Bayesian and Conditional Frequentist Hypothesis Testing and Model Selection

James O. Berger

Duke University, USA

VIII C.L.A.P.E.M. La Habana, Cuba, November 2001

Overall Outline

- Basics
- Motivation for the Bayesian approach to model selection and hypothesis testing
- Conditional frequentist testing
- Methodologies for objective Bayesian model selection
- Testing when there is no alternative hypothesis

Basics

- Brief History of Bayesian statistics
- Basic notions of Bayesian hypothesis testing (through an example)
- Difficulties in interpretation of p-values
- Notation and example for model selection

Brief History of (Bayesian) Statistics

1760 – 1920 : Statistical Inference was primarily Bayesian

Bayes (1764) : Binomial distribution, $\pi(\theta) = 1$.

Laplace (..., 1812) : Many distributions, $\pi(\theta) = 1$.

⋮

Edgeworth

⋮

Pearson

⋮

A curiosity - the name:

1764 – 1838: called "probability theory"

1838 – 1945: called "inverse probability" (named by Augustus de Morgan

1945 – : called "Bayesian analysis"

1929 – 1955 : Fisherian and Frequentist approaches developed and became dominant, because now one could do practical statistics "without the logical flaws resulting from always using $\pi(\theta) = 1$."

Voices in the wilderness:

Harold Jeffreys: fixed the logical flaw in inverse probability (objective Bayesian analysis)

Bruno de Finetti and others: developed the logically sound subjective Bayes school.

1955 – : Reemergence of Bayesian analysis, and development of Bayesian testing and model selection.

Psychokinesis Example

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.

- Stream of particles arrive at a 'quantum gate'; each goes on to either a red or a green light
- Quantum mechanics implies particles are 50/50 to go to each light
- Individual tries to "influence" particles to go to red light

Data and model:

- Each "particle" is a Bernoulli trial (red = 1, green = 0)
  $\theta =$ probability of "1"
  $n = 104,900,000$ trials
  $X = \#$ "successes" ($\#$ of 1's), $\quad X \sim \text{Binomial}(n, \theta)$
  $x = 52,263,000$ is the actual observation

  To test $H_0 : \theta = \frac{1}{2}$ (subject has no influence)

  versus $H_1 : \theta \neq \frac{1}{2}$ (subject has influence)

- P-value $= P_{\theta=\frac{1}{2}}(X \geq x) \approx .0003$.
  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

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

$\pi(\theta) = 1$ (on $0 < \theta < 1$)

Posterior distribution:

$$Pr(H_0|x) = \text{probability that } H_0 \text{ true, given data, } x$$

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

For the objective prior,

$$Pr(H_0|x = 52,263,000) \approx 0.94$$

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

Key ingredients for Bayesian Inference:

- the model for the data
- the prior $\pi(\theta)$
- ability to compute or approximate integrals

Bayes Factor:

An 'objective' alternative to choosing $Pr(H_0) = Pr(H_1) = \frac{1}{2}$ is to report the *Bayes factor*

$$B_{01} = \frac{\text{likelihood of observed data under } \mathrm{H}_0}{\text{'average' likelihood of observed data under } \mathrm{H}_1}$$

$$= \frac{f(x \mid \theta{=}\frac{1}{2})}{\int_0^1 f(x \mid \theta)\pi(\theta)d\theta} \approx 15.4$$

Note: $\underset{\text{(posterior odds)}}{\frac{Pr(H_0|x)}{Pr(H_1|x)}} = \underset{\text{(prior odds)}}{\frac{Pr(H_0)}{Pr(H_1)}} \times \underset{\text{(Bayes factor)}}{B_{01}}$

so B_{01} is often thought of as "the odds of H_0 to H_1 provided by the data"

Bayesian Reporting in Hypothesis Testing

- The complete posterior distribution is given by
  - $Pr(H_0|x)$, the posterior probability of null hypothesis
  - $\pi(\theta|x, H_1)$, the posterior distribution of θ under H_1
- A useful $summary$ of the complete posterior is
  - $Pr(H_0|x)$
  - C, a (say) 95% posterior credible set for θ under H_1
- In the psychokinesis example
  - $Pr(H_0|x) = .94 \rightsquigarrow$ gives the probability of H_0
  - $C = (.50008, .50027) \rightsquigarrow$ shows where θ is if H_1 is true
- For testing precise hypotheses, confidence intervals alone are not a satisfactory inferential summary

Crucial point: In this example, $\theta = .5$ (or $\theta \approx .5$) is plausible. If $\theta = .5$ has no special plausibility, a different analysis will be called for.

Example: Quality control for truck transmissions

- $\theta =$ % of transmissions that last at least 250,000 miles
- the manufacturer wants to report that θ is at least $1/2$
- test $H_0 : \theta \geq 0.5$ vs $H_1 : \theta < 0.5$
  Here $\theta = .5$ has no special plausibility.

Note: whether one does a one-sided or two-sided test is not very important. What is important is whether or not a point null has special plausibility.

Clash between p-values and Bayesian answers

In the example, the p-value $\approx .0003$, but the posterior probability of the null ≈ 0.94 (equivalently, the Bayes factors gives $\approx$ 15.4 odds in favor of the null). Could this conflict be because of the prior distribution used?

- But it was a neutral, $objective$ prior.
- Any sensible prior produces Bayes factors orders of magnitude larger than the p-value. For instance, any symmetric (around $.5$), unimodal prior would produce a Bayes factor $at\ least\ 30$ times larger than the p-value.

P-values also fail frequentist evaluations

An Example

- Experimental drugs $D_1, D_2, D_3, \ldots \rightsquigarrow$ are to be tested (same illness or different illnesses; independent tests)
- For each drug, test

$$H_1: \quad D_i \text{ has negligible effect } \quad vs$$
$$H_2: \quad D_i \text{ is effective}$$

- Observe (independent) data for each test, and compute each p-value (the probability of observing hypothetical data as or more "extreme" than the actual data).

- Results of the Tests of the Drugs:

TREATMENT	D1	D2	D3	D4	D5	D6
P-VALUE	0.41	**0.04**	0.32	0.94	**0.01**	0.28
TRATAMIENTO	D7	D8	D9	D10	D11	D12
P-VALOR	0.11	**0.05**	0.65	**0.009**	0.09	0.66

- Question: How strongly do we believe that D_i has a non negligible effect when:
  (i) the p-value is approximately $.05$?
  (ii) the p-value is approximately $.01$?

A surprising fact:

Suppose it is known that, a priori, about 50% of the D_i will have negligible effect. Then

(i) of the D_i for which **p-value $\approx$ 0.05, at least 25%** (and typically over $\mathbf{50\%}$) will have negligible effect;

(ii) of the D_i for which **p-value $\approx$ 0.01, at least 7%** (and typically over $\mathbf{15\%}$) will have negligible effect;

(Berger and Sellke, 1987, Berger and Delampady, 1987)

An interesting simulation for normal data:

- Generate random data from H_0, and see where the p-values fall.
- Generate random data from H_1, and see where the p-values fall.

Under H_0, such random p-values have a uniform distribution on $(0, 1)$.

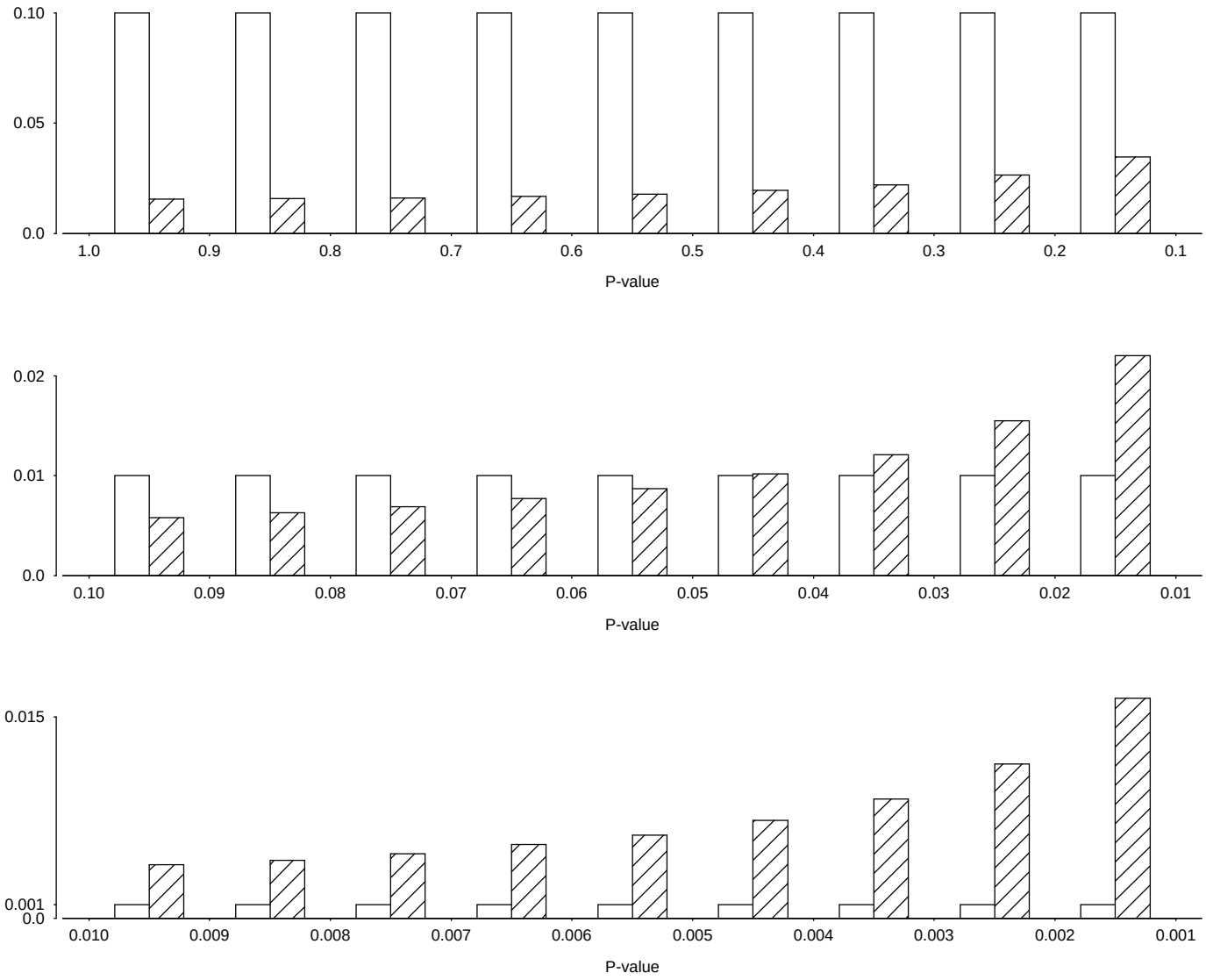
Under H_1, "random data" might arise from either:

(i) Picking a $\theta \neq 0$ and generating the data

(ii) Picking some sequence of θ's and generating a corresponding sequence of data

(i) Picking a distribution $\pi(\theta)$ for θ; generating θ's from π; and then generating data from these θ's

EXAMPLE: Picking $\pi(\theta)$ to be $N(0, 2)$ and $n = 20$, yields the following fraction of p-values in each interval


0.10
0.05
0.0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
P-value
0.02
0.01
0.0
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
P-value
0.015
0.001
0.0
0.010
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
P-value

Surprise : No matter how one generates "random" p-values under H_1, $at\ most$ 3.4% will fall in the interval $(.04, .05)$, so a p-value near .05 (i.e., $|t|$ near 1.96) provides $at\ most$ 3.4 to 1 odds in favor of H_1

(Berger and Sellke, J. Amer.Stat. Assoc., 1987)

Message : Knowing that data is "rare" under H_0 is of little use unless ones determines whether or not it is also "rare" under H_1.

In practice : For moderate or large sample sizes, data under H_1, for which the p-value is between 0.04 and 0.05, is typically as rare or more rare than data under H_0, so that odds of about 1 to 1 are then reasonable.

The previous simulation can be performed on the web, using an applet available at:

http://www.stat.duke.edu/∼berger

Notation for general model selection

Models (or hypotheses) for data $\mathbf{x}$: $M_1, M_2, \ldots, M_q$

Under model M_i :

Density of $\mathbf{X}$: $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$, $\boldsymbol{\theta}_i$ unknown parameters

Prior density of $\boldsymbol{\theta}_i$: $\pi(\boldsymbol{\theta}_i)$

Prior probability of model M_i: $P(M_i)$, ($= \frac{1}{q}$ here)

Marginal density of $\mathbf{X}$: $m_i(\mathbf{x}) = \int f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i$

Posterior density: $\pi(\boldsymbol{\theta}_i|\mathbf{x}) = f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)/m_i(\mathbf{x})$

Bayes factor of M_j to M_i: $B_{ji} = m_j(\mathbf{x})/m_i(\mathbf{x})$

Posterior probability of M_i:

$$P(M_i \,|\, \mathbf{x}) = \frac{P(M_i)m_i(\mathbf{x})}{\sum_{j=1}^{q} P(M_j)m_j(\mathbf{x})} = \left[\sum_{j=1}^{q} \frac{P(M_j)}{P(M_i)} B_{ji}\right]^{-1}$$

Particular case : $P(M_j) = 1/q$:

$$P(M_i \,|\, \mathbf{x}) = \bar{m}_i(\mathbf{x}) = \frac{m_i(\mathbf{x})}{\sum_{j=1}^{q} m_j(\mathbf{x})} = \frac{1}{\sum_{j=1}^{q} B_{ji}}$$

Reporting : It is useful to separately report $\{\bar{m}_i(\mathbf{x})\}$ and $\{P(M_i)\}$. Knowing the $\bar{m}_i$ allows computation of the

$$P(M_i \,|\, \mathbf{x}) = \frac{P(M_i)\, \bar{m}_i(\mathbf{x})}{\sum_{j=1}^{q} P(M_j)\, \bar{m}_j(\mathbf{x})}$$

for any prior probabilities.

Example: (location-scale)

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

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

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

Difficulty: Since these models are not nested, and since there are no low dimensional sufficient statistics for all models, classical model selection is difficult and would typically rely on asymptotics.

Prior distribution: choose

$P(M_i) = \frac{1}{q} = \frac{1}{5}$;

the noninformative prior $\pi(\boldsymbol{\theta}_i) = \pi(\mu, \sigma) = \frac{1}{\sigma}$

Marginal distributions, $m^N(\mathbf{x}|M)$, for these models can then be calculated in closed form.

Here $m^N(\mathbf{x}|M) = \int \prod_{i=1}^n \left[\frac{1}{\sigma}\, g\left(\frac{x_i-\mu}{\sigma}\right)\right] \frac{1}{\sigma} \quad d\mu\, d\sigma$. For the five models, the marginals are

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

Consider four classic data sets:

- Darwin's data ($n = 15$)
- Cavendish's data ($n = 29$)
- Stigler's Data Set 9 ($n = 20$)
- A randomly generated Cauchy sample ($n = 14$)

The objective posterior probabilities of the five models, for each data set,

$$\bar{m}^N(\mathbf{x}|M) = \frac{m^N(\mathbf{x}|M)}{m^N(\mathbf{x}|M_N) + m^N(\mathbf{x}|M_U) + m^N(\mathbf{x}|M_C) + m^N(\mathbf{x}|M_L) + m^N(\mathbf{x}|M_R)},$$

are as follows:

Data	Models				
set	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}

Part I Conclusions

- $Pr(H_0 \mid x)$ or B can be $much\ larger$ than a corresponding p-value.
- Bayes factors are useful for communicating "what the data says," separate from the $Pr(H_i)$.
- When testing, one must think carefully about whether a 'precise null' is believable.
- One cannot generally test from confidence intervals.

Motivation for the Bayesian approach to Model Selection and Hypothesis Testing

Outline

- 'Standard' motivations
- Sequential analysis
- Conditional frequentist testing

Why Bayes Factors (or $\bar{m}_i$) ?

1. Easy to interpret (in contrast with, say, p-values)
2. Consistency
   A. If one of the M_i is true, then $\bar{m}_i \to 1$ as $n \to \infty$
   B. If some other model, M^*, is true, $\bar{m}_i \to 1$ for that model closest to M^* in Kullback-Leibler divergence

   (Berk, 1966, Dmochowski, 1994)

3. Predictive optimality
   A. The $\bar{m}_i$ are the basic quantities in predictive model averaging schemes.
   B. If a single model is to be chosen, that with the largest $\bar{m}_i$ is often best predictively.
      Example. Suppose $q = 2$ and it is desired to predict a future observation Y under loss $L(|\hat{Y} - Y|)$, where L is nondecreasing. If $P(M_1) = P(M_2)$, then the model with the largest $\bar{m}_i$ gives the best prediction.
   C. In other contexts (e.g., nonparametric regression), the optimal predictive model is the *median probability model* (the model that includes only variables whose posterior model inclusion probabilities exceed 1/2).
4. Bayes Factors automatically seek parsimony; no adhoc penalties for model complexity are needed.

5. The approach is viable for any models, nested or non-nested, regular or irregular, large or small sample sizes, two or multiple models.
   Example: In the location-scale example, we chose among five non-nested models, with small sample sizes, irregular asymptotics, and no sufficient statistics (of fixed dimension).
6. Ease of application in sequential scenarios (no need to 'spend α' when looking at the data)
7. In many important situations involving testing of M_1 versus M_2, $\bar{m}_1$ and $\bar{m}_2$ are 'optimal' conditional frequentist error probabilities of Type I and Type II, respectively.

6. Sequential analysis is no more difficult

EX: SEQUENTIALY TESTING MULTIPLE HYPOTHESIS

Data: d_i is the observed treatment difference for subject i treated with two hypotensive agents (Robertson and Armitage, 1959; Armitage, 1975). (Here t denotes the t-statistic and s_d the sample standard deviation.)

Model: The d_i are i.i.d $Normal(\theta, \sigma^2)$, $i = 1, \ldots, 53$.

To Test: $H_1 : \theta = 0$ versus $H_2 : \theta < 0$ versus $H_3 : \theta > 0$.

Prior inputs: $P(H_i) = 1/3$; noninformative prior for (θ, σ^2) appropriately 'trained' ('Encompassing Intrinsic Bayes Factors': Berger & Pericchi, 96, Berger & Mortera, 99).

Posterior Probabilities, P_i, for the H_i:

$$P_1 = \left[1 + \frac{s_1}{t_{n-1}(t)}\left(\frac{1 - T_{n-1}(t)}{s_2} + \frac{T_{n-1}(t)}{s_3}\right)\right]^{-1},$$

$$P_2 = \left[1 + \frac{s_2}{1 - T_{n-1}(t)}\left(\frac{t_{n-1}(t)}{s_1} + \frac{T_{n-1}(t)}{s_3}\right)\right]^{-1},$$

and $P_3 = 1 - P_1 - P_2$, where t_{n-1} and T_{n-1} are the density and c.d.f. of the standard t-distribution with $(n-1)$ d.f.,

$s_3 = \pi n(n-1) - s_2$,

$$s_1 = \frac{s_d}{\sqrt{n}} \sum_{i \neq j} \frac{|d_i - d_j|}{{d_i}^2 + {d_j}^2 + \epsilon}, \quad s_2 = \sum_{i \neq j} \left(\frac{\pi}{2} - \arctan\left(\frac{-(d_i + d_j)}{|d_i - d_j + \epsilon|}\right)\right).$$

($\epsilon \approx 0$ is introduced to avoid numerical indeterminacy)

Pair	Difference	t-statistic	Posterior Probabilities		
n	d_i	t	P_1	P_2	P_3
5	1	1.01	0.360	0.148	0.492
6	12	1.15	0.342	0.142	0.516
7	11	1.26	0.348	0.132	0.519
8	-2	1.23	0.276	0.136	0.589
9	6	1.30	0.283	0.130	0.587
10	14	1.44	0.295	0.115	0.590
11	19	1.63	0.291	0.095	0.615
12	71	2.05	0.203	0.058	0.739
13	-9	1.92	0.229	0.058	0.713
14	7	1.97	0.225	0.054	0.721
15	-19	1.74	0.294	0.061	0.646
20	-9	1.51	0.387	0.056	0.557
25	0	1.35	0.465	0.060	0.475
30	-3	0.831	0.620	0.079	0.301
35	0	0.339	0.669	0.112	0.219
40	0	0.056	0.698	0.134	0.168
50	-3	-0.202	0.736	0.141	0.123
53	-37	-0.396	0.740	0.157	0.103

COMMENTS

(i) Neither multiple hypotheses nor the sequential aspect caused difficulties. There is no penalty (e.g., 'spending α') for looks at the data.

