Objective Bayesian Statistical Inference

James O. Berger
Duke University and the
Statistical and Applied Mathematical Sciences Institute

London, UK
July 6-8, 2005

Outline

- Preliminaries
- History of objective Bayesian inference
- Viewpoints and motivations concerning objective Bayesian analysis
- Classes of objective Bayesian approaches.
- A typical example: determining Cepheid distances
- Final comments

A. Preliminaries

Probability of event A: I assume that the concept is a primitive; a measure of the degree of belief (for an individual or a group) that A will occur. This can include almost any definition that satisfies the usual axioms of probability.

Statistical model for data: I assume it is given, up to unknown parameters θ. (While almost never objective, a model is testable.)

Statistical analysis of data from a model: I presume it is to be Bayesian, although other approaches (e.g. likelihood and frequentist approaches) are not irrelevant.

In subjective Bayesian analysis, prior distributions for θ, $\pi(\theta)$, represent beliefs, which change with data via Bayes theorem.

In objective Bayesian analysis, prior distributions represent ‘neutral’ knowledge and the posterior distribution is claimed to give the probability of unknowns arising from just the data.

A Medical Diagnosis Example (with Mossman, 2001)

The Medical Problem:

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

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

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

Suggested Solution: Assign p_i the Jeffreys-rule objective prior

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

(superior to the uniform prior $\pi(p_i) = 1$). By Bayes theorem, the posterior distribution of p_i given the data, x_i, is

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

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

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

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

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

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

B. A Brief History of Objective Bayesian Analysis

The Reverend Thomas Bayes, began the objective Bayesian theory, by solving a particular problem

- Suppose X is Binomial (n,p); an ‘objective’ belief would be that each value of X occurs equally often.
- The only prior distribution on p consistent with this is the uniform distribution.
- Along the way, he codified Bayes theorem.
- Alas, he died before the work was finally published in 1763.

REV. T. BAYES

The real inventor of Objective Bayes was Simon Laplace (also a great mathematician, astronomer and civil servant) who wrote *Théorie Analytique des Probabilité* in 1812

- He virtually always utilized a ‘constant’ prior density (and clearly said why he did so).
- He established the ‘central limit theorem’ showing that, for large amounts of data, the posterior distribution is asymptotically normal (and the prior does not matter).
- He solved very many applications, especially in physical sciences.
- He had numerous methodological developments, e.g., a version of the Fisher exact test.

6. Laplace in his robes as Chancellor of the Senate.

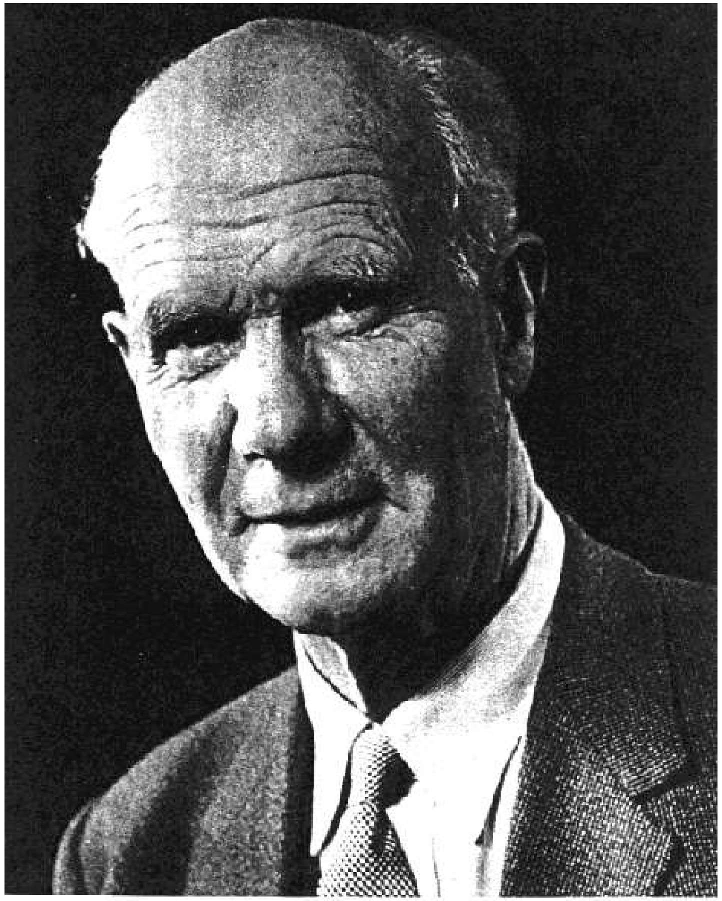
What’s in a name, part I

- It was called *probability theory* until 1838.
- From 1838-1950, it was called *inverse probability,* apparently so named by Augustus de Morgan.
- From 1950 on it was called *Bayesian analysis* (as well as the other names).

