma^2 \boldsymbol{G})$, where $\boldsymbol{G}$ (e.g., the diagonal matrix of measurement variances) is known.

- The recommended prior (from Zellner and Siow, 1980)
  - for θ_0 is $\pi(\theta_0) = 1$;
  - for $\boldsymbol{\theta}$, given σ^2, is Cauchy$(\mathbf{0}, n\sigma^2(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1})$; for computational convenience, this can be written in two stages as

$$\pi(\boldsymbol{\theta} \mid \sigma^2, \tau) \text{ is } \boldsymbol{N}(\mathbf{0}, \tau n\sigma^2(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1});$$
$$\pi(\tau) = \frac{1}{\sqrt{2\pi}\,\tau^{3/2}} \exp(-\frac{1}{2\tau}).$$

Choice of Cepheid Prior Distributions

- The orders of the trigonometric polynomials, (M, N), are given a uniform distribution up to some cut-off (e.g., $(10, 10)$).
- τ_u, τ_v, τ_c, which adjust the measurement standard errors, are given the standard objective priors for 'scale parameters,' namely the Jeffreys-rule priors $\pi(\tau_u) = \frac{1}{\tau_u}$, $\pi(\tau_v) = \frac{1}{\tau_v}$, and $\pi(\tau_c) = \frac{1}{\tau_c}$.
- The mean velocity and luminosity, u_0 and v_0, are 'location parameters' and so can be assigned the standard objective priors $\pi(u_0) = 1$ and $\pi(v_0) = 1$.

- The angular diameter ϕ_0 and the unknown phase shift $\Delta\phi$ are also assigned the objective priors $\pi(\Delta\phi) = 1$ and $\pi(\phi_0) = 1$. It is unclear if these are 'optimal' objective priors but the choice was found to have neglible impact on the answers.
- The Fourier coefficients, $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$, occur in linear models, so the Zellner-Siow priors can be utilized.

- The prior for distance s of the star should account for
  - *Lutz-Kelker bias*: a uniform spatial distribution of Cepheid stars would yield a prior proportional to s^2.
  - The distribution of Cepheids is flattened wrt the galactic plane; we use an exponential distribution.

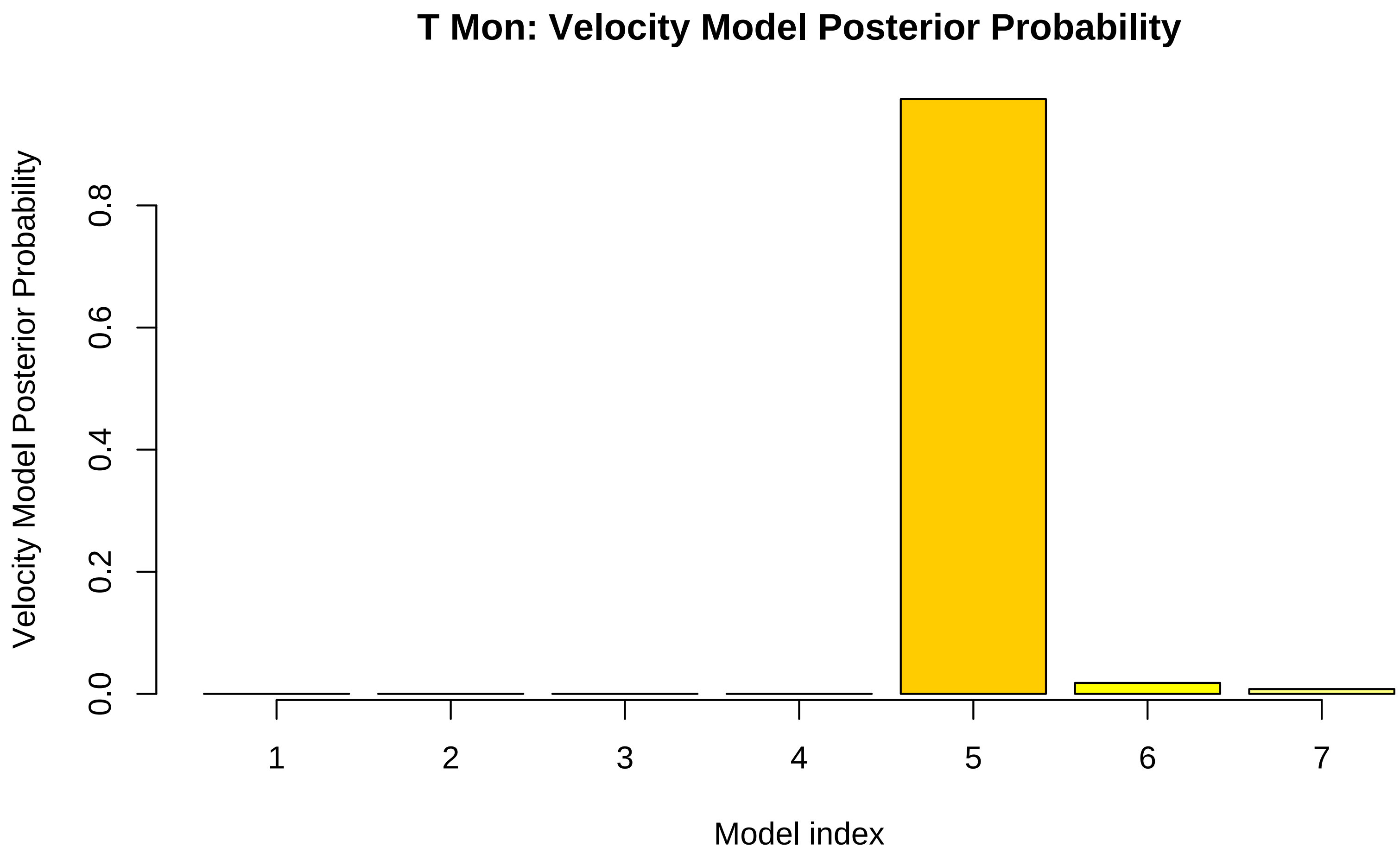
  So, we use $\quad \pi(s) \propto s^2 \exp\left(-|s \sin\beta|/z_0\right)$,
  - β being the known galactic latitude of the star (its angle above the galactic plane),
  - z_0 being the 'scale height,' assigned a uniform prior over the range $z_0 = 97 \pm 7$ parsecs.