(ii) Quantification of the support for $H_1 : \theta = 0$ is direct. At the 12th observation, $t = 2.05$ but $P_1 = 0.203$. At the end, $P_1 = 0.740$.

(iii) At the 12th observation, $P_2 = 0.058$, so H_2 can be effectively ruled out.

(iv) For testing $H_1 : \theta = 0$ versus $H_2 : \theta \neq 0$, the P_i are conditional frequentist error probabilities.

7. Conditional frequentist testing

Motivation:

- Conditional frequentist testing can be the same as Bayesian testing (unification), but is fully frequentist.
- It is more intuitively appealing than unconditional testing.
- It is much easier to apply, especially in sequential clinical trials.

Problem:

Data $\mathbf{X}$;

To test: $H_0 : \mathbf{X} \sim f_0(\mathbf{x})$ versus $H_1 : \mathbf{X} \sim f_1(\mathbf{x})$.

The need for good conditional performance

Artificial Example: Observe X_1 and X_2, where

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2. \end{cases}$$

Consider the confidence set for θ

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2. \end{cases}$$

Unconditional coverage:

$$P_\theta(C(X_1, X_2) \text{ contains } \theta) = 0.75.$$

Conditional coverage:

$$\begin{aligned} P_\theta(C(X_1, X_2) \text{ contains } \theta \,|\, X_1 \neq X_2) &= 1 \\ P_\theta(C(X_1, X_2) \text{ contains } \theta \,|\, X_1 = X_2) &= \tfrac{1}{2}. \end{aligned}$$

Note: The unconditional coverage arises as the expected value of the conditional coverage.

Some philosophy behind Conditional Frequentist Theory

Frequentist Principle: In repeated use of a statistical procedure, the "average" actual error should not be greater than the "average" reported error.

The key to success: Defining a sensible "average."

Example (Normal testing): Averaging over all $|z| > 1.96$ and reporting error $\alpha = 0.05$ is not a sensible "average" since $|z| = 1.96$ and $|z| = 10$ indicate very different levels of error. One should somehow separately "average" over these extremes. (Note: p-values try to do this, but badly violate the Frequentist Principle)

- *Basic question:* What is the sequence of possible data for which to consider frequentist evaluations?
  (Fisher: "relevant subset;" Lehmann: "frame of reference.")
- Implementation: Find a statistic, $S(x)$, whose magnitude indicates the "strength of evidence" in x. Compute error probabilities as

$$\begin{aligned} \alpha(s) &= \text{Type I error prob, given } S(x) = s \\ &= P_0(\text{Reject } H_0 | S(x) = s) \\ \beta(s) &= \text{Type II error prob, given } S(x) = s \\ &= P_1(\text{Accept } H_0 | S(x) = s) \end{aligned}$$

History of conditional frequentist testing

- Many Fisherian precursors based on conditioning on ancillary statistics and on other statistics (e.g., the Fisher exact test, conditioning on marginal totals).
- General theory in Kiefer (1977 JASA); the key step was in understanding that almost *any* conditioning is formally allowed within the frequentist paradigm (ancillarity is not required).
- Brown (1978) found optimal conditioning frequentist tests for 'symmetric', simple hypotheses.
- Berger, Brown and Wolpert (1994 AOS) developed the theory discussed herein for testing simple hypotheses.

Our suggested choice of the conditioning statistic

- Let p_i be the p-value from testing H_i against the other hypothesis; use the p_i as measures of 'strength of evidence,' as Fisher suggested.
- Define the conditioning statistic $S = \max\{p_0, p_1\}$; its use is based on deciding that data (in either the rejection or acceptance regions) with the same p-value has the same 'strength of evidence.'

The conditional frequentist test (T^C)

- Accept H_0 when $p_0 > p_1$, and reject otherwise.
- Compute Type I and Type II conditional error probabilities (CEPs) as

$$\begin{aligned}\alpha(s) &= P_0(\text{rejecting } H_0|S=s) \equiv P_0(p_0 \leq p_1|S(X)=s)\\ \beta(s) &= P_1(\text{accepting } H_0|S=s) \equiv P_1(p_0 > p_1|S(X)=s).\end{aligned}$$

Unification of the Fisherian, frequentist, and Bayesian philosophies in testing

- The evidentiary content of p-values is acknowledged, but 'converted' to error probabilities by conditioning.
- The conditional error probabilities $\alpha(s)$ and $\beta(s)$ are fully data-dependent, yet fully frequentist.
- $\alpha(s)$ and $\beta(s)$ are exactly equal to the (objective) posterior probabilities of H_0 and H_1, respectively, so the conditional frequentists and Bayesians report the same error (Berger, Brown and Wolpert, 1994 AOS).

A simple example

Sellke, Bayarri and Berger (2001 American Statistician)

- $H_0 : X \sim$ Uniform$(0, 1)$ vs. $H_1 : X \sim$ Beta$(1/2, 1)$.
- $B(x) = \frac{1}{(2\sqrt{x})^{-1}} = 2\sqrt{x}$ is the likelihood ratio.
- $p_0 = P_0(X \leq x) = x$ and $p_1 = P_1(X \geq x) = 1 - \sqrt{x}$.
- Accept H_0 when $p_0 > p_1$ (i.e., when $x > .382$) and reject otherwise.
- Define $S = \max\{p_0, p_1\} = \max\{x, 1 - \sqrt{x}\}$ (so it is declared that, say, $x = \frac{3}{4}$ has the same "strength of evidence" as $x = \frac{1}{16}$).

- The conditional test is

$$T^C = \begin{cases} \text{if } x \leq 0.382, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha(x) = (1 + \frac{1}{2}x^{-1/2})^{-1}\,; \\ \text{if } x > 0.382, & \text{accept } H_0 \text{ and report Type II CEP} \\ & \beta(x) = (1 + 2x^{1/2})^{-1}. \end{cases}$$

- Note that $\alpha(x)$ and $\beta(x)$ are precisely the objective Bayesian posterior probabilities of H_0 and H_1, respectively.

Other types of conditioning

- Ancillary S rarely exist and, when they exist, can result in unnatural conditional error probabilities (e.g., in the above example, $\beta(x)$ equals the constant $1/2$ as the likelihood ratio $B(x)$ varies from 1 to 2).
- Birnbaum suggested 'intrinsic significance,' conditioning defined through likelihood concepts. This rarely works (e.g., in the above example $\alpha(x) = 1$ when $B(x)$ varies between 0 and $1/2$).
- Kiefer (1977) suggested 'equal probability continuum' conditioning, which also fails in the above example.

Generalizations

- Berger, Boukai and Wang (1997a, 1997b) generalized to simple versus composite hypothesis testing, including sequential settings.
- Dass (1998) generalized to discrete settings.
- Dass and Berger (2001) generalized to two composite hypotheses, under partial invariance.
- Berger and Guglielmi (2001) used the ideas to create a finite sample exact nonparametric test of fit.

An aside: Using this to calibrate p-values when no alternative hypothesis is specified (Sellke, Bayarri & Berger, 2001)

A *proper* p-value satisfies $H_0 : p(X) \sim \text{Uniform}(0, 1)$.

Consider testing this versus $H_1 : p(X) \sim \text{Beta}(\xi, 1)$.
The same argument, as earlier, yields, when $p < e^{-1}$,

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

This *lower bound* on conditional Type I error also applies under a quite general nonparametric alternative. It thus provides a generic 'calibration' of p-values.

Example: Are gamma ray bursts galactic or extra-galactic in origin?

- data in early 90's were 260 observed burst directions
- H_0: data are uniformly directionally distributed (implying extra-galactic origin)
- standard test for uniformity rejected at $p = 0.027$
- $\alpha(p) \geq (1 + [-e\,(.027)\log(.027)]^{-1})^{-1} = .21$, so the actual error rate in rejecting H_0 is $at\ least$ .21

Example: nonparametric testing of fit

Berger and Guglielmi (2001, JASA)

To test: $H_0 : X \sim \mathcal{N}(\mu, \sigma)$ vs. $H_1 : X \sim F(\mu, \sigma)$,
where F is an unknown location-scale distribution.

- Define a weighted likelihood ratio (or Bayes factor) $B(x)$, by
  - choosing a Polya tree prior for F, centered at H_0;
  - choosing the right-Haar priors $\pi(\mu, \sigma) = 1/\sigma$.
- Choose S based on p-value conditioning.

- Define the conditional frequentist test as before, obtaining

$$T^C = \begin{cases} \text{if } B(\mathbf{x}) \leq c, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha(\mathbf{x}) = B(\mathbf{x})/(1 + B(\mathbf{x}))\,; \\ \text{if } B(\mathbf{x}) > c, & \text{accept } H_0 \text{ and report the 'average'} \\ & \text{Type II CEP} \beta(\mathbf{x}) = 1/(1 + B(\mathbf{x})), \end{cases}$$

where c is the critical value at which the two p-values are equal.

