Bayesian Inference

Byron Hall
STATISTICAT,LLC

January 13, 2011

This article is an introduction to Bayesian inference for users of the `LaplacesDemon` package in `R`, otherwise referred to as Laplace's Demon. A formal introduction to Laplace's Demon is provided in an accompanying vignette entitled "`LaplacesDemon` Tutorial". Merriam-Webster defines 'Bayesian' as follows:

> **Bayesian** : being, relating to, or involving statistical methods that assign probabilities or distributions to events (as rain tomorrow) or parameters (as a population mean) based on experience or best guesses before experimentation and data collection and that apply Bayes' theorem to revise the probabilities and distributions after obtaining experimental data.

In statistical inference, there are two broad categories of interpretations of probability: Bayesian inference and frequentist inference. These views often differ with each other on the fundamental nature of probability. Frequentist inference loosely defines probability as the limit of an event's relative frequency in a large number of trials, and only in the context of experiments that are random and well-defined. Bayesian inference, on the other hand, is able to assign probabilities to any statement, even when a random process is not involved. In Bayesian inference, probability is a way to represent an individual's degree of belief in a statement, or given evidence.

Within Bayesian inference, there are also different interpretations of probability, and different approaches based on those interpretations. The most popular interpretations and approaches are: objective Bayesian inference (Berger, 2006) and subjective Bayesian inference (Goldstein, 2006). These differences are best explored outside of this article[1].

This article is intended as an approachable introduction to Bayesian inference, or as a handy summary for experienced Bayesians. It is assumed that the reader has at least an elementary understanding of statistics, and this article focuses on applied, rather than theoretical material. Equations and statistical notation are included, but it is hopefully presented so the reader does not need an intricate understanding of solving integrals, for example, but should understand the basic concept of integration. Please be aware that it is difficult to summarize Bayesian inference in such a short article. For a more thorough and formal introduction, see Gelman et al. (2004).

[1] If these terms are new to the reader, then please do not focus too much on the words 'objective' and 'subjective', since there is a lot of debate over them. For what it's worth, *STATISTICAT, LLC* and myself, the author of this `R` package entitled `LaplacesDemon`, favor the 'subjective' interpretation.

The material in this vignette has been organized into the following sections, and is intended to be read in succession, since the material at any given point depends on an understanding of the previous material:

- Bayes' Theorem
  - Example 1
  - Example 2
- Model-Based Bayesian Inference
- Components of Bayesian Inference
- Prior Distributions
  - Informative Prior Distributions
  - Uninformative Prior Distributions
- Hierarchical Bayes
- Conjugacy
- Likelihood
  - Terminology: From Inverse Probability to Bayesian Probability
  - The Likelihood Principle
  - Likelihood Functions of a Parameterized Model
- Numerical Approximation
- Prediction
- Bayes Factors
- Model Fit
- Posterior Predictive Checks
  - Bayesian p-values
  - Conditional Predictive Ordinate
  - Predictive Concordance
- Advantages of Bayesian Inference Over Frequentist Inference
- Advantages of Frequentist Inference Over Bayesian Inference
- References

1 Bayes' Theorem

Bayes' theorem shows the relation between two conditional probabilities that are the reverse of each other. This theorem is named after Reverend Thomas Bayes (1702-1761), and is also referred to as Bayes' law or Bayes' rule. Bayes' theorem expresses the conditional probability, or 'posterior probability', of an event A after B is observed in terms of the 'prior probability' of A, prior probability of B, and the conditional probability of B given A. Bayes' theorem is valid in all common interpretations of probability. The two (related) examples below should be sufficient to introduce Bayes' theorem.

1.1 Bayes' Theorem, Example 1

Bayes' theorem provides an expression for the conditional probability of A given B, which is equal to:

$$\Pr(A|B) = \frac{\Pr(B|A)Pr(A)}{\Pr(B)}$$

For example, suppose one asks the question: what is the probability of going to Hell, conditional on consorting (or given that a person consorts) with Laplace's Demon[2]. By replacing A with $Hell$ and B with $Consort$, the question becomes:

$$\Pr(Hell|Consort) = \frac{\Pr(Consort|Hell)\Pr(Hell)}{\Pr(Consort)}$$

Note that a common fallacy is to assume that $\Pr(A|B) = \Pr(B|A)$, which is called the conditional probability fallacy.

1.2 Bayes' Theorem, Example 2

Another way to state Bayes' theorem is:

$$\frac{\Pr(A_i|B) = \Pr(B|A_i)\Pr(A_i)}{\Pr(B|A_i)\Pr(A_i) + ... + \Pr(B|A_n)\Pr(A_n)}$$

Let's examine our *burning* question, by replacing A_i with $Hell$ or $Heaven$, and replacing B with $Consort$:

- $Pr(A_1) = Pr(Hell)$
- $Pr(A_2) = Pr(Heaven)$
- $Pr(B) = Pr(Consort)$
- $Pr(A_1|B) = Pr(Hell|Consort)$

[2]This example is, of course, intended with humor.

- $Pr(A_2|B) = Pr(Heaven|Consort)$
- $Pr(B|A_1) = Pr(Consort|Hell)$
- $Pr(B|A_2) = Pr(Consort|Heaven)$

Laplace's Demon was conjured and asked for some data. He was glad to oblige:

Data

- 6 people consorted out of 9 who went to Hell.
- 5 people consorted out of 7 who went to Heaven.
- 75% of the population goes to Hell.
- 25% of the population goes to Heaven.