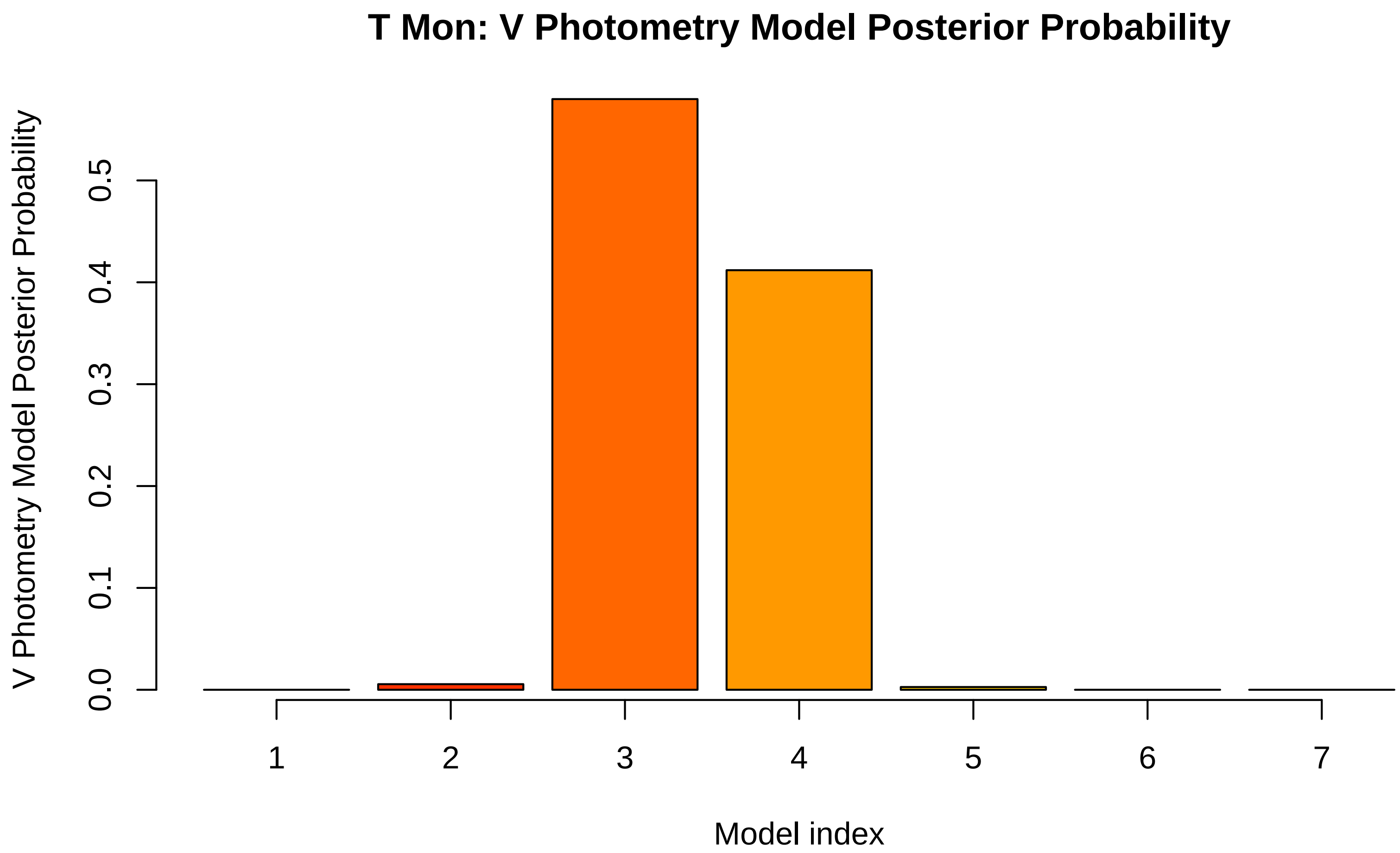
Computation

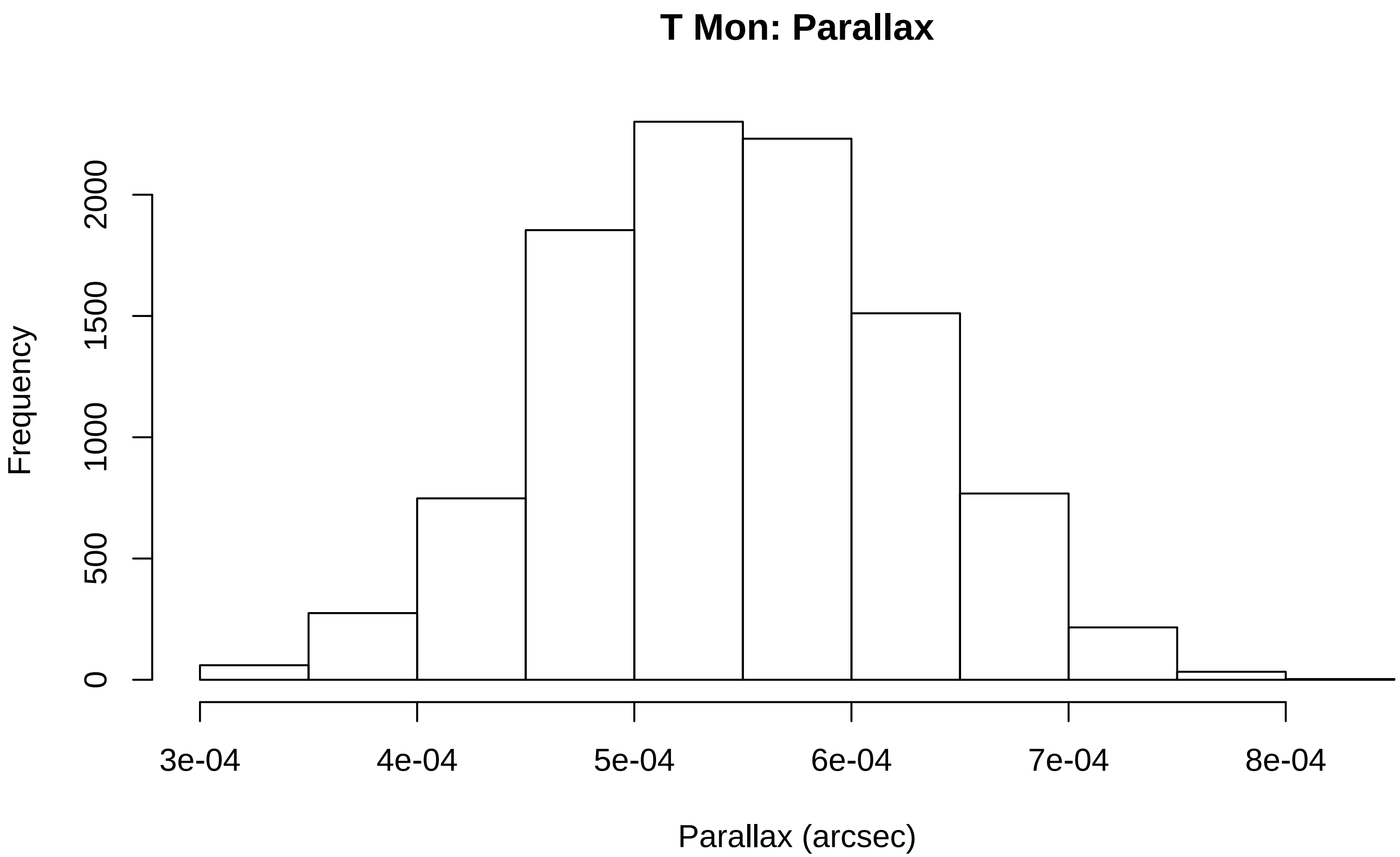
A reversible-jump MCMC algorithm of the type reviewed in Dellaportas, Forster and Ntzoufras (2000) is used to move between models and generate posterior distributions and estimates.