Augustus de Morgan

The importance of inverse probability b.f. (before Fisher): as an example, Egon Pearson in 1925 finding the ‘right’ objective prior for a binomial proportion

- Gathered a large number of estimates of proportions p_i from different binomial experiments
- Treated these as arising from the predictive distribution corresponding to a fixed prior.
- Estimated the underlying prior distribution (an early empirical Bayes analysis).
- Recommended something close to the currently recommended ‘Jeffreys prior’ $p^{-1/2}(1-p)^{-1/2}$.



EGON SHARPE PEARSON

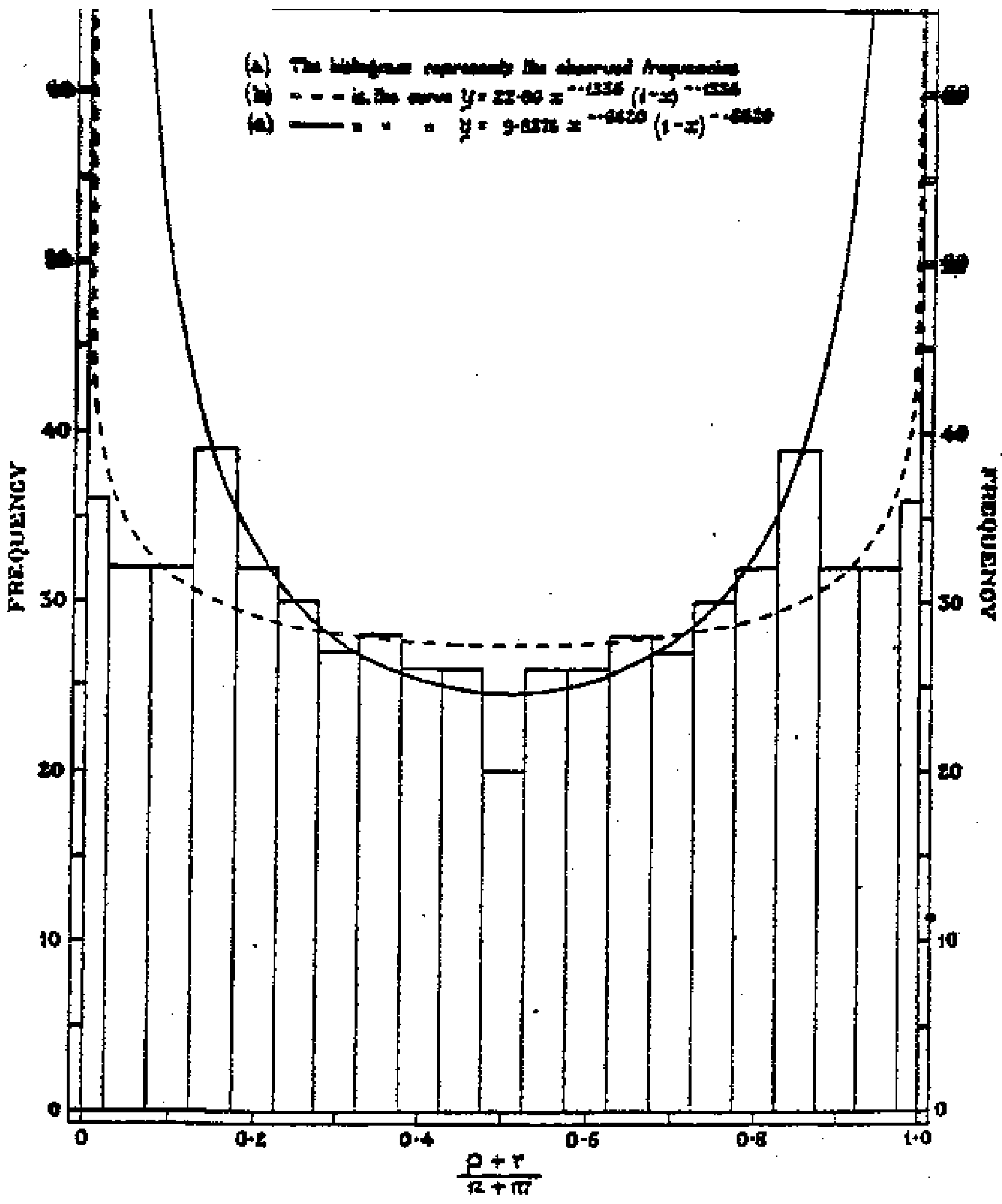
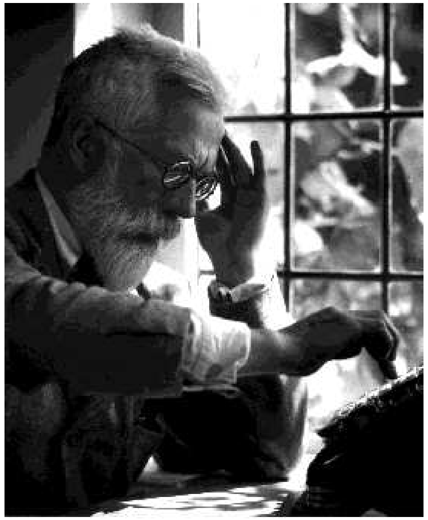