Now, Bayes' theorem is applied to the data. Four pieces are worked out as follows:

- $Pr(Consort|Hell) = 6/9 = 0.666$
- $Pr(Consort|Heaven) = 5/7 = 0.714$
- $Pr(Hell) = 0.75$
- $Pr(Heaven) = 0.25$

Finally, the desired conditional probability $\Pr(Hell|Consort)$ is calculated:

- $Pr(Hell|Consort)$ = (0.666 x 0.75) / (0.666 x 0.75 + 0.714 x 0.25)
- $Pr(Hell|Consort) = 0.737$

The probability of someone consorting with Laplace's Demon and going to Hell is 73.7%, which is less than the prevalence of 75% in the population. According to these findings, consorting with Laplace's Demon does not increase the probability of going to Hell. With that in mind, please continue...

2 Model-Based Bayesian Inference

The basis for Bayesian inference is derived from Bayes' theorem:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

Replacing B with observations y, A with parameter set θ, and probabilities Pr with densities p (or sometimes function f), results in the following:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

where $p(y)$ will be discussed below, p(θ) is the set of prior distributions of parameter set θ before y is observed, $p(y|\theta)$ is the likelihood of y under a model, and $p(\theta|y)$ is the joint posterior distribution, sometimes called the full posterior distribution, of parameter set θ that expresses uncertainty about parameter set θ after taking both the prior and data into account. Since there are usually multiple parameters, θ represents a set of j parameters, and may be considered hereafter in this article as:

$$\theta = \theta_1, ..., \theta_j$$

The denominator:

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

defines the "marginal likelihood" of y, or the "prior predictive distribution" of y, and may be set to an unknown constant c. The prior predictive distribution indicates what y should look like, given the model, before y has been observed. Only the set of prior probabilities and the model's likelihood function are used for the marginal likelihood of y. The presence of the marginal likelihood of y ensures that the joint posterior distribution, $p(\theta|y)$, is a proper distribution and integrates to one.

By replacing $p(y)$ with c, which is short for a 'constant of proportionality', the model-based formulation of Bayes' theorem becomes:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{c}$$

By removing c from the equation, the relationship changes from 'equals' (=) to 'proportional to' ($\propto$)[3]:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

This form can be stated as the joint posterior being proportional to the likelihood times the prior. However, the goal in model-based Bayesian inference is usually not to summarize the joint posterior distribution, but to summarize the marginal distributions of the parameters. The full parameter set θ can typically be partitioned into

$$\theta = \{\phi, \lambda\}$$

where ϕ is the sub-vector of interest, and λ is the complementary sub-vector of θ, often referred to as a vector of nuisance parameters. In a Bayesian framework, the presence of

[3]For those unfamiliar with $\propto$, this symbol simply means that two quantities are proportional if they vary in such a way that one is a constant multiplier of the other. This is due to the constant of proportionality c in the equation. Here, this can be treated as 'equal to'.

nuisance parameters does not pose any formal, theoretical problems. A nuisance parameter is a parameter that exists in the joint posterior distribution of a model, though it is not a parameter of interest. The marginal posterior distribution of ϕ, the parameter of interest, can simply be written as:

$$p(\phi|y) = \int p(\phi, \lambda|y)d\lambda$$

In model-based Bayesian inference, Bayes' theorem is used to estimate the joint posterior distribution, and finally the user can assess the marginal posterior distributions.

3 Components of Bayesian Inference

The components of Bayesian inference are:

1. $p(\theta)$ is the set of prior distributions for parameter set θ, and uses probability as a means of quantifying uncertainty about θ before taking the data into account.
2. $p(y|\theta)$ is the likelihood or likelihood function, in which all variables are related in a full probability model.
3. $p(\theta|y)$ is the joint posterior distribution that expresses uncertainty about parameter set θ after taking both the prior and the data into account. If parameter set θ is partitioned into a parameter of interest ϕ and the remaining parameters are considered nuisance parameters, then $p(\phi|y)$ is the marginal posterior distribution.

4 Prior Distributions

In Bayesian inference, a prior probability distribution, often called simply the prior, of an uncertain parameter θ or latent variable is a probability distribution that expresses uncertainty about θ before the data are taken into account[4]. The parameters of a prior distribution are called hyperparameters, to distinguish them from the parameters (θ) of the model.

When applying Bayes' theorem, the prior is multiplied by the likelihood function and then normalized to estimate the posterior probability distribution, which is the conditional distribution of θ given the data.

[4]One so-called version of Bayesian inference is 'empirical Bayes', which sounds enticing because anything 'empirical' seems desirable. However, empirical Bayes is a term for the use of data-dependent priors, where the prior is first modeled usually with maximum likelihood and then used in the Bayesian model. This is an undesirable double-use of the data and is most problematic with small sample sizes (Berger, 2006). It also seems to violate the elementary concept that a prior probability distribution expresses uncertainty about θ *before* the data are taken into account. It has been claimed that "empirical Bayes methods are not Bayesian" (Bernardo, 2008)

4.1 Informative Prior Distributions

When prior information is available about θ, it should be included in the prior distribution of θ. For example, if the present model form is similar to a previous model form, and the present model is intended to be an updated version based on more current data, then the posterior distribution of θ from the previous model may be used as the prior distribution of θ for the present model.

In this way, each version of a model is not starting from scratch, based only on the present data, but the cumulative effects of all data, past and present, can be taken into account. If the present data is very similar to the previous data, then the precision of the posterior distribution increases when including more and more information from previous models. If the present data differs considerably, then the posterior distribution of θ may be in the tails of the prior distribution for θ, so the prior distribution contributes less density in its tails. Moreover, the prior distribution affects the posterior distribution.

4.2 Uninformative Prior Distributions

When prior information is unavailable about θ, which is more common, an uninformative prior distribution could be used, or the prior distribution could in turn be estimated from hyperprior distributions in a hierarchical context.