- The full conditional distributions for the variance and precision parameters and hyperparameters are standard gamma and inverse-gamma distributions and are sampled with Gibbs sampling.
- For $\Delta\phi$, ϕ_0 and s, we employ a random-walk Metropolis algorithm using, as the proposal distribution, a multivariate normal distribution centered on the current values and with a covariance matrix found from linearizing the problem for these three parameters.
- The Fourier coefficients $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_v$, as well as u_0 and v_0, are also sampled via Metropolis. The natural proposal distributions are found by combining the normal likelihoods with the normal part of the Zellner-Siow priors, leading to conjugate normal posterior distributions.

- Proposal for moves between models:
  - A 'burn-in' portion of the MCMC with uniform model proposals yielded initial posterior model probabilities, which were then used as the proposal for subsequent model moves.
  - Simultaneously, new values were proposed for the Fourier coefficients.
- The proposal distributions listed above lead to a well-mixed Markov chain, so that only 10,000 iterations of the MCMC computation needed to be performed.


T Mon: Velocity Model Posterior Probability
Velocity Model Posterior Probability
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
Model index


T Mon: V Photometry Model Posterior Probability
V Photometry Model Posterior Probability
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
Model index


T Mon: Parallax
Frequency
0
500
1000
1500
2000
3e-04
4e-04
5e-04
6e-04
7e-04
8e-04
Parallax (arcsec)