Fig. 3. Distribution of Frequencies of $\frac{p+r}{n+m}$ in 300 samples (made symmetrical).

1930's: 'inverse probability' gets 'replaced' in mainstream statistics by two alternatives

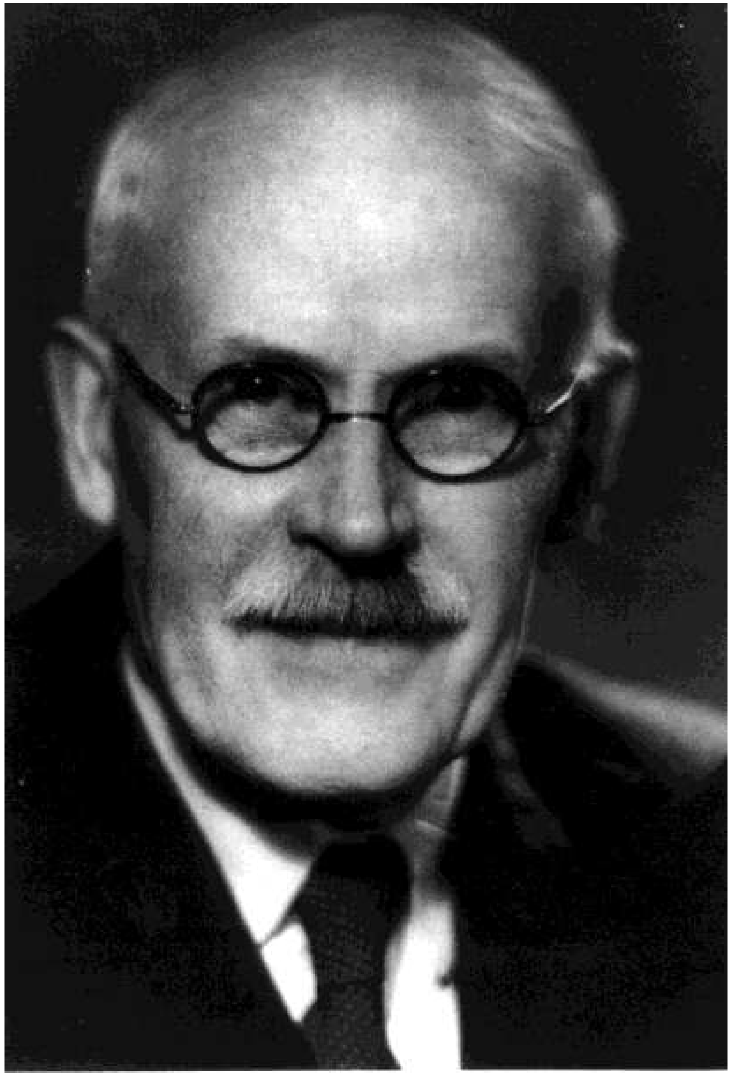
- For 50 years, Boole, Venn and others had been calling use of a constant prior logically unsound (since the answer depended on the choice of the parameter), so alternatives were desired.
- R.A. Fisher's developments of 'likelihood methods,' 'fiducial inference,' … appealed to many.
- Jerzy Neyman's development of the frequentist philosophy appealed to many others.