Uninformative prior distributions (also called diffuse, minimal, non-informative, objective, reference, uniform, or vague priors[5]) attempt to minimize the impact of the selection of the prior distribution. The rationale for using uninformative prior distributions is often said to be 'to let the data speak for themselves'. The epitome of uninformative prior distributions is the unbounded, uniform distribution, often called a flat prior, such as:

$$\theta \sim U(-\infty, \infty)$$

where θ is uniformly-distributed from negative infinity to positive infinity. Although this allows the posterior distribution to be affected soley by the data with no impact from prior information, this should generally be avoided because the posterior distribution is improper, meaning it will not integrate to one, because the integral of the assumed $p(\theta)$ is infinity, which violates the assumption that the probabilities sum to one.

Reverend Thomas Bayes (1702-1761) was the first to use inverse probability, and used a flat prior for his billiard example so that all possible values of θ are equally likely *a priori* (Gelman, 2004, p. 34-36). Pierre-Simon Laplace (1749-1827) also used the flat prior to estimate the proportion of female births in a population. Laplace's use of this prior distribution was later referred to as the 'principle of insufficient reason', and is now called the flat prior (Gelman, 2004, p. 39).

There is often at least some information about θ to include in a prior distribution of θ,

[5]These terms are not all equivalent. For example, Bernardo (2000) proposes 'reference priors', which are better explored elsewhere. The idea is to maximize the expected Kullback-Leibler divergence of the posterior distribution relative to the prior. This maximizes the expected posterior information about y when the prior density is $p(y)$. In some sense, $p(y)$ is the 'least informative' prior about y. Reference priors are often the objective prior of choice in multivariate problems, since other rules (e.g., Jeffreys' rule) may result in priors with problematic behavior.

such as θ must be positive, or θ must be less than some limit. It is popular, for good reasons, to center and scale all continuous predictors (Gelman, 2008). Although centering and scaling predictors is not discussed here, it should be obvious that the potential range of the posterior distribution of θ for a centered and scaled predictor should be small. A popular, uninformative prior distribution for a centered and scaled predictor may be:

$$\theta \sim N(0, 10000)$$

where θ is normally-distributed according to a mean of 0 and a variance of 10,000, which is equivalent to a standard deviation of 100, or precision of 1.0E-4. In this case, the density for θ is nearly flat. Nonetheless, the fact that it is not perfectly flat yields good properties for numerical approximation algorithms. In both Bayesian and frequentist inference, it is possible for numerical approximation algorithms to become stuck in regions of flat density, which become more common as sample size decreases or model complexity increases. Numerical approximation algorithms in frequentist inference function as though a completely flat, uninformative prior were used, so numerical approximation algorithms in frequentist inference become stuck more frequently than numerical approximation algorithms in Bayesian inference. Prior distributions that are not completely flat provide enough information for the numerical approximation algorithm to continue to explore the target density, the posterior distribution.

5 Hierarchical Bayes

Prior distributions may be estimated within the model via hyperprior distributions, which are usually uninformative and nearly flat. Using hyperprior distributions to estimate prior distributions is known as hierarchical Bayes. In theory, this process could continue further, using hyper-hyperprior distributions to estimate the hyperprior distributions. Recall that the posterior distribution is proportional to the likelihood times the prior distribution:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

The simplest hierarchical Bayes model takes the form:

$$p(\theta, \phi|y) \propto p(y|\theta)p(\theta|\phi)p(\phi)$$

where ϕ is a set of hyperprior distributions. By reading the equation from right to left, it begins with hyperpriors ϕ, which are used conditionally to estimate $p(\theta|\phi)$, which in turn is used, as per usual, to estimate the likelihood $p(y|\theta)$, and finally the posterior is $p(\theta, \phi|y)$.

6 Conjugacy

When the posterior distribution $p(\theta|y)$ is in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood. For example, the Gaussian family is conjugate

to itself (or self-conjugate) with respect to a Gaussian likelihood function: if the likelihood function is Gaussian, then choosing a Gaussian prior for the mean will ensure that the posterior distribution is also Gaussian. All probability distributions in the exponential family have conjugate priors. See Gelman (2004) for a catalog.

Although the gamma distribution is the conjugate prior distribution for the precision of a normal distribution, better properties for hierarchical Bayes are yielded with the non-congjuate, proper, half-Cauchy distribution (Gelman, 2006), such as:

$$\sigma \sim U(0, 100)$$
$$\tau = \sigma^{-2}$$

This is noteworthy because most examples of Bayesian inference use the older form:

$$\tau \sim \Gamma(0.001, 0.001)$$

Conjugacy is mathematically convenient in that the posterior distribution follows a known parametric form (Gelman, 2004, p. 40). It is obviously easier to summarize a normal distribution than a complex, multi-modal distribution with no known form. If information is available that contradicts a conjugate parametric family, then it may be necessary to use a more realistic, inconvenient, prior distribution.

The basic justification for the use of conjugate prior distributions is similar to that for using standard models (such as the binomial and normal) for the likelihood: it is easy to understand the results, which can often be put in analytic form, they are often a good approximation, and they simplify computations. Also, they are useful as building blocks for more complicated models, including many dimensions, where conjugacy is typically impossible. For these reasons, conjugate models can be good starting points (Gelman, 2004, p. 41).

Although they can make interpretations of posterior inferences less transparent and computation more difficult, nonconjugate prior distributions do not pose any new conceptual problems. In practice, for complicated models, conjugate prior distributions may not even be possible (Gelman, 2004, p. 41-42).

When conjugate distributions are used, a summary statistic for a posterior distribution of θ may be represented as $t(y)$ and said to be a sufficient statistic (Gelman, 2004, p.42). When nonconjugate distributions are used, a summary statistic for a posterior distribution is usually not a sufficient statistic. A sufficient statistic is a statistic that has the property of sufficiency with respect to a statistical model and the associated unknown parameter. The quantity $t(y)$ is said to be a sufficient statistic for θ, because the likelihood for θ depends on the data y only through the value of $t(y)$. Sufficient statistics are useful in algebraic manipulations of likelihoods and posterior distributions.

7 Likelihood

In order to complete the definition of a Bayesian model, both the prior distributions and the likelihood must be fully specified.