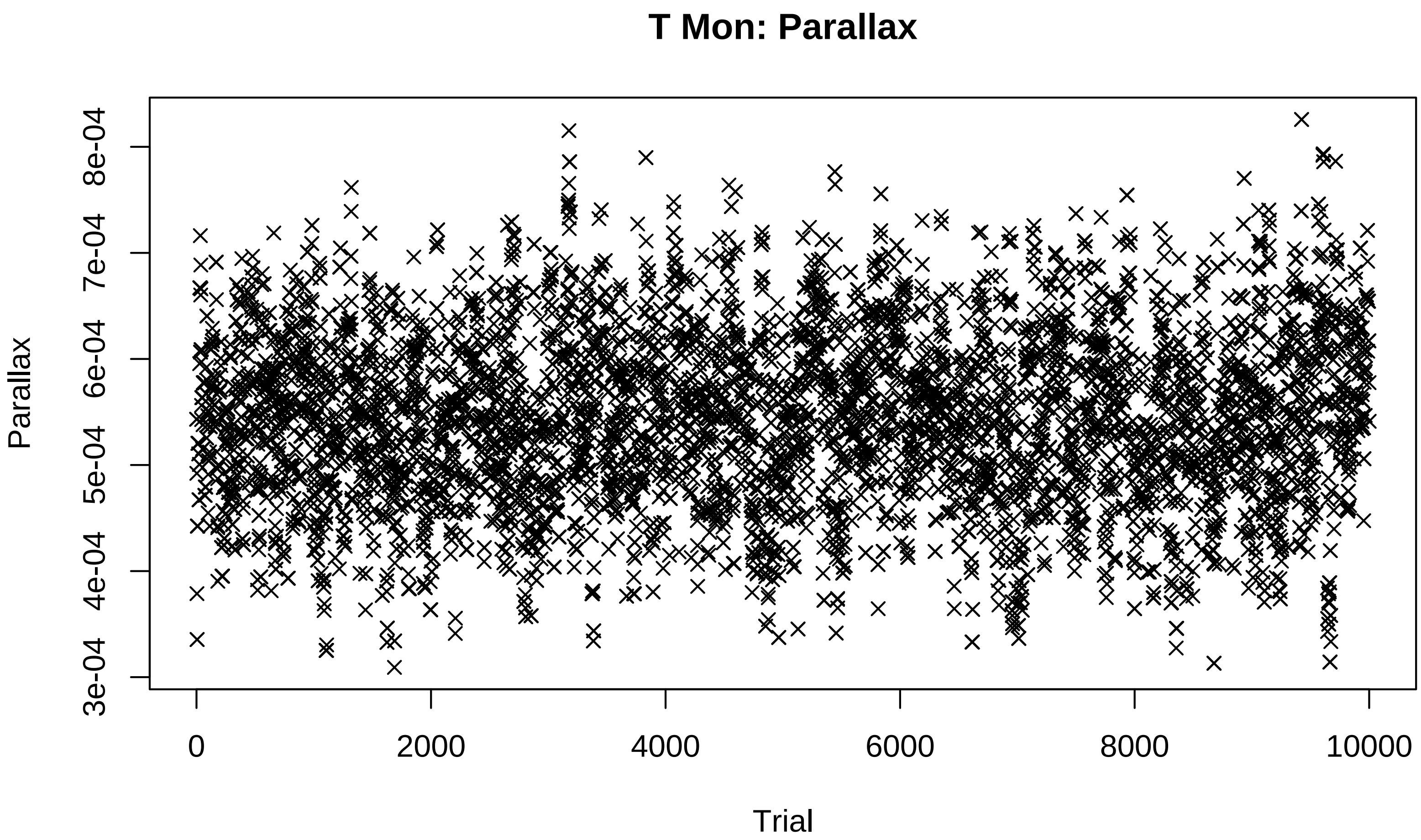

T Mon: Parallax
Parallax
3e-04
4e-04
5e-04
6e-04
7e-04
8e-04
0
2000
4000
6000
8000
10000
Trial

Table 6: For T-Moncerotis, the estimated posterior means [95% posterior intervals] under the EP Prior, g-Prior and ZSN-Prior.

	EP Prior	g-Prior	ZSN-Prior
s	1307 [1226,1398]	1306 [1226,1378]	1294 [1221,1366]
Δ	-0.120 [-0.175,-0.068]	-0.113 [-0.165,-0.056]	-0.116 [-0.178,-0.067]
ϕ_0	1.053 [1.049,1.057]	1.053 [1.048,1.057]	1.053 [1.048,1.058]
τ_C	0.604 [0.479,0.742]	0.529 [0.418,0.645]	0.534 [0.432,0.661]
u_0	22.92 [22.65,23.18]	22.86 [22.59,23.11]	22.91 [22.57,23.19]
v_0	6.18 [6.18,6.19]	6.18 [6.18,6.19]	6.18 [6.18,6.19]
$P(M = 7)$	0.031	0.064	0.080
$P(M = 8)$	0.969	0.936	0.920
$P(N = 5)$	0.333	0	0
$P(N = 6)$	0.667	0.984	0.876
$P(N = 7)$	0	0.016	0.124

V. HIV Vaccine 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 (evidence-based?) 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 – evidence-based – 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:

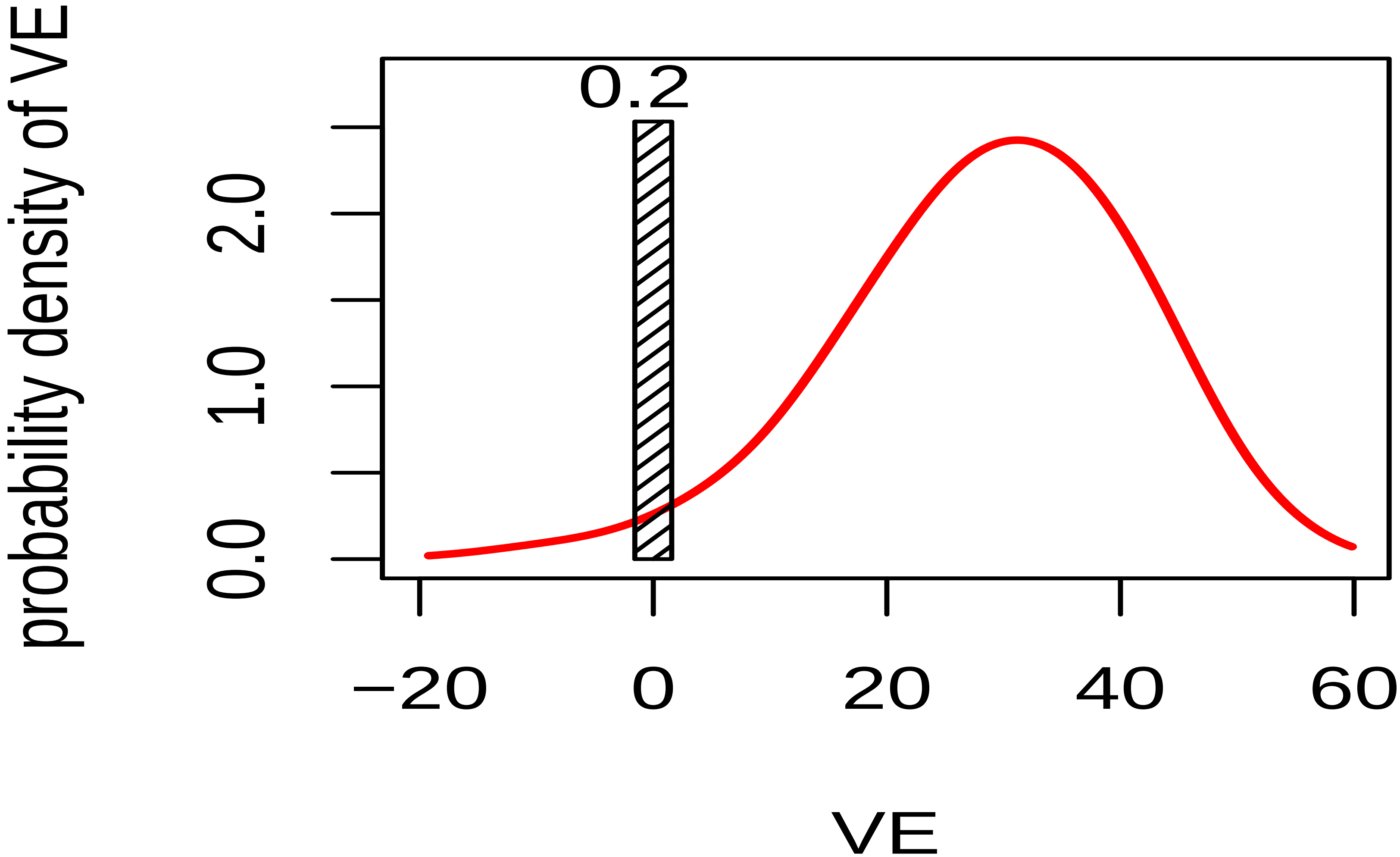
$$Pr(H_0 \mid \text{data}) = \frac{Pr(H_0)B_{01}}{Pr(H_0)B_{01} + Pr(H_1)}\,,$$