Notes:

- The Type I CEP again exactly equals the objective Bayesian posterior probability of H_0.
- The Type II CEP depends on (μ, σ). However, the 'average' Type II CEP is the objective Bayesian posterior probability of H_1.
- This gives $exact$ conditional frequentist error probabilities, even for small samples.
- Computation of $B(x)$ uses the (simple) exact formula for a Polya tree marginal distribution, together with importance sampling to deal with (μ, σ).

Other features of conditional frequentist testing with p-value conditioning

- Since the CEPs are also the objective Bayesian posterior probabilities of the hypotheses,
  - they do not depend on the stopping rule in sequential analysis, so
    - their computation is much easier,
    - one does not 'spend α' to look at the data;
  - there is no danger of misinterpretation of a p-value as a frequentist error probability, or misinterpretation of either as the posterior probability that the hypothesis is true.

- One does not need to compute (or even mention) S.
  - The computation of the CEPs can be done directly, using the (often routine) objective Bayesian approach; the theory ensures the validity of the conditional frequentist interpretation.
  - Likewise, in elementary courses, it is not necessary to introduce S; one can simply give the test.
- One has $exact$ frequentist tests for numerous situations where only approximations were readily available before:
  - Computation of a CEP is easier than computation of an unconditional error probability.

- For T^C, the rejection and acceptance regions together span the entire sample space.
  - So that one might 'reject' and report CEP 0.40.
- One could, instead:
  - Specify an ordinary rejection region R (say, at the $\alpha = 0.05$ level); note that then $\alpha = \mathsf{E}[\alpha(s)1_R]$.
  - Find the 'matching' acceptance region, and define the region in the middle to be the $no\ decision$ region.
  - Construct the corresponding conditional test.
  - Note that the CEPs would not change.

Example: Sequential t-test

(Berger, Boukai and Wang, 1998)

Data $X_1, X_2, \ldots$ are i.i.d. $N(\theta, \sigma^2)$, both unknown

To test: $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$

Note: Exact unconditional frequentist tests do not exist for this problem

Default Bayes test (Jeffreys, 1961):

Prior distribution : $Pr(H_0) = Pr(H_1) = 1/2$

Under H_0, prior on σ^2 is $g_0(\sigma^2) = 1/\sigma^2$.

Under H_1, prior on (μ, σ^2) is $g_1(\mu, \sigma^2) = \frac{1}{\sigma^2}\, g_1(\mu|\sigma^2)$, where $g_1(\mu|\sigma^2)$ is Cauchy(θ_0, σ).

Motivation: Under H_1, the additional parameter θ is given a Cauchy default testing prior (centered at θ_0 and scaled by σ), while the common (orthogonal) parameter σ^2 is given the usual noninformative prior (see Jeffreys, 1961)

Bayes factor of H_0 to H_1, if one stops after observing $X_1, X_2, \ldots X_n$ $(n \geq 2)$, is

$$B_n = \left[\int_0^\infty \frac{(1+n\xi)^{(n-1)/2}\ e^{1/(2\xi)}}{\left(1+\frac{(n-1)n\xi}{n-1+t_n^2}\right)^{n/2} \sqrt{2\pi}\,\xi^{3/2}}\, d\xi \right]^{-1},$$

where t_n is the usual t-statistic.

A simplification : It is convenient to consider the sequential test in terms of the sequence $B_1, B_2, \ldots,$ instead of the original data. We monitor this sequence as data arrive, deciding when to stop the experiment and make a decision.

Note: If H_0 and H_1 have equal prior probabilities of $1/2$, then the posterior probability of H_0 is $B_n/(1+B_n)$

A surprising fact: A slight modification of this Bayesian t-test is a conditional frequentist test such that

$$\begin{aligned} \alpha(s) &= P_{\theta_0}(\text{Type I error}|s) \\ &= \frac{B_N}{1+B_N} \quad \text{(also the posterior probability of } H_0\text{)}, \end{aligned}$$

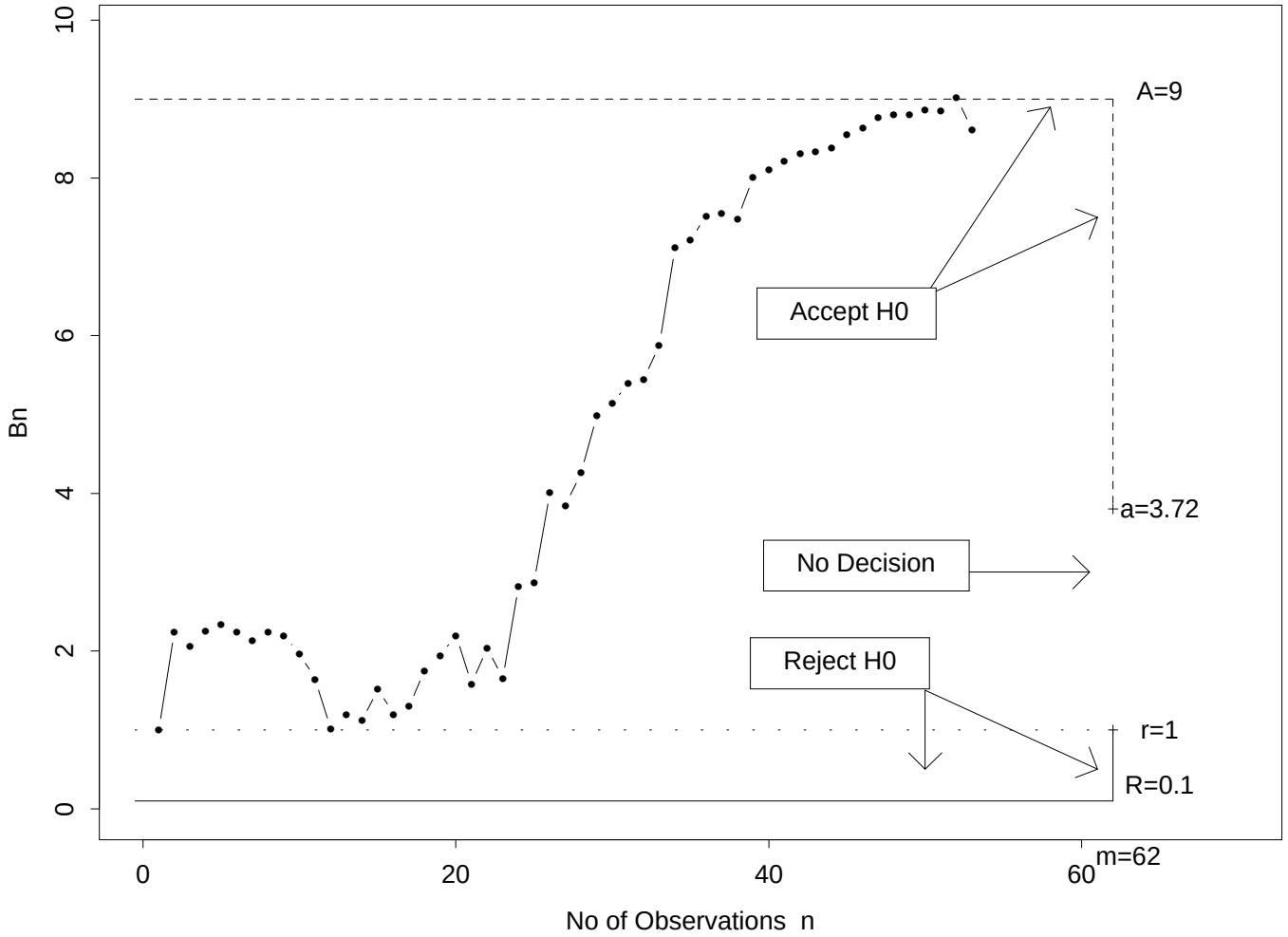
and the "average" Type II error is $\frac{1}{1+B_N}$, also the posterior probability of H_1.

(The modification, T_1^*, is the introduction of a "no decision" region in the acceptance region, where if $1 < B_N < a$, one states "no decision".)

(Berger, Boukai and Wang, *Stat.Sci* 1997, *Biometrika* 1999)

Example: The data arose as differences in time to recovery between paired patients who were administered different hypotensive agents.

Testing $H_0 : \theta = 0$ versus $H_0 : \theta \neq 0$ is thus a test to detect a mean difference in treatment effects.


Bn
No of Observations n
A=9
a=3.72
r=1
R=0.1
m=62
Accept H0
No Decision
Reject H0
0
2
4
6
8
10
0
20
40
60

The stopping rule (an example):

If $B_n \leq R$, $B_n \geq A$ or $n = M$, then stop the experiment.

Intuition:

R = "odds of H_0 to H_1" at which one would wish to stop and reject H_0.

A = "odds of H_0 to H_1" at which one would wish to stop and accept H_0.

M = maximum number of observations that can be taken

Example:

R = 0.1 (i.e., stop when 1 to 10 odds of H_0 to H_1)

A = 9 (i.e., stop when 9 to 1 odds of H_0 to H_1)

M = 62

Note: *any* stopping rule could be used just as easily.

The test T_1^* : If N denotes the stopping time,

$$T_1^* = \begin{cases} \text{if } B_N \leq 1, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha^*(B_N) = B_N/(1+B_N)\,; \\ \text{if } 1 < B_N < a, & \text{make no decision;} \\ \text{if } B_N \geq a, & \text{accept } H_0 \text{ and report the 'average'} \\ & \text{Type II CEP } \beta^*(B_N) = 1/(1+B_N). \end{cases}$$

Example: $a = 3.72$ (found by simulation). For the actual data, the stopping boundary would have been reached at time $n = 52$ $(B_{52} = 9.017 > A = 9)$, and the conclusion would have been to accept H_0, and report error probability $\beta^*(B_{52}) = 1/(1+9.017) \approx 0.100$.