JERZY NEYMAN

Harold Jeffreys (also a leading geophysicist) revived the Objective Bayesian viewpoint through his work, especially the *Theory of Probability* (1937, 1949, 1963)

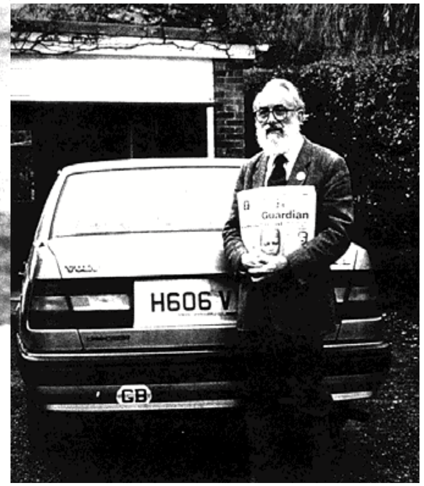
- The now famous *Jeffreys prior* yielded the same answer no matter what parameterization was used.
- His priors yielded the ‘accepted’ procedures in all of the standard statistical situations.
- He began to subject Fisherian and frequentist philosophies to critical examination, including his famous critique of p-values: “An hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred.”



What's in a name, part II

- In the 50's and 60's the *subjective* Bayesian approach was popularized (de Finetti, Rubin, Savage, Lindley, …)
- At the same time, the *objective* Bayesian approach was being revived by Jeffreys, but Bayesianism became incorrectly associated with the subjective viewpoint. Indeed,
  - only a small fraction of Bayesian analyses done today heavily utilize subjective priors;
  - objective Bayesian methodology dominates entire fields of application today.

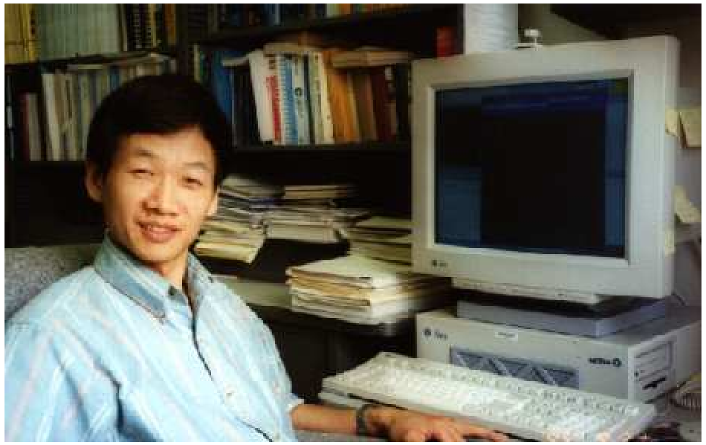


What's in a name, part III

- Some contenders for the name (other than Objective Bayes):
  - Probability
  - Inverse Probability
  - Noninformative Bayes
  - Default Bayes
  - Vague Bayes
  - Matching Bayes
  - Non-subjective Bayes
- But 'objective Bayes' has a website and soon will have *Objective Bayesian Inference*

(coming soon to a bookstore near you)



C. Viewpoints and Motivations Concerning Objective Bayesian Analysis

Four common philosophical positions:

- A major goal of science is to find a completely *coherent* objective Bayesian methodology for learning from data.
- Objective Bayesian analysis is the *best* method for objectively synthesizing and communicating the uncertainties that arise in a problem, but is not coherent according to the usual definitions of coherency.
- Objective Bayesian analysis is a convention we should adopt in scenarios requiring ‘objectivity.’
- Objective Bayesian analysis is simply a collection of adhoc but useful methodologies for learning from data.

More on Coherency

Numerous axiomatic systems seek to define coherent inference or coherent decision making. They all seem to lead to some form of Bayesianism.

- The most common conclusion of the systems is that subjective Bayes is the coherent answer.
  - But it assumes infinitely accurate specification of (typically) an infinite number of things.
- The most convincing coherent system is *robust Bayesian analysis*, where subjective specifications are intervals (e.g. $P(A) \in (0.45, 0.5)$) and conclusions are intervals (see e.g., Walley, Berger, ...)
  - But it has not proven to be practical; the interval answers are typically too wide to be useful.