where the Bayes factor of H_0 to H_1 is

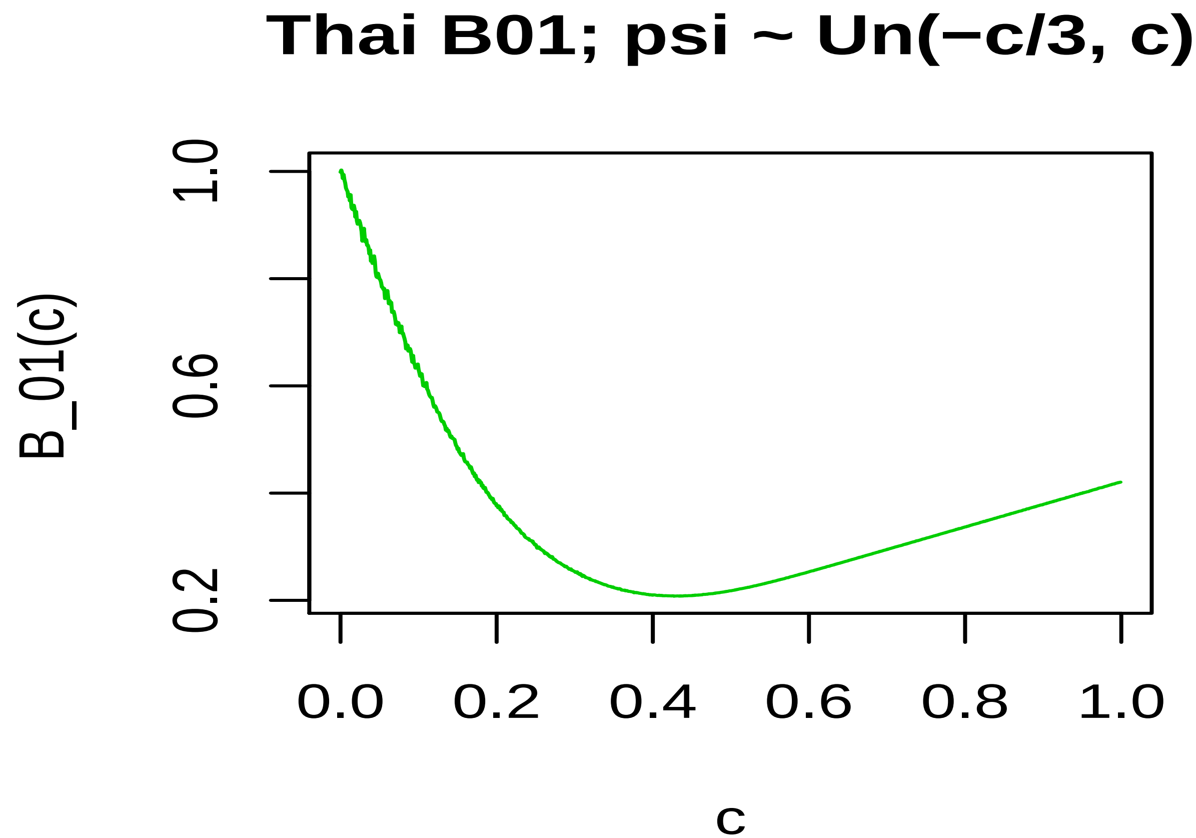
$$B_{01} = \frac{\int \text{Binomial}(74 \mid p_1, 8198)\,\text{Binomial}(51 \mid p_1, 8197)\pi_0(p_1)dp_1}{\int \text{Binomial}(74 \mid p_1, 8198)\,\text{Binomial}(51 \mid p_2, 8197)\pi_0(p_1)\pi_1(p_2 \mid p_1)dp_1dp_2}\,.$$

- 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



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.

Alternative frequentist perspective:

Let α and $(1-\beta(\theta))$ be the Type I error and power for testing H_0 versus H_1 with, say, rejection region $\mathcal{R} = \{z : z > 1.645\}$. Then

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

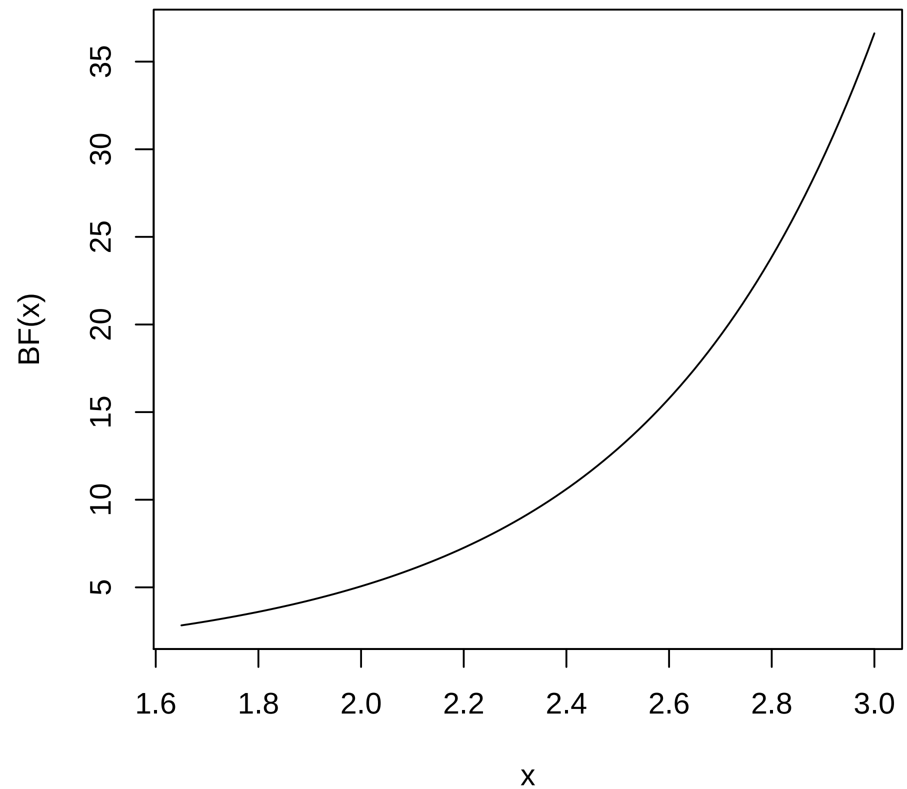
where $(1-\bar{\beta}) = \int (1-\beta(\theta))\pi(\theta)d\theta$ is average power wrt the prior $\pi(\theta)$.