COMMENTS

- T_1^* is fully frequentist (and Bayesian).
- The conclusions and stopping boundary all have simple intuitive interpretations.
- Computation is easy (except possibly computing a, but it is rarely needed in practice):
  - No stochastic process computations are needed.
  - Computations do not change as the stopping rule changes.
  - Sequential testing is as easy as fixed sample size testing.

- T_1^* essentially follows the Stopping Rule Principle, which states that, upon stopping experimentation, inference should not depend on the reason experimentation stopped.
- This equivalence of conditional frequentist and Bayesian testing applies to most classical testing scenarios.

Methodologies for Objective Bayesian Model Selection

Main Difficulty with Bayesian model selection:

Noninformative priors ($\int \pi(\theta)d\theta = \infty$) cannot generally be used for model selection or hypothesis testing. (They can be used in Bayesian estimation and predictive problems, and in hypothesis testing or model selection where the differing models have the same type of parameters, e.g., in the location-scale example.)

Solutions:

- Conventional proper priors: we illustrate with an astronomical example
- Asymptotics, such as BIC
- Training sample methods: here we discuss the "intrinsic Bayes factor" approach (Berger and Pericchi, 1996).

Conventional Proper Prior Approach

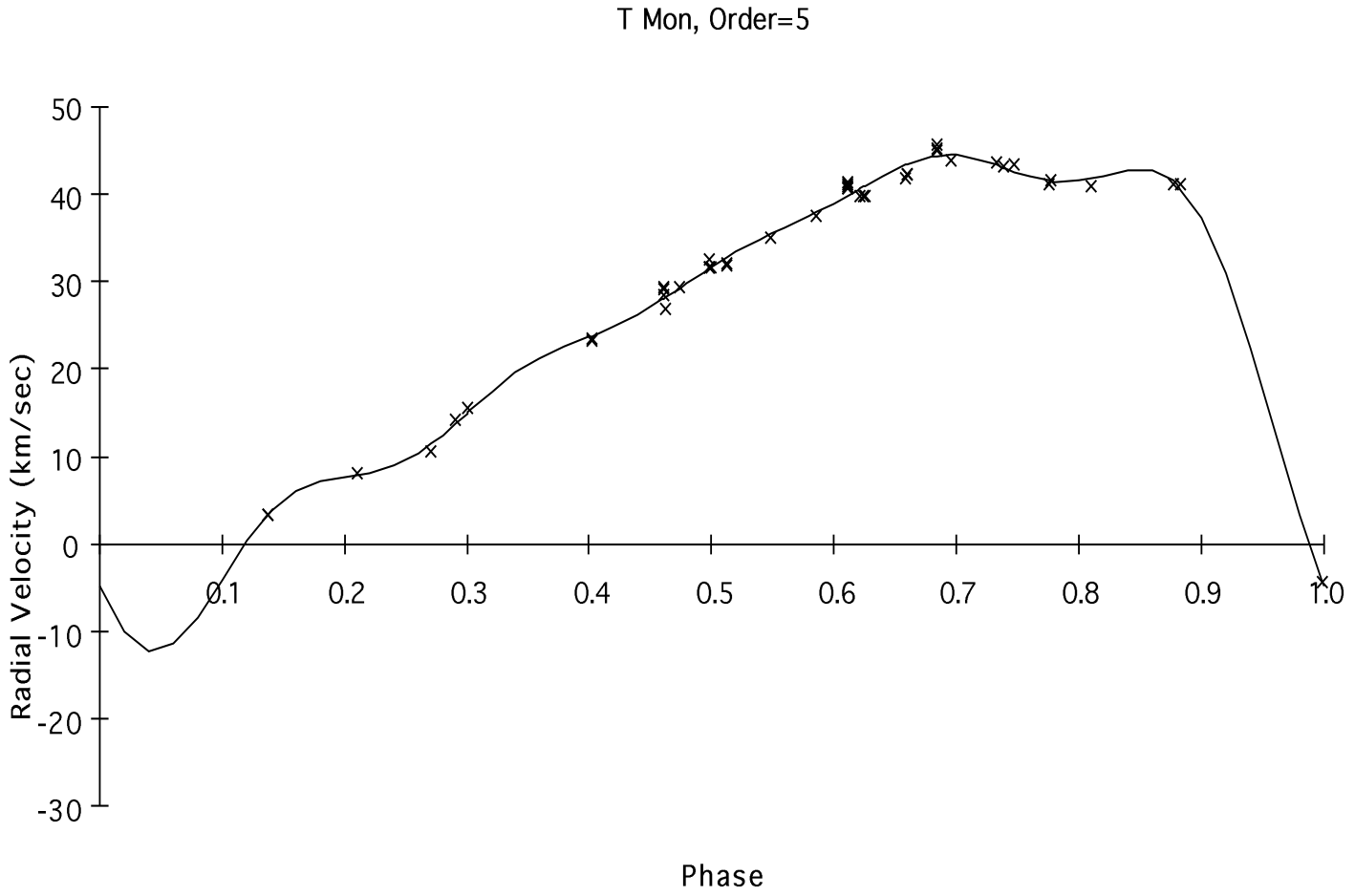
Example: *Bayesian Model Selection and Analysis for Cepheid Star Oscillations*

(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
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

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 $f_i(\boldsymbol{y} \,|\, \boldsymbol{\theta}_i)$.
- $\boldsymbol{\theta}_i$ 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

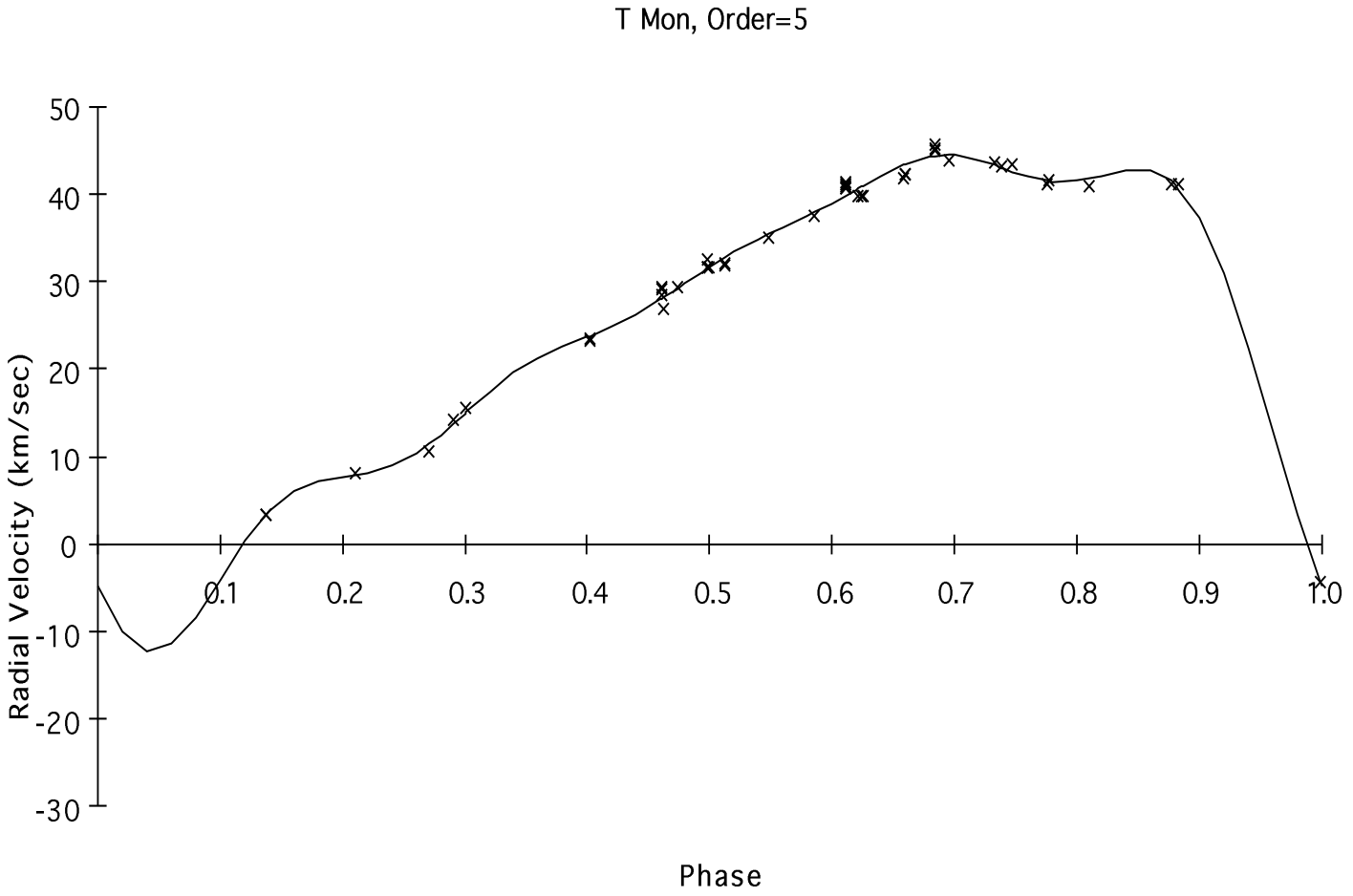
- Data:
  - m observed radial velocities U_i, $i = 1, \ldots, m$.
  - n vectors of photometry data consisting of luminosity V_i, $i = 1, \ldots, n$, and color index C_i, $i = 1, \ldots, n$.
- Specified standard deviations σ_{U_i}, σ_{V_i}, and σ_{C_i}; unknown adjustment factors are inserted, leading 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:

$$U_i \sim N(u_i, \sigma^2_{U_i}/\tau_u),$$

$$V_i \sim N(v_i, \sigma^2_{V_i}/\tau_v),$$
$$C_i \sim N(c_i, \sigma^2_{C_i}/\tau_c),$$

where u_i, v_i, and c_i denote the true unknown mean velocity, luminosity, and color index.

