An Overview of Objective Bayesian Analysis

James O. Berger
Duke University
visiting the University of Chicago

Department of Statistics
Spring Quarter, 2011
Lectures

- Lecture 1. Objective Bayesian Analysis: Introduction and a Casual History
- Lecture 2. Objective Bayesian Estimation
- Lecture 3. Objective Bayesian Hypothesis Testing and Conditional Frequentist Testing
- Lecture 4. Essentials of Objective Bayesian Model Uncertainty
- Lecture 5. Methodology of Objective Bayesian Model Uncertainty
Outline for Lecture 1

- A. Preliminaries and an introductory example
- B. Brief history of objective Bayesian analysis
- C. Adhoc versus formal objective Bayesian analysis
- D. The Jeffreys Rule prior
A. Preliminaries and Introductory Example

**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 (first 2 lectures), up to unknown parameters $\theta$. (While almost never objective, a model is testable.)

**In subjective Bayesian analysis,** prior distributions for $\theta$, $\pi(\theta)$, represent personal beliefs; these could, however, be based on agreed communal knowledge (“evidence-based priors,” “scientific priors,” ...); the posterior distribution then reflects how this prior knowledge is altered by the data.

**In objective Bayesian analysis,** prior distributions represent ‘neutral’ knowledge and the posterior distribution is viewed as giving 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 diagnostic test results in either a Positive (P) or Negative (N) reading.
- $p_1 = \Pr(P \mid \text{patient has } D)$.
- $p_2 = \Pr(P \mid \text{patient does not have } D)$.

It follows from Bayes theorem that

$$
\theta = \Pr(D \mid P) = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2}.
$$
The Statistical Problem: The $p_i$ are unknown. Based on (independent) data $X_i \sim \text{Binomial}(n_i, p_i)$ (arising from medical studies), find a $100(1 - \alpha)\%$ confidence set for $\theta$.

Suggested Solution: Assign $p_i$ the Jeffreys-rule 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

$$
\pi(p_i \mid x_i) = \frac{p_i^{-1/2}(1 - p_i)^{-1/2} \times \left( \begin{array}{c} n_i \\ x_i \end{array} \right) p_i^{x_i}(1 - p_i)^{n_i-x_i}}{\int p_i^{-1/2}(1 - p_i)^{-1/2} \times \left( \begin{array}{c} n_i \\ x_i \end{array} \right) p_i^{x_i}(1 - p_i)^{n_i-x_i} dp_i},
$$

which is the Beta($x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2}$) distribution.
Finally, compute the desired confidence set (formally, the $100(1 - \alpha)\%$ equal-tailed posterior credible set) by

- drawing random $p_i$ from the Beta($x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2}$) distributions, $i = 0, 1, 2$;
- computing the associated $\theta = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2}$;
- repeating this process 10,000 times;
- using the $\frac{\alpha}{2}\%$ upper and lower percentiles of these generated $\theta$ to form the desired confidence limits.
Table 1: The 95\% equal-tailed posterior credible interval for $\theta = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2}$, for various values of the $n_i$ and $x_i$.

<table>
<thead>
<tr>
<th>$n_0 = n_1 = n_2$</th>
<th>$(x_0, x_1, x_2)$</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>(2,18,2)</td>
<td>(0.107, 0.872)</td>
</tr>
<tr>
<td>20</td>
<td>(10,18,0)</td>
<td>(0.857, 1.000)</td>
</tr>
<tr>
<td>80</td>
<td>(20,60,20)</td>
<td>(0.346, 0.658)</td>
</tr>
<tr>
<td>80</td>
<td>(40,72,8)</td>
<td>(0.808, 0.952)</td>
</tr>
</tbody>
</table>
The actual goal of the scientist was to find frequentist confidence intervals for

\[
\theta = Pr(D | P) = \frac{p_0p_1}{p_0p_1 + (1 - p_0)p_2}.
\]