- $\frac{(1-\bar{\beta})}{\alpha} = \frac{\text{average power}}{\text{type 1 error}}$ is the *experimental odds* of correct rejection to incorrect rejection.
- For vaccine example, $(1-\bar{\beta}) = 0.45$ and $\alpha = 0.05$ (the error probability corresponding to $\mathcal{R}$), so $\frac{(1-\bar{\beta})}{\alpha} = 9$.

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 vaccine example, here is $B_{10}(z)$ (recall $\frac{(1-\bar{\beta})}{\alpha} = 9$):



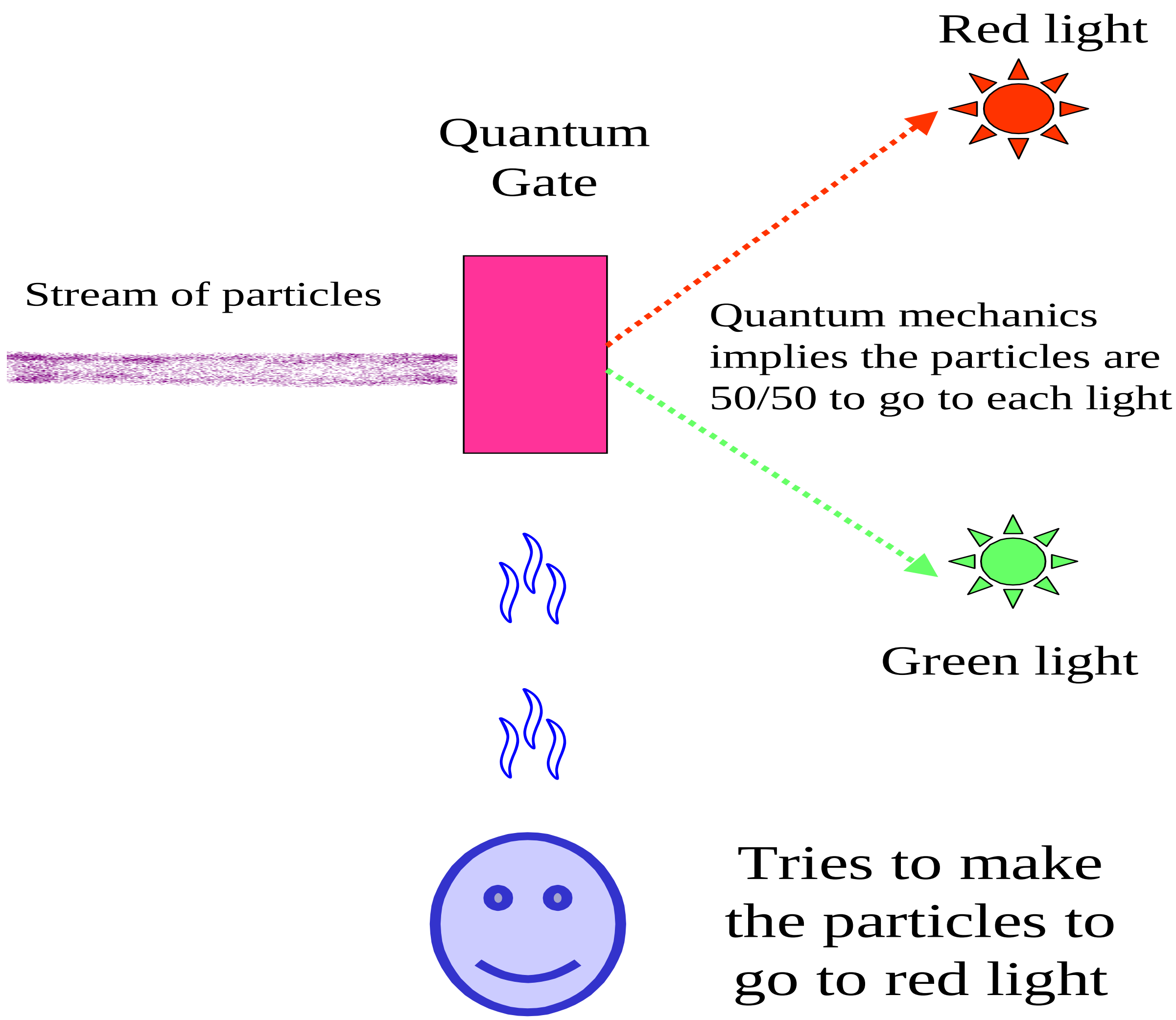
- For simple nulls (or nulls that are simple for the test statistic) $E[B_{10}(Z) \mid H_0, \mathcal{R}] = \frac{(1-\bar{\beta})}{\alpha}$, so reporting $B_{10}(z)$ is a valid conditional frequentist procedure. (Kiefer, 1977 JASA; Brown, 1978 AOS)

VI. Psychokinesis Example

Do people have the ability to perform *psychokinesis*, affecting objects with thoughts?

The experiment:

Schmidt, Jahn and Radin (1987) used electronic and quantum-mechanical random event generators with visual feedback; the subject with alleged psychokinetic ability tries to "influence" the generator.