- Being coherent by itself is worthless: e.g., it is fully coherent to *always* estimate $\theta \in (0, \infty)$ by 17.35426.
- When the goal is communication of information, defining coherency is not easy:
  - *Example:* Suppose we observe data $x \sim N(\theta, 1)$, and the goal is to objectively communicate the probability that $\theta < x$. Almost any objective approach (Bayesian, frequentist, ...) would say the probability is 1/2. This is incoherent under definitions of coherency that involve betting and 'dutch books.'
  - *Example:* Suppose we want confidence intervals for the correlation coefficient in a bivariate normal distribution. There is an objective Bayesian answer that is simultaneously correct from Bayesian, frequentist, and fiducial perspectives, but it is incoherent according to the 'marginalization paradox.'

Motivations for Objective Bayesian Analysis

- The appearance of objectivity is often required.
- Even subjectivists should make extensive use of objective Bayesian analysis.
  - An objective Bayesian analysis can provide a reference for the effect of the prior in a subjective analysis.
  - It is rare that subjective elicitation can be thoroughly done for all unknowns in a problem, so that some utilization of objective priors is almost inevitable.
  - An objective Bayesian analysis can be run initially, to assess if subjective priors are even needed. (Perhaps the data will 'swamp the prior'.)
  - Through study of objective priors, one can obtain insight into possibly bad behavior of standard (e.g., conjugate) subjective priors.

- One still wants the many benefits of Bayesian analysis, even if a subjective analysis cannot be done:
  - Highly complex problems can be handled, via MCMC.
  - Very different information sources can be combined, through hierarchical modeling.
  - Multiple comparisons are automatically accommodated.
  - Bayesian analysis is an automatic 'Ockham's razor,' naturally favoring simpler models that explain the data.
  - Sequential analysis (e.g. clinical trials) is much easier.
- Unification of statistics: objective Bayesian methods
  - have very good frequentist properties;
  - solve the two major problems facing frequentists:
    - * How to properly condition.
    - * What to do for small sample sizes.
- Teaching of statistics is greatly simplified.

D. Classes of Objective Bayesian Approaches

- Information-theoretic approaches
  - Maximum entropy: fine when $\boldsymbol{\Theta}$ is finite, but otherwise
    - it often doesn't apply;
    - for a continuous parameter, the entropy of $\pi(\theta)$, $\text{En}(\pi) \equiv -\int_\Theta \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta$, is only defined relative to a base or reference density π_0.
  - Minimum description length (or minimum message length) priors have more or less the same issues.
  - *Reference priors* (Bernardo): choose $\pi(\theta)$ to minimize the 'asymptotic missing information.'
    - It depends on the inferential goal and on the model, this last making it difficult to achieve traditional coherency.
    - It can be viewed as maximum entropy, together with a way to determine π_0.

* Formally, let $\boldsymbol{x}$ be the data from the model $\mathcal{M} = \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$.
* Let $\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ be the result of k conditionally independent replications of the original experiment, so that $p(\boldsymbol{x}^{(k)} \,|\, \theta) = \prod_{j=1}^{k} p(\boldsymbol{x}_j \,|\, \theta), \quad \boldsymbol{x}_j \in \boldsymbol{\mathcal{X}}, \quad \theta \in \Theta$.
* The amount of information about θ which repeated sampling from $\mathcal{M}$ will provide is the functional of the proper prior $\pi(\theta)$,

$$I[\boldsymbol{\mathcal{X}}^k, \pi\,] \equiv \int_\Theta \int_{\boldsymbol{\mathcal{X}}^k} \pi(\theta) p(\boldsymbol{x}^{(k)} \,|\, \theta) \; \log \left[\frac{\pi(\theta \,|\, \boldsymbol{x}^{(k)})}{\pi(\theta)} \right] d\boldsymbol{x}^{(k)} \, d\theta \,,$$

where $\pi(\theta \,|\, \boldsymbol{x}^{(k)}) \propto p(\boldsymbol{x}^{(k)} \,|\, \theta)\, \pi(\theta)$ is the posterior distbn.
* As $k \to \infty$, perfect knowledge about θ will be approached so that, $I[\boldsymbol{\mathcal{X}}^k, \pi\,]$ will approach the missing information about θ which corresponds to $\pi(\theta)$.
* The reference prior, $\pi^*(\theta \,|\, \mathcal{M})$, is that which maximizes the missing information.

- Invariance/Geometry approaches
  - Jeffreys prior
  - Transformation to local location structure (Box and Tiao, ... recently Fraser et. al. in a data-dependent fashion)
  - *Invariance to group operations*
    * Right-Haar priors and left-Haar (structural) priors (Fraser)
    * Fiducial distributions (Fisher) and specific invariance (Jaynes)
- Frequentist-based approaches
  - *Matching priors*, that yield Bayesian confidence sets that are (at least approximately) frequentist confidence sets.
  - Admissible priors, that yield admissible inferences.
- Objective testing and model selection priors.
- Nonparametric priors.