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

<table>
<thead>
<tr>
<th>((p_0, p_1, p_2))</th>
<th>O-Bayes</th>
<th>Log Odds</th>
<th>Gart-Nam</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\frac{1}{4}, \frac{3}{4}, \frac{1}{4}))</td>
<td>2.86, 2.71</td>
<td>1.53, 1.55</td>
<td>2.77, 2.57</td>
<td>2.68, 2.45</td>
</tr>
<tr>
<td>((\frac{1}{10}, \frac{9}{10}, \frac{1}{10}))</td>
<td>2.23, 2.47</td>
<td>0.17, 0.03</td>
<td>1.58, 2.14</td>
<td>0.83, 0.41</td>
</tr>
<tr>
<td>((\frac{1}{2}, \frac{9}{10}, \frac{1}{10}))</td>
<td>2.81, 2.40</td>
<td>0.04, 4.40</td>
<td>2.40, 2.12</td>
<td>1.25, 1.91</td>
</tr>
</tbody>
</table>
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.
The real inventor of Objective Bayes was Simon Laplace (also a great mathematician, astronomer and civil servant) who wrote *Théorie Analytique des Probabilités* 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.
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).
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 \( \frac{\hat{p} - p}{\sigma_p} \) in 300 samples (made symmetric).
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.
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
  – Flat prior Bayes

• Should we even use the word ‘objective,’ when model choice is rarely objective? Yes:
  – Models are testable, at least in principle.
  – Use of models is embedded in the major statistical paradigms, and Bayesians can use the word ‘objective’ if the others do.
C. Adhoc versus Formal O-Bayes Analysis

Objective Bayesian analysis is used all the time in Bayesian practice and has become central to numerous application areas and a number of other disciplines.

But much of it is *ad hoc objective Bayesian analysis*, characterized by use of either

- vague proper priors (dangerous, unless it’s a good approximation to a formal objective prior);

- proper priors chosen over ‘reasonable ranges’ or in a data-dependent fashion (but is one “using the data twice”).

*Ad hoc* objective Bayesian analysis can be successful if validated by experience or extensive sensitivity studies, but this is rarely done.
In contrast, *Formal objective Bayesian analysis* seeks methodology that comes with a ‘guarantee’ of success.

**First effort:** Laplace just used a constant prior but justified it by
- the central limit theorem;
- the recommendation of choosing the parameterization in which beliefs are nearly uniform;
- practical experience that the prior choice was not very relevant.

**Second effort:** Egon Pearson’s effort to empirically determine the objective prior for a binomial parameter.

**Third effort:** The *Jeffreys-rule prior* was developed by Harold Jeffreys to produce objective Bayesian answers that were invariant to the parameterization used for the problem.
D. The Jeffreys Rule Prior

If the data model density is $p(x \mid \theta)$ the Jeffreys-rule prior for the unknown $\theta = \{\theta_1, \ldots, \theta_k\}$ has the form

$$|I(\theta)|^{1/2}d\theta_1 \ldots d\theta_k$$

where $I(\theta)$ is the $k \times k$ matrix Fisher’s information matrix with $(i, j)$ element

$$I(\theta)_{ij} = \mathbb{E}_{x \mid \theta} \left[ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(x \mid \theta) \right].$$
**Example: Binomial Model.**

\( X_i \mid p \sim \text{Bernoulli}(p), \ i = 1, \ldots, n, \) independent.

- Fisher information: \( I(p) = \frac{n}{p(1-p)} \)
- \( \pi(p) \propto \frac{1}{\sqrt{p(1-p)}}. \)

**Example: Normal Model.**

\( X_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2), \ i = 1, \ldots, n, \) independent, \( \Theta = (\mu, \sigma^2) \)

- Fisher Information: \( I(\Theta) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix} \)
- Jeffreys prior: \( \pi(\Theta) \propto (\sigma^2)^{-3/2} \)
- Independent Jeffreys prior: \( \pi(\Theta) \propto 1/\sigma^2 \)
  (ultimately recommended by Jeffreys)
Strengths of the Jeffreys Prior

- Almost always defined

- Transforms properly with 1-1 transformations
  - If $\psi$ is a 1-1 transformation of $\theta$, we need (for answers to be unique)
    \[ \pi(\psi) = \pi(\theta) \det J, \] where \( J_{ij} = \partial \theta_i / \partial \psi_j \).
  - Since \( I^*(\psi) = J I(\theta) J^t \sim \det I^* = \det I \det J^2 \),
    Jeffrey’s prior \( \pi(\psi) \propto [\det I^*(\psi)]^{1/2} \) does so transform.