- Define $\boldsymbol{G}_u$ and $\boldsymbol{G}_v$ to be the known diagonal matrices of the variances $\sigma^2_{U_i}$ 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

$$\pi(\boldsymbol{\theta} \,|\, \sigma^2, \tau) \text{ is } \boldsymbol{N}(\mathbf{0}, \tau n\sigma^2(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1});$$

$$\pi(\tau) = \frac{1}{\sqrt{2\pi}\,\tau^{3/2}} \exp(-\frac{1}{2\tau}).$$

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 $\pi(s) \propto s^2 \exp\left(-|s \sin\beta|/z_0\right)$,
  - 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.

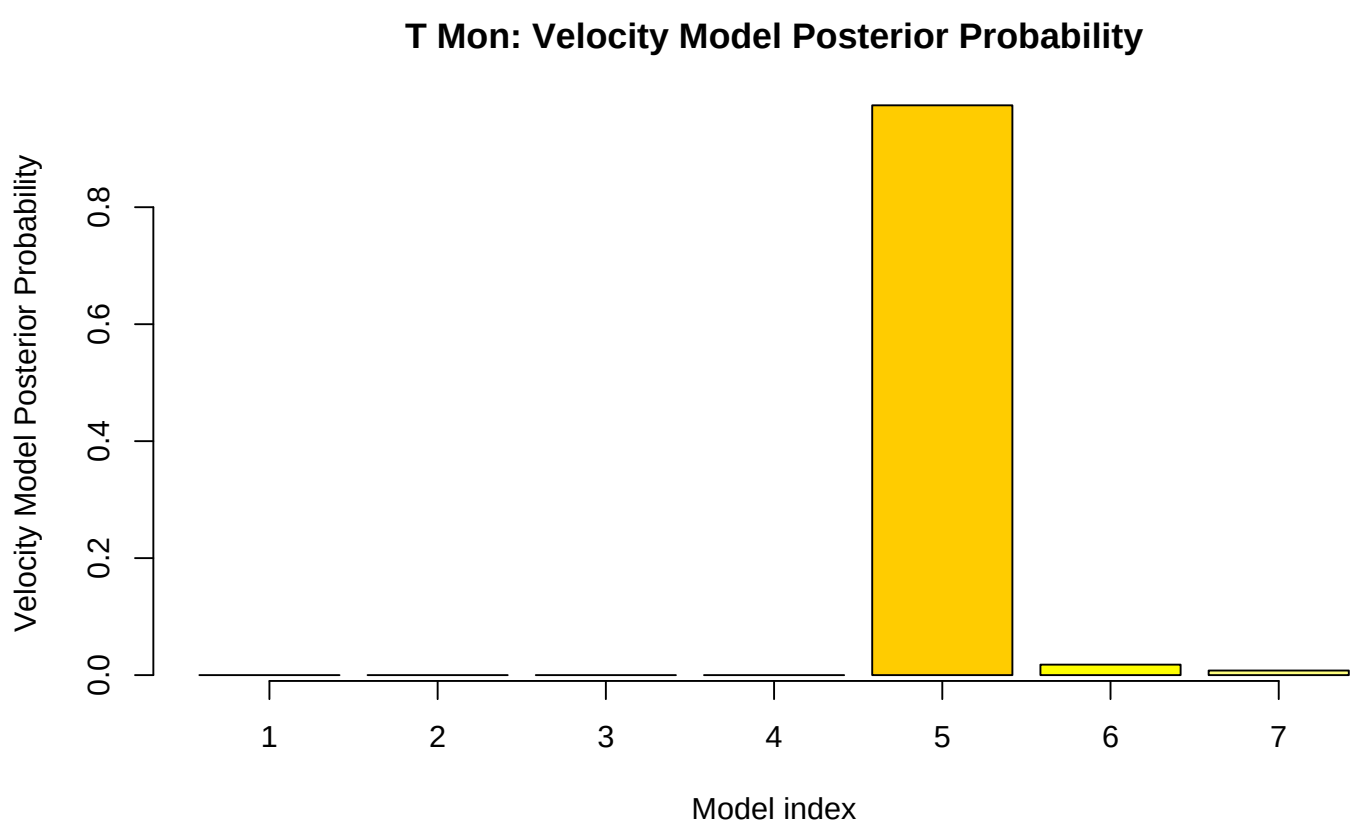
Computation

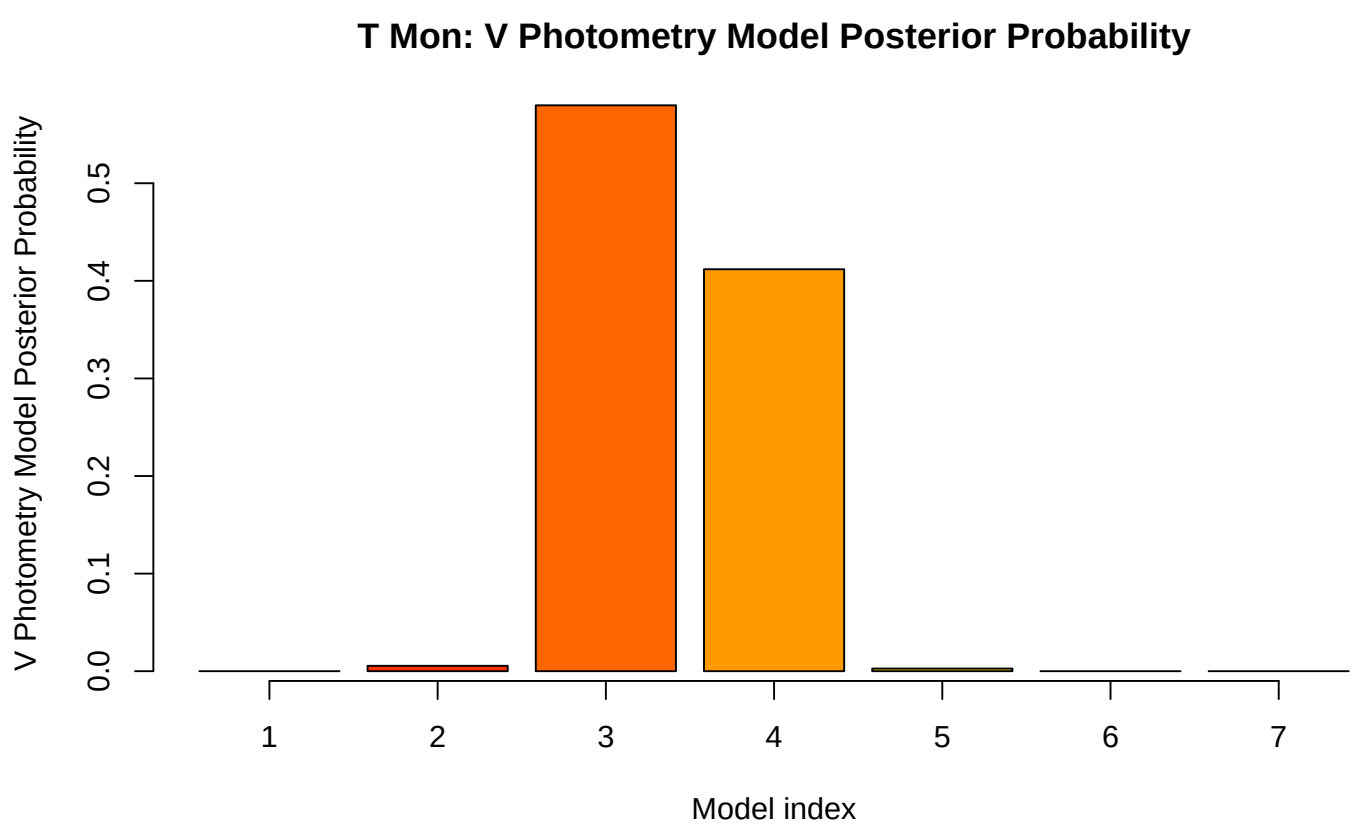
A reversible-jump MCMC algorithm of the type reviewed in Dellaportas, Forster and Ntzoufras (2000) is used to move between models and generate posterior distributions and estimates.

- 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.

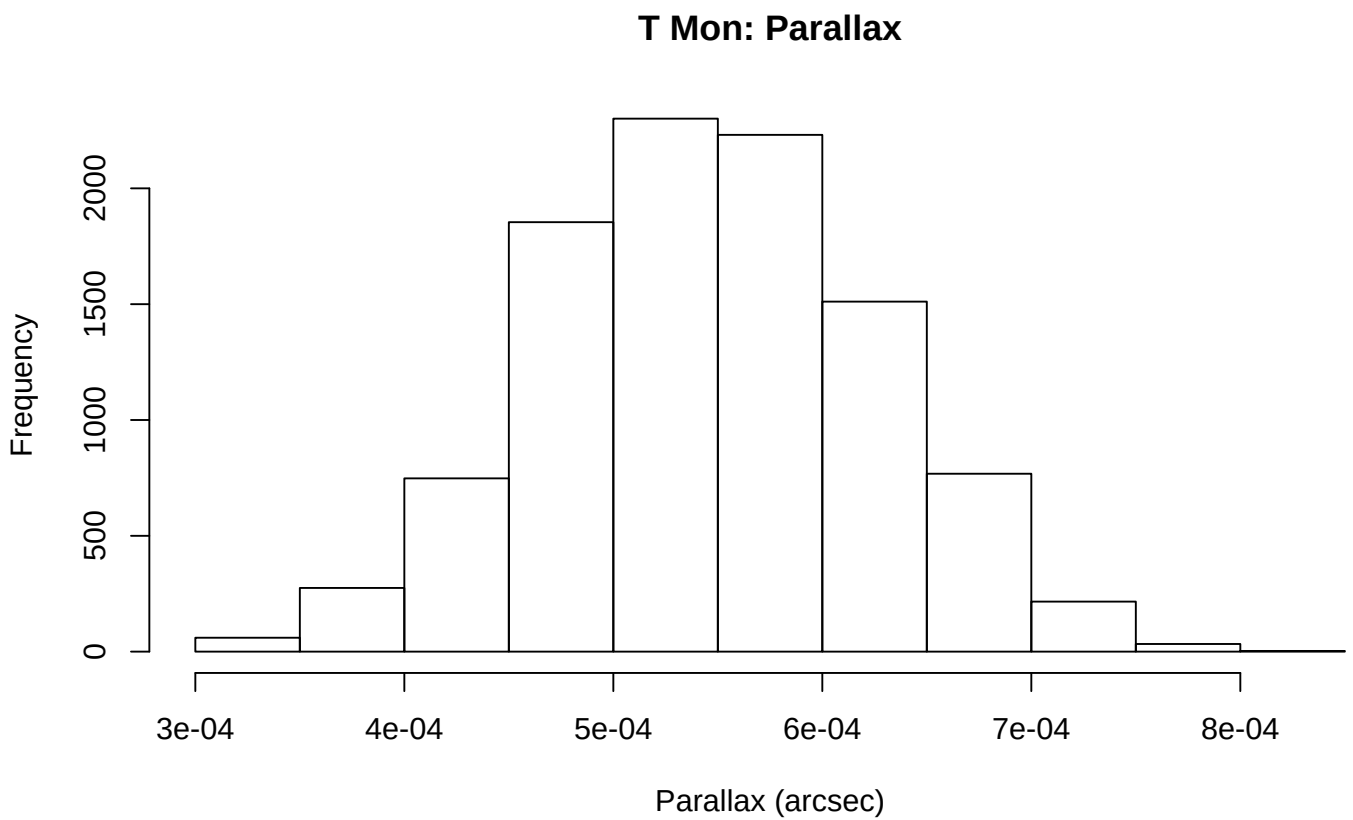
- Proposal for moves between models:
  - A 'burn-in' portion of the MCMC with uniform model proposals yielded initial posterior model probabilities, which were then used as the proposal for subsequent model moves.
  - Simultaneously, new values were proposed for the Fourier coefficients.
- The proposal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