Matching Priors (Peers; Datta and Mukerjee, 2004)

An objective prior is often evaluated by the frequentist coverage of its credible sets (when interpreted as confidence intervals). If ξ is the parameter of interest (with $\boldsymbol{\theta}$ the entire parameter), it suffices to study one-sided intervals $(-\infty, q_{1-\alpha}(\mathbf{x}))$, where $q_{1-\alpha}(\mathbf{x})$ is the posterior quantile of ξ, defined by

$$P(\xi < q_{1-\alpha}(\mathbf{x}) \mid \mathbf{x}) = \int_{-\infty}^{q_{1-\alpha}(\mathbf{x})} \pi(\xi \mid \mathbf{x})\, d\xi = 1 - \alpha.$$

Of interest is the frequentist coverage of the one-sided intervals

$$C(\boldsymbol{\theta}) = P(q_{1-\alpha}(\mathbf{X}) > \xi \mid \boldsymbol{\theta}).$$

Definition 1 *An objective prior is* exact matching *for a parameter ξ, if it's* $100(1-\alpha)\%$ *one-sided posterior credible sets for ξ have frequentist coverage equal to* $1-\alpha$*. An objective prior is* matching *if this is true asymptotically up to a term of order* $1/n$*.*

Medical Diagnosis Example: Recall that the goal was to find confidence sets for

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

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

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

Invariance Priors

Generalizes 'invariance to parameterization' to other transformations that seem to leave a problem unchanged. There are many illustrations of this in the literature, but the most systematically studied (and most reliable) invariance theory is 'invariance to a group operation.'

An example: location-scale group operation on a normal distribution:

- Suppose $X \sim N(\mu, \sigma)$.
- Then $X^* = aX + b \sim N(\mu^*, \sigma^*)$, where $\mu^* = a\mu + b$ and $\sigma^* = a\sigma$.

Desiderata:

- Final answers should be the same for the two problems.
- Since the X and X^* problems have identical 'structure,' $\pi(\mu, \sigma)$ should have the same form as $\pi(\mu^*, \sigma^*)$.

Mathematical consequence: use an *invariant measure* corresponding to the 'group action' of the problem, the *Haar measure* if unique, and the *right-Haar measure* otherwise (optimal from a frequentist perspective). For the example, $\pi^{RH}(\mu, \sigma) = \sigma^{-1} d\mu d\sigma$ (independence-Jefferys prior).

Admissible Priors

Objective priors can be too diffuse, or can be too concentrated. Some problems this can cause:

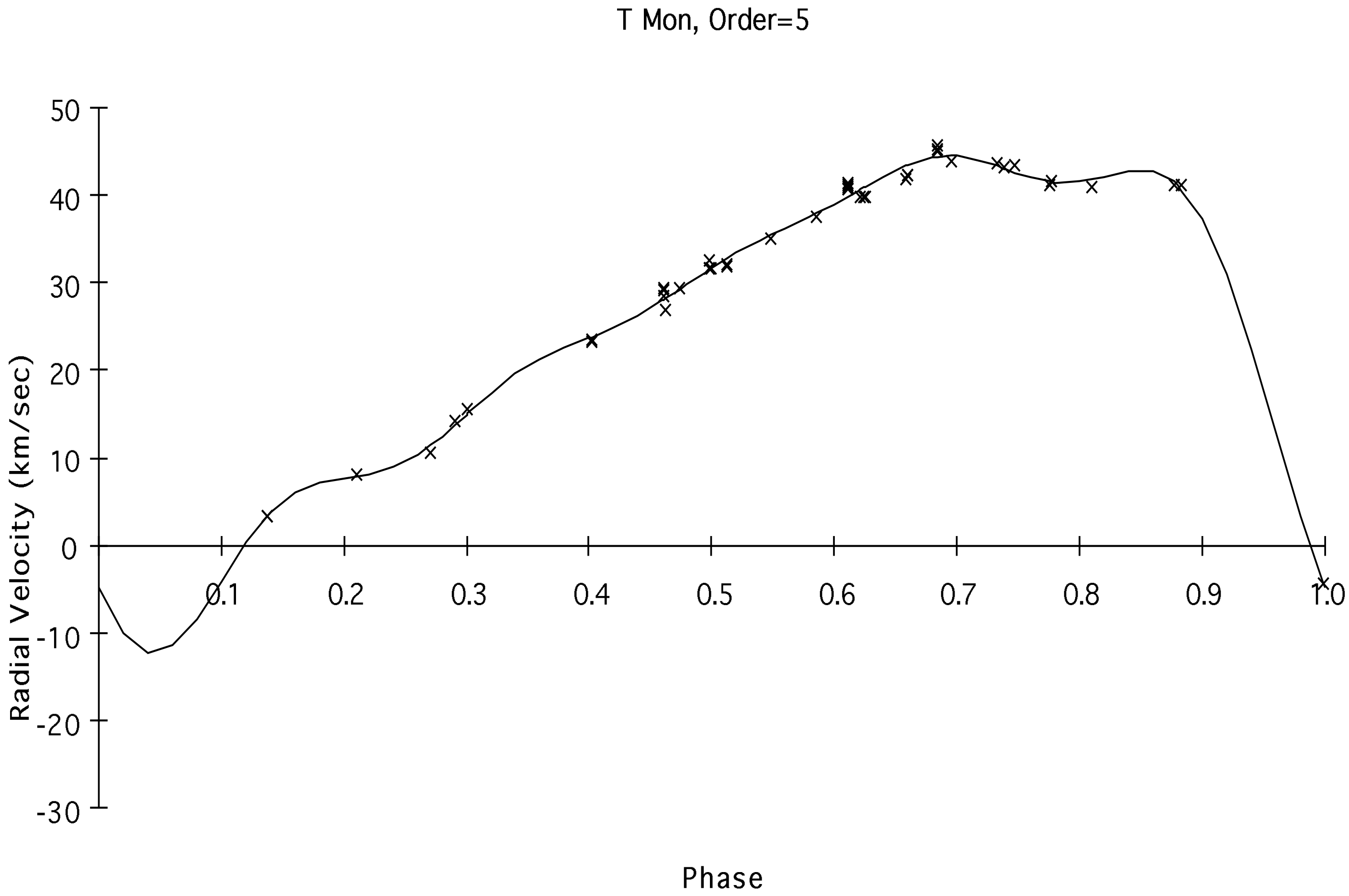
1. **Posterior Impropriety:** If an objective prior does not yield a proper posterior for reasonable sample sizes, it is grossly defective. (One of the very considerable strengths of reference priors – and to an extent Jeffreys priors – is that they almost never result in this problem.)

2. **Inconsistency:** Inconsistent behavior can result as the sample size $n \to \infty$. For instance, in the Neyman-Scott problem of observing $X_{ij} \sim N(\mu_i, \sigma^2), i = 1, \ldots, k; j = 1, 2$, the Jeffrey-rule prior, $\pi(\mu_1, \ldots, \mu_k, \sigma) = \sigma^{-k}$, leads to an inconsistent estimator of σ^2 as $n = 2k \to \infty$; the reference prior is fine.

3. **Priors Overwhelming the Data:** As an example, in large sparse contingency tables, priors will often be much more influential than the data, if great care is not taken.