Red light
Quantum
Gate
Stream of particles
Quantum mechanics
implies the particles are
50/50 to go to each light
Green light
Tries to make
the particles to
go to red light

Data and model:

- Each "particle" is a Bernoulli trial (red = 1, green = 0)
  θ = probability of "1"
  $n = 104,490,000$ trials
  X = # "successes" (# of 1's), $\quad X \sim \text{Binomial}(n, \theta)$
  $x = 52,263,470$ is the actual observation

$$\begin{aligned} \text{To test } H_0 : \theta &= \tfrac{1}{2} \text{ (subject has no influence)} \\ \text{versus } H_1 : \theta &\neq \tfrac{1}{2} \text{ (subject has influence)} \end{aligned}$$

- P-value $= P_{\theta=\frac{1}{2}}(|X - \frac{n}{2}| \geq |x - \frac{n}{2}|) \approx .0003.$
  Is there strong evidence against H_0 (i.e., strong evidence that the subject influences the particles) ?

Bayesian Analysis: (Jefferys, 1990)

Prior distribution:

$Pr(H_i)$ = prior probability that H_i is true, $i = 0, 1$;

On $H_1 : \theta \neq \frac{1}{2}$, let $\pi(\theta)$ be the prior density for θ.

Subjective Bayes: choose the $Pr(H_i)$ and $\pi(\theta)$ based on personal beliefs

Objective (or default) Bayes: choose

$Pr(H_0) = Pr(H_1) = \frac{1}{2}$

$\pi(\theta) = 1 \quad (\text{on } 0 < \theta < 1)$

Posterior probability of hypotheses:

$$Pr(H_0|x) = \frac{f(x\,|\,\theta = \frac{1}{2})\, Pr(H_0)}{Pr(H_0)\, f(x\,|\,\theta = \frac{1}{2}) + Pr(H_1)\int f(x\,|\,\theta)\pi(\theta)d\theta}$$

For the objective prior,

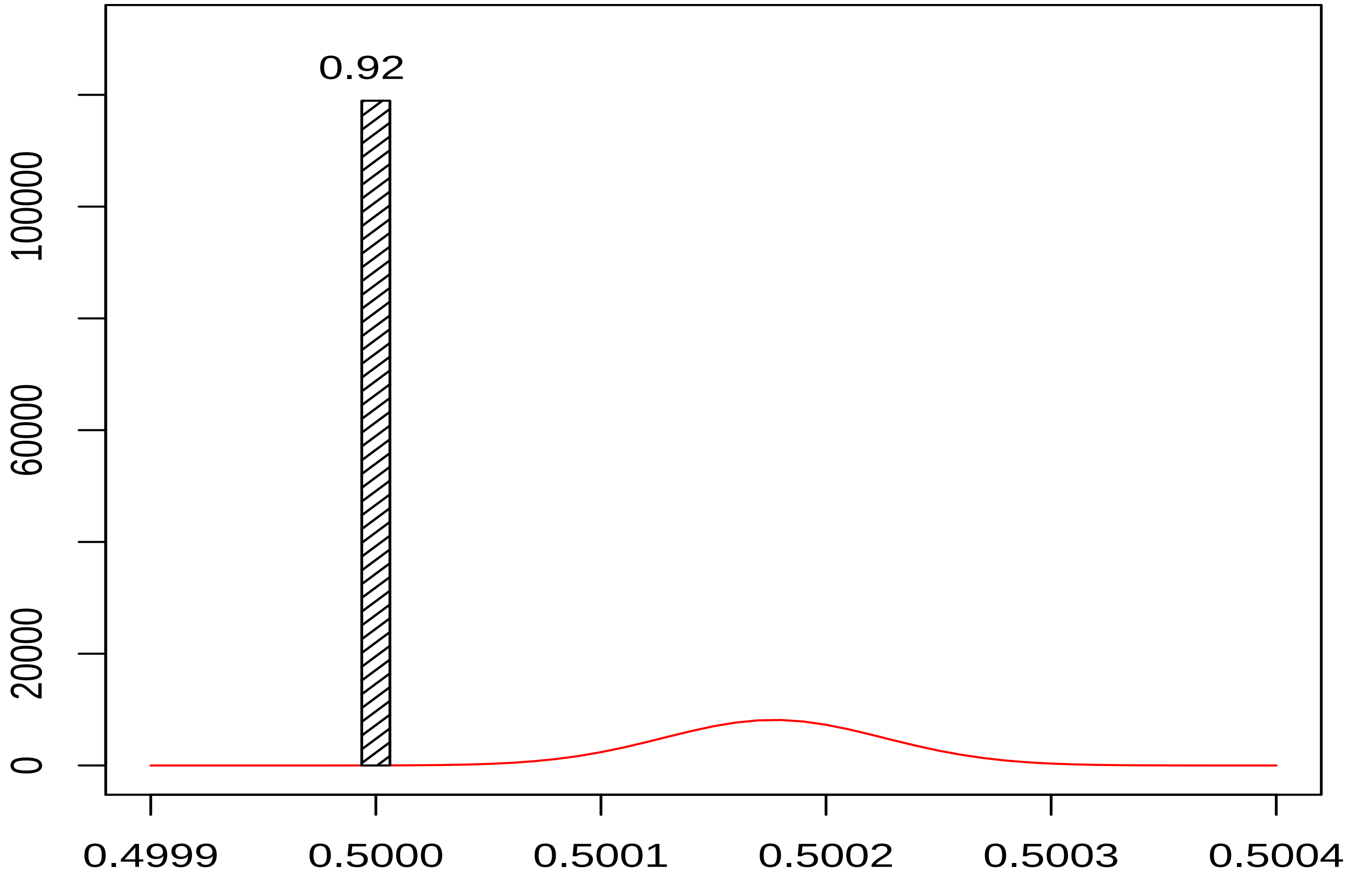
$Pr(H_0 \mid x = 52,263,470) \approx 0.92$

(recall, p-value $\approx .0003$)

Posterior density on $H_1 : \theta \neq \frac{1}{2}$ is

$$\pi(\theta|x, H_1) \propto \pi(\theta) f(x \mid \theta) \propto 1 \times \theta^x (1-\theta)^{n-x},$$

the $Be(\theta \mid 52,263,471\ ,\ 52,226,531)$ density.