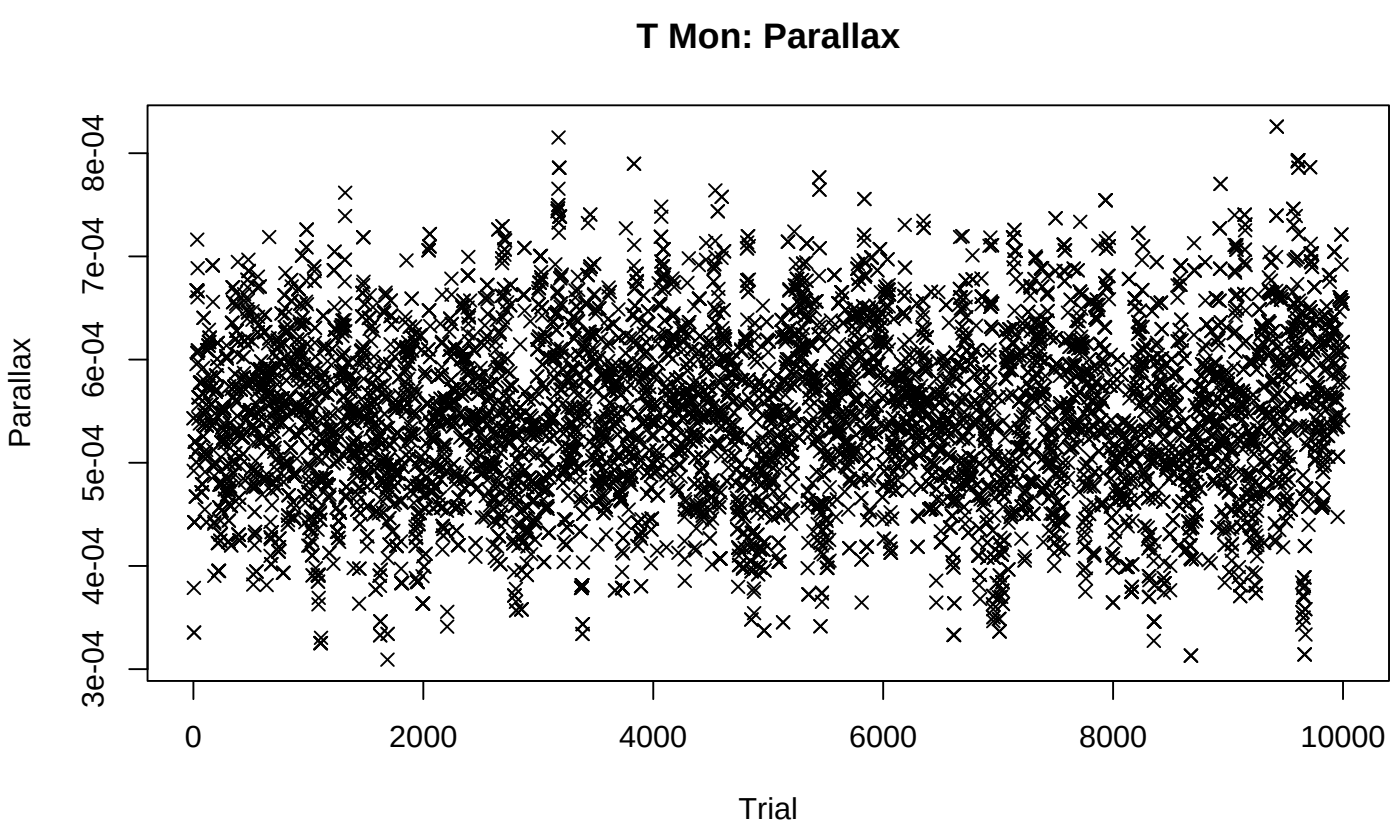

T Mon: Velocity Model Posterior Probability
Velocity Model Posterior Probability
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
Model index


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

Results for other parameters

- Estimates, standard errors, etc., for any of the unknowns or parameters in the model are also available from the MCMC computation.
- The MCMC computational strategy discussed above will automatically perform 'model-averaging' over these models, when computing posterior quantities of interest.
- The simulation history of a parameter can be graphed to check the mixing of the Markov chain (and thus assess the validity of the computation).


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

Asymptotic Approaches to Model Selection

Training Sample Methods of Model Selection

Example: the *median intrinsic Bayes factor*

(Berger and Pericchi, 1998)

Data: $X_1, X_2, \ldots, X_n$ are $N(\theta, 1)$

Models: $M_1 : \theta = 0, \quad M_2 : \theta \neq 0$

Noninformative prior:

$Pr(M_1) = Pr(M_2) = \frac{1}{2}$;

Under M_2, $\pi_2(\theta) = 1$

(Formal) Bayes factor:

$$B = \frac{f(\mathbf{x}|0)}{\int f(\mathbf{x}|\theta)\,(1)d\theta} = \frac{1}{\sqrt{n}}\, e^{-\frac{n}{2}\bar{x}^2}$$

(Formal) posterior probability of M_1:

$$P_1 = \frac{1}{1 + B^{-1}} = (1 + \sqrt{n}\, e^{-\frac{n}{2}\bar{x}^2})^{-1}$$

Redo this; probably drop the posterior probability thing, since median may make no sense here.

Note: This is not legitimate, since $\pi(\theta)$ was improper and M_1 and M_2 do not both have an unknown location parameter.

Obtaining a proper prior by use of a training sample:

Choose one observation, say x_i, and compute

$$\pi_2(\theta|x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x_i)^2} \text{ (proper)}.$$

Use this on the remaining data, $\mathbf{x}^{(i)}$, to compute the posterior probability of M_j:

$$P_1 = 1 - P_2 = [1 + \sqrt{n}\, e^{\frac{n}{2}\bar{x}^2}\, e^{-\frac{1}{2}x_i^2}]^{-1}.$$

Find the "Median Intrinsic" posterior probability:

Choose the median of P_1 (and P_2) over all choices of x_i, leading to

$$P_1^{med} = 1 - P_2^{med} = [1 + \sqrt{n}\, e^{\frac{n}{2}\bar{x}^2}\, e^{-\frac{1}{2}\mathrm{med}\{x_i^2\}}]^{-1}$$

as the recommended conventional posterior probabilities of M_1 and M_2.

General Algorithm:

- Begin with noninformative priors, π_i^N, for the parameters θ_i in the model M_i ($\pi_i^N(\theta_i) = 1$ is okay).
- Define a "minimal training sample," $\mathbf{x}^* = (x_1^*, \ldots, x_l^*)$, as the smallest subset of the data such that the posterior distributions $\pi_i^N(\theta_i|\mathbf{x}^*)$ are all proper. (Usually, $l = \#$ parameters.)
- Compute the posterior probabilities of all models using the $\pi_i^N(\theta_i|\mathbf{x}^*)$ as priors, for the remaining data.
- Do this for every possible minimal training sample, $\mathbf{x}^*$, and take the median of the results.

Formula:

$$\begin{aligned} P_i &= \text{'intrinsic median' posterior probability of} M_i \\ &= \text{Median} \left\{ \frac{m_i^N(\mathbf{x})}{m_i^N(\mathbf{x}^*) \sum_{j=1}^m (m_j^N(\mathbf{x}) / m_j^N(\mathbf{x}^*))} \right\}, \end{aligned}$$

where $m_i^N(\mathbf{x}) = \int f_i(\mathbf{x}|\theta_i) \pi_i^N(\theta_i) d\theta_i$.

Note: When the number of possible training samples is large, one need only sample from them and take the median posterior probability over those sampled. Indeed, if n is the sample size of the data, it usually suffices to draw just n (sets) of training samples.

Example: Hald regression data

Possible regressors: X_1, X_2, X_3, X_4

Models: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$,

$\epsilon \sim N(0, \sigma^2)$

Models $\rightsquigarrow$ choose some subset of regressors

Notation: Model {1,3,4} means

$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$

Noninformative priors: $\pi(\boldsymbol{\beta}, \sigma) = 1/\sigma$

Minimal training samples: $\mathbf{x}^*$ consists of any subset of six distinct observations (since there are a maximum of six unknown parameters, including σ^2).