The likelihood, likelihood function, or $p(y|\theta)$, contains the available information provided by the sample. The likelihood is:

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

The data y affects the posterior distribution $p(\theta|y)$ only through the likelihood function $p(y|\theta)$. In this way, Bayesian inference obeys the likelihood principle, which states that for a given sample of data, any two probability models $p(y|\theta)$ that have the same likelihood function yield the same inference for θ. The likelihood principle is covered in more detail later in a section entitled "The Likelihood Principle".

7.1 Terminology: From Inverse Probability to Bayesian Probability

A gambler's dispute in 1654 led to the creation of a mathematical theory of probability by two famous French mathematicians, Blaise Pascal and Pierre de Fermat.

Reverend Thomas Bayes (1702-1761) discovered Bayes' theorem, published posthumously in 1763, in which he was the first to use inverse probability. 'Inverse probability' refers to assigning a probability distribution to an unobserved variable, and is in essence, probability in the opposite direction of the usual sense. For example, the probability of obtaining heads on the next coin flip in a Bayesian context would be the predicted probability, $p(y^{new}|y, \theta)$, but to estimate this predicted probability, the probability distribution of θ must first be estimated, using coin toss data y to estimate the parameter θ by the likelihood function $p(y|\theta)$, which contains the likelihood $p(\theta|y)$, where θ is estimated from the data, y. Therefore, the data, y, is used to estimate the most probable θ that would lead to a data-generating process for y.

Unaware of Bayes, Pierre-Simon Laplace (1749-1827) independently developed Bayes' theorem and first published his version in 1774, eleven years after Bayes, in one of Laplace's first major works (Laplace, 1774, p. 366-367).

In 1812, Laplace (1749-1827) introduced a host of new ideas and mathematical techniques in his book, Theorie Analytique des Probabilites. Before Laplace, probability theory was solely concerned with developing a mathematical analysis of games of chance. Laplace applied probabilistic ideas to many scientific and practical problems.

Then, in 1814, Laplace published his "Essai philosophique sur les probabilites", which introduced a mathematical system of inductive reasoning based on probability. In it, the Bayesian interpretation of probability was developed independently by Laplace, much more thoroughly than Bayes, so some "Bayesians" refer to Bayesian inference as Laplacian inference.

The term "inverse probability" appears in an 1837 paper of Augustus De Morgan, in reference to Laplace's method of probability (Laplace, 1774, 1812), though the term "inverse probability" does not occur in these works. Bayes' theorem has been referred to as "the principle of inverse probability".

Terminology has changed, so that today, Bayesian probability (rather than inverse probability) refers to assigning a probability distribution to an unobservable variable. The "dis-

tribution" of an unobserved variable given data is the likelihood function (which is not a distribution), and the distribution of an unobserved variable, given both data and a prior distribution, is the posterior distribution.

The term "Bayesian", which displaced "inverse probability", was in fact introduced by R. A. Fisher as a derogatory term.

In modern terms, given a probability distribution $p(y|\theta)$ for an observable quantity y conditional on an unobserved variable θ, the "inverse probability" is the posterior distribution $p(\theta|y)$, which depends both on the likelihood function (the inversion of the probability distribution) and a prior distribution. The distribution $p(y|\theta)$ itself is called the direct probability.

However, $p(y|\theta)$ is also called the likelihood function, which can be confusing, seeming to pit the definitions of probability and likelihood against each other. A quick introduction to the likelihood principle follows, and finally all of the information on likelihood comes together in the section entitled "Likelihood Function of a Parameterized Model".

7.2 The Likelihood Principle

An informal summary of the likelihood principle may be that inferences from data to hypotheses should depend on how likely the actual data are under competing hypotheses, not on how likely imaginary data would have been under a single "null" hypothesis or any other properties of merely possible data.

A more precise interpretation may be that inference procedures which make inferences about simple hypotheses should not be justified by appealing to probabilities assigned to observations that have not occurred.

The usual interpretation is that any two probability models with the same likelihood function yield the same inference for θ.

Some authors mistakenly claim that frequentist inference, such as using MLE, obeys the likelihood, though it does not. Some authors claim that the largest contention between Bayesians and frequentists regards prior probability distributions. Other authors argue that, although the subject of priors gets more attention, the true contention between frequentist and Bayesian inference is the likelihood principle, which Bayesian inference obeys, and frequentist inference does not.

There have been many frequentist attacks on the likelihood principle, and have been shown to be poor arguments. Some Bayesians have argued that Bayesian inference is incompatible with the likelihood principle on the grounds that there is no such thing as an isolated likelihood function (Bayarri and DeGroot, 1987). They argue that in a Bayesian analysis there is no principled distinction between the likelihood function and the prior probability function. The objection is motivated, for Bayesians, by the fact that prior probabilities are needed in order to apply what seems like the likelihood principle. Once it is admitted that there is a universal necessity to use prior probabilities, there is no longer a need to separate the likelihood function from the prior. Thus, the likelihood principle is accepted 'conditional' on the assumption that a likelihood function has been specified, but it is denied that specifying a likelihood function is necessary. Nonetheless, the likelihood principle is seen as a useful Bayesian weapon to combat frequentism.

Following are some interesting qutoes from prominent statisticians:

"Using Bayes' rule with a chosen probability model means that the data y affect posterior inference 'only' through the function $p(y|\theta)$, which, when regarded as a function of θ, for fixed y, is called the 'likelihood function'. In this way Bayesian inference obeys what is sometimes called the 'likelihood principle', which states that for a given sample of data, any two probability models $p(y|\theta)$ that have the same likelihood function yield the same inference for θ" (Gelman, 2004, p. 9).

"The likelihood principle is reasonable, but only within the framework of the model or family of models adopted for a particular analysis" (Gelman, 2004, p. 9).