0.92
0
20000
60000
100000
0.4999
0.5000
0.5001
0.5002
0.5003
0.5004

Bayes Factor:

$$B_{01} = \frac{\text{likelihood of observed data under H}_0}{\text{`average' likelihood of observed data under H}_1}$$
$$= \frac{f(x \mid \theta = \frac{1}{2})}{\int_0^1 f(x \mid \theta)\pi(\theta)d\theta} \approx 12$$

Crash of frequentist and Bayesian conclussions is dramatic.

- $\alpha_0 \leq 0.5$ requires $\pi_1 \geq 0.92$,
- alternatively $\rightsquigarrow$ require a $\pi(\theta)$ under H_1 *extremely* concentrated around $H_0 : \theta = 0.5$ (that is, both hypothesis would then be very precise)

Choice of the prior density or weight function, π, on $\{\theta : \theta \neq \frac{1}{2}\}$

Consider $\pi_r(\theta) = U(\theta \mid \frac{1}{2} - r, \frac{1}{2} + r)$ the uniform density on $(\frac{1}{2} - r, \frac{1}{2} + r)$

Subjective interpretation: r is the largest chance in success probability that you would expect, given that ESP exists. And you give equal probability to all θ in the interval $(\frac{1}{2} - r, \frac{1}{2} + r)$.

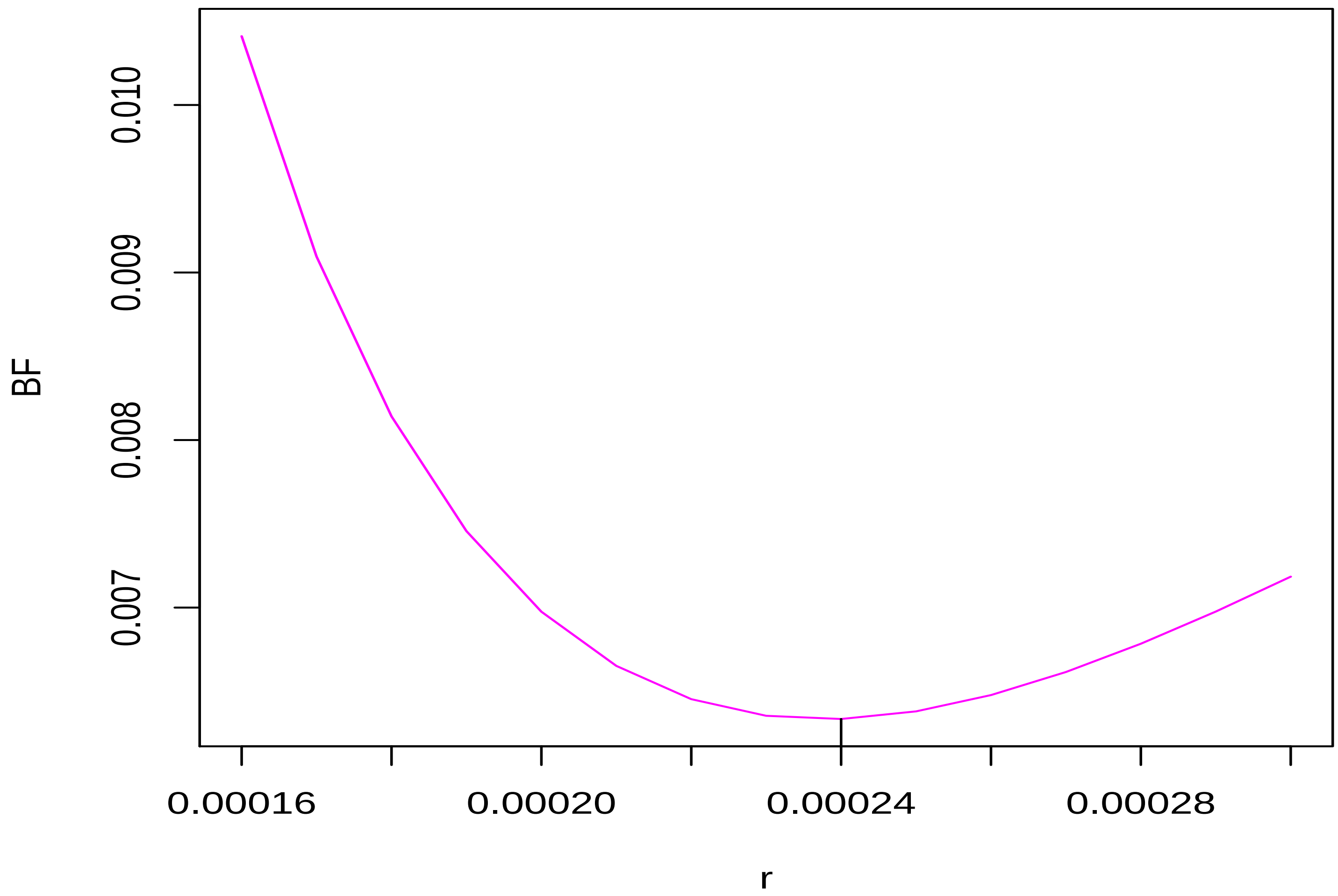
Resulting Bayes factor (letting $FBe(\cdot \mid a, b)$ denote the CDF of the Beta(a, b) distribution)

$$B(r) = \frac{f(x \mid 1/2)}{\int_0^1 f(x \mid \theta)\pi_r(\theta)\, d\theta} = \binom{n}{x} \frac{(n+1)\, r}{2^{n-1}} \left[FB_2 - FB_1\right]^{-1}$$

where

$FB_2 = FBe(\frac{1}{2} + r \mid x+1, n-x+1)$ and
$FB_1 = FBe(\frac{1}{2} - r \mid x+1, n-x+1)$

For example, $B(0.25) \approx 6$



BF
0.010
0.009
0.008
0.007
0.00016
0.00020
0.00024
0.00028
r

r = largest increase in success probability that would be expected, given ESP exists.

the minimum value of $B(r)$ is $\frac{1}{158}$, attained at the minimizing value of $r = .00024$

Conclusion: Although the p-value is small (.0003), for typical prior beliefs the data would provide evidence *for* the simpler model H_0 : no ESP. Only if one believed a priori that $|\theta - \frac{1}{2}| \leq .0021$, would the evidence for H_1 be at least 20 to 1.