E. A Typical Application: Determining Cepheid Distances (with Jefferys and Müller)

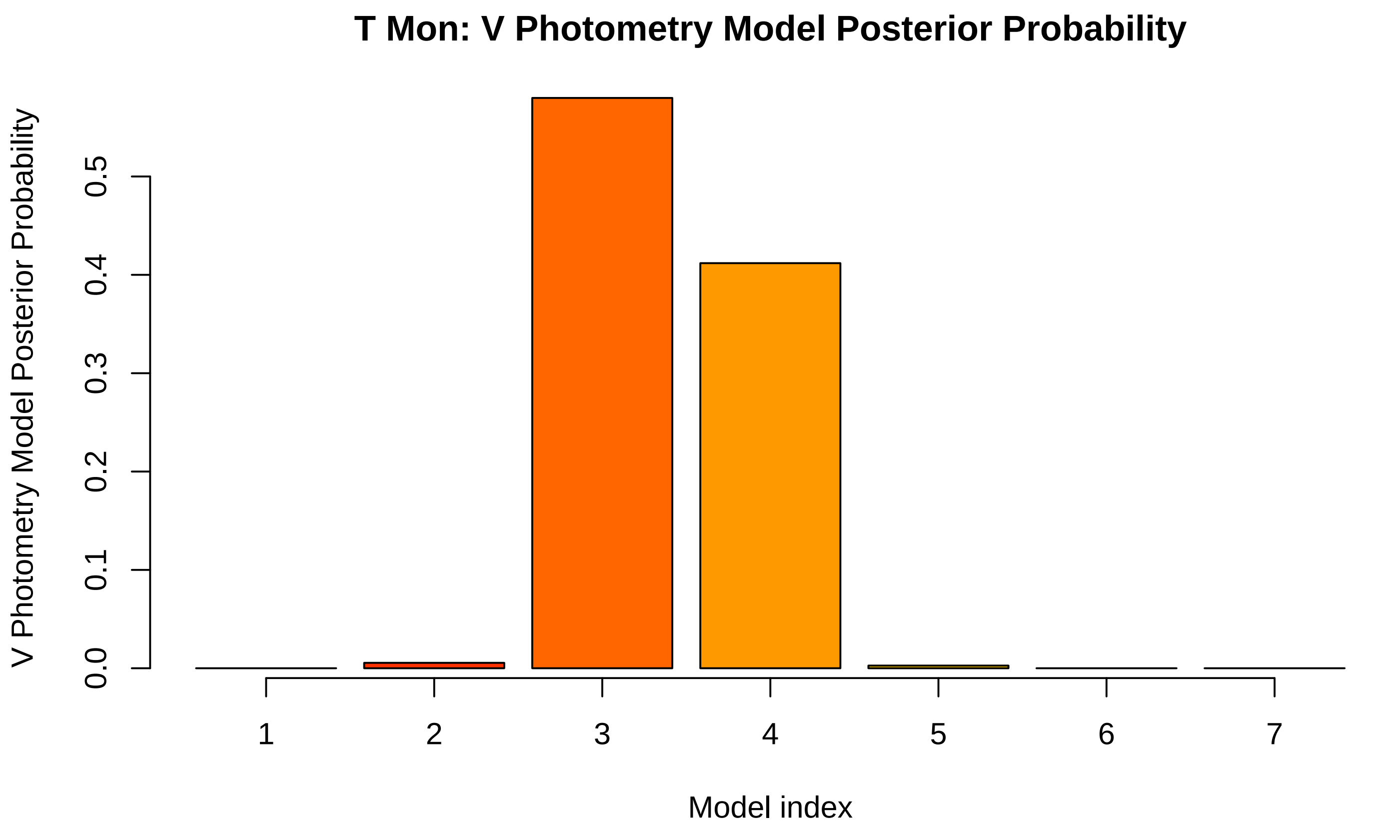
- A Cepheid star pulsates, regularly varying its luminosity (light output) and size.
- From the Doppler shift as the star surface pulsates, one can compute surface velocities at certain phases of the star's period.
- From the luminosity and 'color' of the star, one can learn about the angular radius 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

Choice of Prior Distributions

- The orders, (M, N), of the trigonometric polynomials used to fit the velocity curves, 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 $p(\tau_u) = \frac{1}{\tau_u}$, $p(\tau_v) = \frac{1}{\tau_v}$, and $p(\tau_c) = \frac{1}{\tau_c}$.
- The mean velocity and luminosity, u_0 and v_0, are 'location parameters' and so can be assigned the standard objective priors $p(u_0) = 1$ and $p(v_0) = 1$.
- The angular diameter ϕ_0 and the unknown phase shift $\Delta\phi$ are also assigned the objective priors $p(\Delta\phi) = 1$ and $p(\phi_0) = 1$. It is unclear if these are 'optimal' objective priors but the choice was found to have neglible impact on the answers.

- The Fourier coefficients arising from the curve fitting (which is done by model selection from among trigonmetric polynomials) occur in linear models, so Zellner-Siow *conventional* model selection priors were 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, constrained by subjective knowledge as to the extent of flattening.
  - So, we use $p(s) \propto s^2 \exp(-|s \sin\beta|/z_0)$,
    - 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.


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

Final Comments

- Objective Bayesian analysis has been at the core of statistics and science for nearly 250 years.
- Practical objective Bayesian inference is thriving today, but it still has shallow foundations.
  - Ordinary definitions of coherency are not oriented towards the goal of objective communication.
  - There are a host of logical 'paradoxes' to sort through.
  - Any foundational justification would likely need to acknowledge that 'communication' must be context dependent, including not only the purpose of the communication but the background (e.g., statistical model).
- The best way to deal with any subjective knowledge *seems* to be to view the subjective specifications as constraints, and operate in an objective Bayesian fashion conditional on those constraints.