Frequentist "procedures typically violate the likelihood principle, since long-run behavior under hypothetical repetitions depends on the entire distribution $p(y|\theta), y \in Y$ and not only on the likelihood" (Bernardo and Smith, 2000, p. 454).

There is "a general fact about the mechanism of parametric Bayesian inference which is trivially obvious; namely 'for any specified $p(\theta)$, if the likelihood functions $p_1(y_1|\theta), p_2(y_2|\theta)$ are proportional as functions of θ, the resulting posterior densities for θ are identical'. It turns out...that many non-Bayesian inference procedures do not lead to identical inferences when applied to such proportional likelihoods. The assertion that they 'should', the so-called 'Likelihood Principle', is therefore a controversial issue among statisticians. In contrast, in the Bayesian inference context...this is a straightforward consequence of Bayes' theorem, rather than an imposed 'principle' " (Bernardo and Smith, 2000, p. 249).

"Although the likelihood principle is implicit in Bayesian statistics, it was developed as a separate principle by Barnard (1949), and became a focus of interest when Birnbaum (1962) showed that it followed from the widely accepted sufficiency and conditionality principles" (Bernardo and Smith, 2000, p. 250).

"The likelihood principle, by itself, is not sufficient to build a method of inference but should be regarded as a minimum requirement of any viable form of inference. This is a controversial point of view for anyone familiar with modern econometrics literature. Much of this literature is devoted to methods that do not obey the likelihood principle..." (Rossi, Allenby, and McCulloch, 2005, p. 15).

"Adherence to the likelihood principle means that inferences are 'conditional' on the observed data as the likelihood function is parameterized by the data. This is worth contrasting to any sampling-based approach to inference. In the sampling literature, inference is conducted by examining the sampling distribution of some estimator of θ, $\hat{\theta} = f(y)$. Some sort of sampling experiment results in a distribution of y and therefore, the estimator is viewed as a random variable. The sampling distribution of the estimator summarizes the properties of the estimator 'prior' to observing the data. As such, it is irrelevant to making inferences given the data we actually observe. For any finite sample, this dinstinction is extremely important. One must conclude that, given our goal for inference, sampling distributions are simply not useful" (Rossi, Allenby, and McCulloch, 2005, p. 15).

7.3 Likelihood Function of a Parameterized Model

In non-technical parlance, "likelihood" is usually a synonym for "probability", but in statistical usage there is a clear distinction: whereas "probability" allows us to predict unknown outcomes based on known parameters, "likelihood" allows us to estimate unknown parameters based on known outcomes.

In a sense, likelihood can be thought a reversed version of conditional probability. Reasoning forward from a given parameter theta, the conditional probability of y is the density $p(y|\theta)$.

With θ as a parameter, here are relationships in expressions of the likelihood function:

$$\mathscr{L}(\theta|y) = p(y|\theta) = f(y|\theta)$$

where y is the observed outcome of an experiment, and the likelihood ($\mathscr{L}$) of θ given y is equal to the density $p(y|\theta)$ or function $f(y|\theta)$. When viewed as a function of y with θ fixed, it is not a likelihood function $\mathscr{L}(\theta|y)$, but merely a probability density function $p(y|\theta)$. When viewed as a function of θ with y fixed, it is a likelihood function and may be denoted as $\mathscr{L}(\theta|y)$, $p(y|\theta)$, or $f(y|\theta)$[6].

For example, in a Bayesian linear regression with an intercept and two independent variables, the model may be specified as:

$$y_i \sim N(\mu_i, \tau^{-1})$$
$$\mu_i = \beta_1 + \beta_2 X_{i,1} + \beta_3 X_{i,2}$$

The dependent variable y, indexed by $i = 1, ..., n$, is stochastic, and normally-distributed according to the expectation vector μ, and the inverse of the residual precision τ, where the scalar τ is the inverse of the variance. Expectation vector μ is an additive, linear function of a vector of regression parameters, β, and the design matrix $\mathbf{X}$.

Since y is normally-distributed, the probability density function (PDF) of a normal distribution will be used, and is usually denoted as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp[(-\frac{1}{2}\sigma^2)(y_i - \mu_i)^2]; -\infty < y < \infty$$

By replacing the variance, σ^2, with its inverse, τ, and by considering a conditional distribution, the equivalent record-level likelihood in Bayesian notation is:

$$p(y_i|\theta) = \sqrt{\frac{\tau}{2\pi}} \exp[(-\frac{\tau}{2})(y_i - \mu_i)^2]; -\infty < y < \infty$$

In both theory and practice, and in both frequentist and Bayesian inference, the log-likelihood is used instead of the likelihood, on both the record- and model-level. The model-level product of record-level likelihoods can exceed the range of a number that can be stored by a computer, which is usually affected by sample size. By estimating a record-level log-likelihood, rather than likelihood, the model-level log-likelihood is the sum of the record-level log-likelihoods, rather than a product of the record-level likelihoods.

$$\log[p(y|\theta)] = \sum_{i=1}^{n} \log[p(y_i|\theta)]$$

[6]Note that $\mathscr{L}(\theta|y)$ is not the same as the probability that those parameters are the right ones, given the observed sample.

rather than

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

As a function of θ, the joint posterior distribution is the product of the likelihood function and the prior distributions. To continue with the example of Bayesian linear regression, here is the joint posterior distribution:

$$p(\theta|y) = p(y|\theta)p(\beta_1)p(\beta_2)p(\beta_3)p(\tau)$$

More usually, the log of the joint posterior distribution is used, which is the sum of the log-likelihood and prior distributions. Here is the log of the joint posterior distribution for this example:

$$\log[p(\theta|y)] = \log[p(y|\theta)] + \log[p(\beta_1)] + \log[p(\beta_2)] + \log[p(\beta_3)] + \log[p(\tau)]$$

The log of the joint posterior distribution is maximized with numerical approximation.