- Almost always yields a proper posterior
  - Mixture models are the main known examples in which improper posteriors result.

- Great for one-dimensional parameters, in Bayesian and frequentist senses.

- It arises from many other approaches, such as minimum description length and various entropy approaches.
Weaknesses of the Jeffreys Prior

- It depends on the statistical model, and hence appears to violate the likelihood principle.

*Example:* Suppose $X$ is Negative Binomial, i.e.

$$f(x | r, p) = \frac{\Gamma(x + r)}{\Gamma(x + 1)\Gamma(r)} p^r (1 - p)^x, \quad \text{for} \quad x = 0, 1, \ldots$$

The Jeffreys prior is $\pi(p) = 1/[p\sqrt{(1 - p)}]$, which differs from the prior for the binomial model, even though the two models yield proportional likelihood functions.

*Note:* All reasonable objective Bayesian theories similarly depend on the model.

- It requires the Fisher information to exist.

  *Example of non-existence:* Uniform$[0, \theta]$ distribution.

- Often fails badly for higher-dimensional parameters.
Example of Failure – the Neyman-Scott problem:

Suppose we observe

\[ X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \quad i = 1, \ldots, n; \quad j = 1, 2. \]

Defining \( \bar{x}_i = (x_{i1} + x_{i2})/2, \) \( \bar{x} = (\bar{x}_1, \ldots, \bar{x}_n), \) \( S^2 = \sum_{i=1}^{n} (x_{i1} - x_{i2})^2, \) and \( \mu = (\mu_1, \ldots, \mu_n), \) the likelihood function can be written as

\[
L(\mu, \sigma^2) \propto \frac{1}{\sigma^{2n}} \cdot \exp \left[ -\frac{1}{\sigma^2} (|\bar{x} - \mu|^2 + \frac{S^2}{4}) \right].
\]

The Fisher information matrix is

\[
I(\mu, \sigma^2) = \text{diag}\{2/\sigma^2, \ldots, 2/\sigma^2, 2n/\sigma^4\},
\]

the last entry corresponding to the information for \( \sigma^2. \)
• Jeffreys-rule prior: \( \pi^J(\mu, \sigma^2) \propto 1/\sigma^{n+2} \). This is bad; for instance, the marginal posterior for \( \sigma^2 \) is

\[
\pi^J(\sigma^2 | \bar{x}, S^2) = \frac{1}{(\sigma^2)^{(n+1)}} \cdot \exp\left[-\frac{S^2}{4\sigma^2}\right],
\]

which concentrates around \( \sigma^2/2 \).

- Indeed the posterior mean is \( S^2/[4(n - 1)] \), while the frequentist density of \( S^2/[2\sigma^2] \) is Chi-Squared with \( n \) degrees of freedom, so that, as \( n \to \infty \), \( S^2/[4(n - 1)] \to \sigma^2/2 \).

- Thus the posterior distribution of \( \sigma^2 \) is inconsistent, i.e., it does not concentrate around the true value.

(Recall that this example was originally created to show that the MLE can be terrible; here \( \hat{\sigma}^2 = S^2/(4n) \).)
Where things stood in the mid-1960’s, after 200 years of O-Bayes

- The domination of O-Bayes (inverse probability) as the preferred approach to statistical inference has ended (although it is still alive in some fields - e.g. high-energy physics).
- Fisherian and frequentist statistics are dominating statistical practice, although coming under increasing criticism as having their own serious logical flaws.
- Subjective Bayes is suddenly becoming hot, at least philosophically, as the only statistical paradigm apparently free of logical flaws.
- The O-Bayes revival of Jeffreys is well underway, but is stalled because of the problem with the Jeffreys rule prior in multiparameter problems.