Formula for m_i^N:

$$m_i^N(\mathbf{x}) = \frac{\pi^{k_i/2}\Gamma((n-k_i)/2)}{\sqrt{det(X_{(i)}^t X_{(i)})}R_i^{(n-k_i)/2}},$$

where, for model M_i, k_i is the number of regressors plus one, $X_{(i)}$ is the design matrix, and R_i is the residual sum of squares.

Answers:

model	posterior probability
{1,2,3,4}	0.049
{1,2,3}	0.171
{1,2,4}	0.190
{1,3,4}	0.160
{2,3,4}	0.041
{1,2}	0.276
{1,4}	0.108
{3,4}	0.004
others	< 0.0003

Do a slide giving the variable inclusion probabilities, and also do a slide that does the Stewart thing: point mass at zero, giving the probability that a parameter is zero, plus the posterior distribution of the parameter given that it is included. Perhaps refer to utilities, saying that by looking at this graph one can also decide to set a variable to zero because the posterior is so close to zero. Do bivariate inclusion probabilities? (Luis's student?)

Testing when there is no alternative hypothesis

Model under consideration: $H_0 : \mathbf{X} \sim f(\mathbf{x} \mid \theta)$;
observed data is $\mathbf{x}_{obs}$

Available prior (usually noninformative): $\pi(\theta)$

Test statistic: $T = t(\mathbf{X})$, large values being 'surprising' and indicating doubt concerning the model.

Basic question: How should one quantify the surprise in a large value of T?

The difficulty: There is no alternative, H_1, to H_0, so a Bayesian analysis is precluded.

Methods of Measuring Surprise

see (Bayarri and Berger, 1998) for general comparison

- Embed $f(\mathbf{x}|\theta)$ in a nonparametric default alternative H_1, and perform a Bayesian test of H_0 versus H_1.
- Compute a p-value: $p = \mathrm{P}r(t(\mathbf{X}) \geq t(\mathbf{x}_{obs}))$.
  Key issue: what probability measure should be used when θ is unknown?
- Many other surprise measures based on likelihood ratio and other ideas.

  (Weaver, 1948; Good, 1956, 1983, 1988; Berger, 1985; Evans, 1997)
- We base our surprise measure on p-values but recalibrate for easier interpretation.

Proposed Calibration: $\underline{\mathrm{B}}(p) = -ep\log(p)$ for $p < e^{-1}$ and interpret this as the "odds" of $H_0 : \mathbf{X} \sim f(\mathbf{x}|\theta)$, to H_1: f is wrong

Frequentist motivation: It exactly corresponds to the calibration proposed from the conditional frequentist perspective: if both hypothesis are equally likely a priori, then $\underline{\mathrm{B}}(p)$ gives rise to a posterior probability for the null equal to the lower bound on the Type I CEP.

Bayesian motivation: This converts a p-value into a Bayes factor, in particular, into a reasonable *lower bound* on the Bayes factor of H_0 to H_1. (Sellke, Bayarri, Berger, 2001)

Example of a lower bound on the Bayes factor:

Null model: $H_0 : X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0,1)$

Test statistic: $T(\mathbf{X}) = |\bar{X}|$

P-value: $p = \mathsf{P}r(|\bar{X}| > |\bar{x}_{obs}|) = 2(1 - \Phi(\sqrt{n}|\bar{x}_{obs}|)$

Suppose we had the alternative

$$H_1 : X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1), \quad \theta \neq 0$$

Consider the prior on θ under H_1

$$\pi(\theta) \in \Gamma_{US} = \{ \text{ unimodal and symmetric priors } \}$$

Compute $\underline{\mathsf{B}} = \inf_{\pi \in \Gamma_{US}}(\text{ Bayes factor of } H_0 \text{ to } H_1)$

p	0.1	0.05	0.01	0.001
$\underline{\mathsf{B}}$	0.6393	0.4084	0.1223	0.01833
$B(p) = -ep \log p$	0.6259	0.4072	0.1252	0.01878

The calibration is also reasonable in high-dimensional problems.

Predictive p-values

Issue: How to compute

$$p = \mathsf{P}r(t(\mathbf{X}) \geq t(\mathbf{x}_{obs}))$$

when $\mathbf{X} \sim f(\mathbf{x}|\theta)$ and θ is unknown.

Classical Approaches:

1. Replace θ by $\hat{\theta}$
2. Find a statistic $U(\mathbf{X})$ such that $f(T|U,\theta) = f(T|U)$ does not depend on θ
3. Modifications based on asymptotics
4. Posterior predictive p-value (Guttman, 1967, Rubin, 1984)

Partial Posterior Predictive p-values

Definition

$$\pi(\theta\,|\mathbf{x}_{obs}\backslash t_{obs}) \;\propto\; f(\mathbf{x}_{obs}|t_{obs},\theta)\pi(\theta) \propto \frac{f(\mathbf{x}_{obs}|\theta)\pi(\theta)}{f(t_{obs}|\theta)}$$

$$m(t\,|\mathbf{x}_{obs}\backslash t_{obs}) \;=\; \int f(t|\theta)\pi(\theta|\mathbf{x}_{obs}\backslash t_{obs})\,d\theta\,,$$

$$p_{ppp} = \mathsf{P}r^{m(\cdot\,|\mathbf{x}_{obs}\backslash t_{obs})}(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})\,.$$

Nice Features

- Emphasizes model (not prior) checking
- Noninformative priors can be used
- Effectively, there is no double use of the data

- No extra difficulty in discrete problems
- Computations are relatively easy, especially when $f(t|\theta)$ is in closed form

Not so nice features

- $m(t\,|\mathbf{x}_{obs}\backslash t_{obs})$ does not have a clear Bayesian interpretation

Example: $H_0 : X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Ex(\lambda)$

- Order statistics: $X_{(1)} < \ldots, X_{(n)}$; $S = \sum_{i=1}^{n} X_i$
- Discrepancy statistic: $T = X_{(1)}$ (checking left tail)
- Noninformative prior $\pi(\lambda) = 1/\lambda$
- $\pi(\lambda | \mathbf{x}_{obs} \backslash t_{obs})$:

$$\frac{f(\mathbf{x}_{obs}|\lambda)\pi(\lambda)}{f(t_{obs}|\lambda)} = \frac{\lambda^{n-2} e^{-\lambda(s_{obs} - n t_{obs})}}{\Gamma(n-1)(s_{obs} - n t_{obs})^{-(n-1)}}$$

- $m(t\,|\mathbf{x}_{obs} \backslash t_{obs})$:

$$\int f(t|\lambda)\pi(\lambda|\mathbf{x}\backslash t)\, d\lambda = \frac{n(n-1)(s_{obs} - n t_{obs})^{n-1}}{(nt + s_{obs} - n t_{obs})^{n}}$$

- partial predictive p-value: $p_1 = \left(1 - \frac{nt_{obs}}{s_{obs}}\right)^{(n-1)}$
- plug-in p-value: $p_2 = \exp(-n^2\, t_{obs}/s_{obs})$
- posterior predictive p-value: $p_3 = \left(1 + \frac{nt_{obs}}{s_{obs}}\right)^{-n}$
- Note: as $nt_{obs}/s_{obs} \to 1$ ("sure" evidence against H_0), $p_1 \to 0$, $p_2 \to e^{-n}$, $p_3 \to 2^{-n}$.

Computation

$$p_{ppp} = \int \Pr\left(t(\mathbf{X}) \geq t(\mathbf{x}_{obs})|\theta\right) \pi(\theta\,|\mathbf{x}_{obs} \backslash t_{obs})\, d\theta$$

$$\pi(\theta\,|\mathbf{x}_{obs} \backslash t_{obs}) = \frac{f(\mathbf{x}_{obs}|\theta)\pi(\theta)/f(t_{obs}|\theta)}{\int f(\mathbf{x}_{obs}|\theta)\pi(\theta)/f(t_{obs}|\theta)\, d\theta}$$

Straight Monte Carlo:

- Generate $\theta_j \sim \pi(\theta\,|\mathbf{x}_{obs} \backslash t_{obs}),\ j = 1, \ldots, m$, by, say, Metropolis using $\pi(\theta\,|\mathbf{x}_{obs})$ as probing and moving from θ_j to θ^* with probability $\min\{1, f(t_{obs}|\theta_j)/f(t_{obs}|\theta^*)\}$.
- Generate $X_j \sim f(\mathbf{x}_{obs}|\theta_j),\ j = 1, 2, \ldots, m$
- $\hat{p}_{ppp} = \frac{1}{m}$ (number of $t(\mathbf{X}_j) \geq t(\mathbf{x}_{obs})$)

Importance Sampling:

- Generate $\theta_j \sim \pi(\theta \,|\mathbf{x}_{obs})$
- $\hat{p}_{ppp} = \frac{\sum_{i=1}^{m} \Pr(T \geq t_{obs}|\theta_j)/f(t_{obs}|\theta_j)}{\sum_{i=1}^{m} 1/f(t_{obs}|\theta_j)}$

Theorem (Robins, 1998): Under regularity conditions, the partial predictive p-value is (asymptotically) a true frequentist p-value. The plug-in p-value and the posterior predictive p-value are conservative (meaning they can fail to detect a bad model), with the posterior predictive p-value being worst.

Note: The plug-in and the posterior predictive p-values can be arbitrarily bad, even $p(x) \equiv \frac{1}{2}$ in the limit, if T is chosen poorly.

Conclusions

IF

(i) You want to check compatibility of data with a model without (initially) considering alternatives;

(ii) You have a statistic $T(\mathbf{X})$ which you feel would measure this compatibility;

THEN

(i) Compute the partial posterior predictive p-value;

(ii) Compute $\underline{\mathrm{B}}(p) = -e\,p\,\log(p)$ and interpret this as a lower bound on the odds in favor of H_0.