8 Numerical Approximation

The technical problem of evaluating quantities required for Bayesian inference typically reduces to the calculation of a ratio of two integrals (Bernardo and Smith, 2000, p. 339). In all cases, the technical key to the implementation of the formal solution given by Bayes' theorem is the ability to perform a number of integrations (Bernardo and Smith, 2000, p. 340). Except in certain rather stylized problems, the required integrations will not be feasible analytically and, thus, efficient approximation strategies are required.

"One approach would be to take various asymptotic approximations to these integrals (such as the Laplace Approximation). Unless these asymptotic approximations can be shown to be accurate, we should be very cautious about using them. In contrast, much of the econometrics and statistics literature uses asymptotic approximations to the sampling distributions of estimators and test statistics without investigating accuracy...Fortunatley, we do not have to rely on asymptotic approximations in modern Bayesian inference" (Rossi, Allenby, and McCulloch, 2005, p. 17).

Numerical approximation algorithms are used to maximize the joint posterior distribution of a model, given the likelihood function and prior distributions. There are too many different types of numerical approximation algorithms in Bayesian inference to cover in any detail in this article. An incomplete list of broad categories of Bayesian numerical approximation may include Approximate Bayesian Computation (ABC), Iterative Quadrature, Laplace Approximation, Markov chain Monte Carlo (MCMC), and Variational Bayes (VB).

8.1 Laplace Approximation

Also known as Laplace's Method, there are a variety of Laplace Approximation algorithms that attempt to estimate distributions, usually Gaussian, for posterior moments. Laplace Approximations are second in popularity only to MCMC. These algorithms are being developed and refined. These approximations are deterministic and asymptotic, and are subject to most of the same limitations as Maximum Likelihood Estimation (MLE) in frequentist inference. The advantage to using Laplace Approximation is that estimation is nearly as fast as with MLE, and it is possible to work with large data sets. The Laplace Approximation algorithm is detailed elsewhere.

8.2 Markov chain Monte Carlo

The most common method of numerical approximation for Bayesian inference is to use a Markov chain Monte Carlo (MCMC) algorithm, of which there are many. In a Bayesian context, the most common MCMC algorithms are Gibbs sampling and Metropolis-Hastings. MCMC algorithms are stochastic and are not asymptotic. Currently, the `R` package entitled `LaplacesDemon` offers four MCMC algorithms. For more information, see the vignette entitled "`LaplacesDemon` Tutorial".

8.3 Recommendations

If sample size is sufficiently large and the model is not extremely complicated, then Laplace Approximation may be the most practical method of numerical approximation, allowing fast convergence and reasonable approximations with large data sets. If sample size is small or the model is extremely complicated, then MCMC is usually the most practical method of numerical approximation, and allows exact estimation with respect to sample size.

9 Prediction

The "posterior predictive distribution" is either the replication of y given the model (usually represented as y^{rep}), or the prediction of a new and unobserved y (usually represented as y^{new} or y'), given the model. This is the likelihood of the replicated or predicted data, averaged over the posterior distribution $p(\theta|y)$:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta$$

or

$$p(y^{new}|y) = \int p(y^{new}|\theta)p(\theta|y)d\theta$$

If y has missing values, then the missing y's can be estimated with the posterior predictive distribution as y^{new} from within the model.

The posterior predictive distribution is easy to derive, since it is the same thing as the expectation μ in:

$$y^{new} \sim N(\mu, \tau^{-1})$$

where $\mu = \mathbf{X}\beta$, and μ is the conditional mean, while τ is the residual precision, which is equal to the inverse of the variance, σ^2.

10 Bayes Factors

Introduced by Harold Jeffreys (1961), a 'Bayes factor' is a Bayesian alternative to frequentist hypothesis testing that is most often used for the comparison of multiple models by hypothesis testing, usually to determine which model better fits the data. Bayes factors are notoriously difficult to compute, and the Bayes factor is only defined when the marginal density of y under each model is proper. Gelman finds Bayes factors generally to be irrelevant, because they compute the relative probabilities of the models conditional on one of them being true. Gelman prefers approaches that measure the distance of the data to each of the approximate models (Gelman, 2004, p. 180).

Two of many possible alternatives are to use:

1. pseudo Bayes factors (PsBF) based on a ratio of pseudo marginal likelihoods (PsML's), which in turn are the exponentiated sum of the log-marginal likelihoods
2. Deviance Information Criterion (DIC)

DIC is the most popular method of assessing model fit and comparing models.

11 Model Fit

In Bayesian inference, the most common method of assessing the goodness of fit of an estimated statistical model is a generalization of the frequentist Akaike Information Criterion (AIC). The Bayesian method, like AIC, is not a test of the model in the sense of hypothesis testing, though Bayesian inference has Bayes factors for such purposes. Instead, like AIC, Bayesian inference provides a model fit statistic that is to be used as a tool to refine the current model or select the better-fitting model of different methodologies.

To begin with, model fit can be summarized with deviance, which is defined as -2 times the log-likelihood (Gelman, 2004, p. 180):

$$D(y, \theta) = -2 \log[p(y|\theta)]$$

Just as with the likelihood, $p(y|\theta)$, or log-likelihood, the deviance exists at both the record- and model-level. Due to the development of `BUGS` software, deviance is defined differently in Bayesian inference than frequentist inference. In frequentist inference, deviance is -2 times

the log-likelihood ratio of a reduced model compared to a full model, whereas in Bayesian inference, deviance is simply -2 times the log-likelihood. In Bayesian inference, the lowest expected deviance has the highest posterior probability (Gelman, 2004, p. 181).

It is possible to have a negative deviance. Deviance is derived from the likelihood, which is derived from probability density functions (PDF). Evaluated at a certain point in parameter space, a PDF can have a density larger than 1 due to a small standard deviation or lack of variation. Likelihoods greater than 1 lead to negative deviance, and are appropriate.

On its own, the deviance is an insufficient model fit statistic, because it does not take model complexity into account. The effect of model fitting, pD, is used as the 'effective number of parameters' of a Bayesian model. The sum of the differences between the posterior mean of the model-level deviance and the deviance at each draw i of θ_i is the pD.

A related way to measure model complexity is as half the posterior variance of the model-level deviance, known as pV (Gelman, 2004, p. 182):

$$pV = var(D)/2$$

The effect of model fitting, pD or pV, can be thought of as the number of 'unconstrained' parameters in the model, where a parameter counts as: 1 if it is estimated with no constraints or prior information; 0 if it is fully constrained or if all the information about the parameter comes from the prior distribution; or an intermediate value if both the data and the prior are informative (Gelman, 2004, p. 182). Therefore, by including prior information, Bayesian inference is more efficient in terms of the effective number of parameters than frequentist inference. Hierarchical, mixed effects, or multilevel models are even more efficient regarding the effective number of parameters.

Model complexity, pD or pV, should be positive. Although pV must be positive since it is related to variance, it is possible for pD to be negative, which indicates one or more problems: log-likelihood is non-concave, a conflict between the prior and the data, or that the posterior mean is a poor estimator (such as with a bimodal posterior).

The sum of both the mean model-level deviance and the model complexity (pD or pV) is the Deviance Information Criterion (DIC), a model fit statistic:

$$DIC = \bar{D} + pV$$

DIC may be compared across different models and even different methods, as long as the dependent variable does not change between models, making DIC the most flexible model fit statistic. DIC is a hierarchical modeling generalization of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Like AIC and BIC, it is an asymptotic approximation as the sample size becomes large. DIC is valid only when the posterior distribution is approximately multivariate normal. Models should be preferred with smaller DIC. Since DIC increases with model complexity (pD or pV), simpler models are preferred.

It is difficult to say what would constitute an important difference in DIC. Very roughly, differences of more than 10 might rule out the model with the higher DIC, differences between 5 and 10 are substantial, but if the difference in DIC is, say, less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC.

The expected predictive deviance has been suggested as a criterion of model fit when the goal is to pick a model with the best out-of-sample predictive power. It can be approximately estimated with a predictive DIC, estimated by comparing the deviance between y^{new} and $y^{holdout}$.

12 Posterior Predictive Checks

Comparing the predictive distribution, y^{rep}, to the observed data y is generally termed a "posterior predictive check". This type of check includes the uncertainty associated with the estimated parameters of the model, unlike frequentist statistics.

Posterior predictive checks (via the predictive distribution) involve a double-use of the data, which violates the likelihood principle. Meng (1994) argues in favor of posterior predictive checks, provided that usage is limited to measures of discrepancy to study model adequacy, not for model comparison and inference.

Gelman recommends at the most basic level to compare y^{rep} to y, looking for any systematic differences, which could indicate potential failings of the model (Gelman, 2004, p. 159). It is often first recommended to compare graphical plots, such as the distribution of y and y^{rep}. There are many posterior predictive checks that are not included in this article, but an introduction to a selection of them appears below.

12.1 Bayesian p-values

A Bayesian form of p-value may be estimated with a variety of test statistics (Gelman et al., 1996). Usually the minimum or maximum observed y is compared to the minimum or maximum y^{rep}.

Frequentist p-values have many problems, but here it will only be noted that the frequentist p-value estimates $p(data|hypothesis)$, while in this case the Bayesian form estimates $p(hypothesis|data)$. The frequentist estimates the wrong probability, because the frequentist is forced to consider the parameters to be fixed and the data random, projecting long-run frequencies of what should happen with future, repeated sampling of similar data, given a fixed parameter, or in this case hypothesis. Even the term hypothesis testing suggests you want to test the hypothesis given the data, not the data given the hypothesis[7].

12.2 Conditional Predictive Ordinate

Although the full predictive distribution $p(y^{rep}|y)$ is useful for prediction, its use for model-checking is questionable because of the double-use of the data, and causes predictive performance to be overestimated. Geisser and Eddy (1979) proposed using the leave-one-out cross-validation predictive density. This is also known as the Conditional Predictive Ordinate or CPO (Gelfand, 1996).

The CPO is:

[7] Numerous problems with frequentist p-values, confidence intervals, point estimates, and hypothesis testing are worth exploring, but not detailed in this article.

$$p(y_i|y_{[i]}) = \int p(y_i|\theta)p(\theta|y_{[i]})d\theta$$

where y_i is each instance of an observed y, and $y_{[i]}$ is y without the current observation i.

The CPO is easy to calculate with MCMC numerical approximation. By considering the inverse likelihood across T iterations, the CPO for each individual i is:

$$CPO_i = \frac{1}{T^{-1}\sum_{t=1}^{T} p(y_i|\theta_t)^{-1}}$$

The CPO is a handy posterior predictive check because it may be used to identify outliers, influential observations, and for hypothesis testing across different non-nested models. However, it may be difficult to calculate with latent mixtures.

The CPO expresses the posterior probability of observing the value (or set of values) of y_i when the model is fitted to all data except y_i, with a larger value implying a better fit of the model to y_i, and very low CPO values suggest that y_i is an outlier and an influential observation. A Monte Carlo estimate of the CPO is obtained without actually omitting y_i from the estimation, and is provided by the harmonic mean of the likelihood for y_i. Specifically, the CPO_i is the inverse of the posterior mean of the inverse likelihood of y_i.

The CPO is connected with the frequentist studentized residual test for outlier detection. Data with large studentized residuals have small CPO's and will be detected as outliers. An advantage of the CPO is that observations with high leverage will have small CPO's, independently of whether or not they are outliers. The Bayesian CPO is able to detect both outliers and influential points, whereas the frequentist studentized residual is unable to detect high leverage outliers.

Ntzoufras (2009, p. 376) asserts that inverse-CPO's (ICPO's) larger than 40 can be considered as possible outliers, and higher than 70 as extreme values. In Bayesian Models for Categorical Data (2006), Congdon recommends scaling CPO's by dividing each by its individual maximum (after the posterior mean) and considering observations with scaled CPO's under 0.01 to be outliers. The range in scaled CPO's is useful as an indicator of a good-fitting model.

The sum of the logged CPO's can be an estimator for the log marginal likelihood, sometimes called the log pseudo marginal likelihood (LPsML). A ratio of PsML's is a surrogate for the Bayes factor, sometimes known as the pseudo Bayes factor (PsBF). In this way, non-nested models may be compared with a hypothesis test to determine the better model, if one exists, based on leave-one-out cross-validation.

12.3 Predictive Concordance

Gelfand (1996) suggests considering any y_i that is in either 2.5% tail area of y_i^{rep} to be an outlier. For each i, I am calling this the predictive quantile (PQ), which is calculated as:

$$PQ_i = p(y_i^{rep} > y_i)$$

which is somewhat similar to the Bayesian p-value. The percentage of y_i's that are not outliers is called the 'Predictive Concordance'. Gelfand suggests the goal is to attempt to achieve 95% predictive concordance. In the case of, say 80% predictive concordance, the discrepancy between the model and data is undesirable because the model does not fit the data well and many outliers have resulted. On the other hand, if the predictive concordance is too high, say 100%, then overfitting may have occurred, and it may be worth considering a more parsimonious model. Kernel density plots of each y_i^{rep} distribution are useful in this case with the actual y_i included as a vertical bar to show its position.

13 Advantages Of Bayesian Inference Over Frequentist Inference

Following is a short list of advantages of Bayesian inference over frequentist inference.

- Bayesian inference allows informative priors so that prior knowledge or results of a previous model to be used to inform the current model.
- Bayesian inference can avoid problems with model identification by manipulating prior distributions (usually in complex models). Frequentist inference with any numerical approximation algorithm does not have prior distributions, and can become stuck in regions of flat density, causing problems with model identification.
- Bayesian inference considers the data to be fixed (which it is), and parameters to be random because they are unknowns. Frequentist inference considers the unknown parameters to be fixed, and the data to be random, estimating not based on the data at hand, but the data at hand plus hypothetical repeated sampling in the future of similar data. "The Bayesian approach delivers the answer to the right question in the sense that Bayesian inference provides answers conditional on the observed data and not based on the distribution of estimators or test statistics over imaginary samples not observed" (Rossi, Allenby, and McCulloch, 2005, p. 4).
- Bayesian inference estimates a full probability model. Frequentist inference does not. There is no probability distribution associated with parameters or hypotheses.
- Bayesian inference estimates $p(hypothesis|data)$. Frequentist inference estimates $p(data|hypothesis)$. Even the term 'hypothesis testing' suggests it should be the hypothesis that is tested, given the data, not the other way around.
- Bayesian inference has an acxiomatic foundation that is uncontested by frequentists. Therefore, Bayesian inference is coherent to a frequentist, but frequentist inference is incoherent to a Bayesian.
- Bayesian inference includes uncertainty in the probability model, yielding more realistic predictions. Frequentist inference does not include uncertainty of the parameter estimates, yielding less realistic predictions.
- Bayesian inference may use DIC to compare models with different methods including hierarchical models, where frequentist model fit statistics cannot compare different methods or hierarchical models.

- Bayesian inference obeys the likelihood principle. Frequentist inference, including Maximum Likelihood Estimation (MLE) and the General Method of Moments (GMM) or Generalized Estimating Equations (GEE), violates the likelihood principle. "The likelihood principle, by itself, is not sufficient to build a method of inference but should be regarded as a minimum requirement of any viable form of inference. This is a controversial point of view for anyone familiar with modern econometrics literature. Much of this literature is devoted to methods that do not obey the likelihood principle..." (Rossi, Allenby, and McCulloch, 2005, p. 15).
- Bayesian inference uses observed data only. Frequentist inference uses both observed data and unobserved, hypothetical, future data.
- Bayesian inference uses prior distributions, so more information is used and 95% probability intervals of posterior distributions should be narrower than 95% confidence intervals of frequentist point-estimates.
- Bayesian inference uses probability intervals to state the probability that θ is between two points. Frequentist inference uses confidence intervals, which must be interpreted with probability of zero or one that θ is in the region, and the frequentist never knows whether it is or is not, but can only say that if 100 repeated samples were drawn in the future, that it would be in the region for 95 samples.
- Bayesian inference via MCMC algorithms allows more complicated models that frequentists are unable to estimate (Laplace Approximations work better with simpler models).
- Bayesian inference via MCMC has a theoretic guarantee than the MCMC algorithm will converge if run long enough. Frequentist inference with Maximum Likelihood Estimation (MLE) has no guarantee of convergence.
- Bayesian inference via MCMC is unbiased with respect to sample size and can accommodate any sample size no matter how small. Frequentist inference becomes more biased as sample size decreases from infinity, and is often wildly biased with small samples, so minimum sample size is an issue. Conversely, frequentist inference with large sample sizes biases p-values to indicate that insignificant effects are significant.
- Bayesian inference via MCMC uses exact estimation. Frequentist inference uses approximate estimation that relies on asymptotic theory.
- Bayesian inference with correlated predictors sometimes allows the prior parameters to be distributed multivariate-normal, therefore including such correlation into the MCMC algorithm to improve estimation. Frequentist inference does not use prior distributions, so standard errors are wider.
- Bayesian inference with proper priors is immune to singularities and near-singularities with matrix inversions, unlike frequentist inference.

14 Advantages Of Frequentist Inference Over Bayesian Inference

Following is a short list of advantages of frequentist inference over Bayesian inference.

- Frequentist models are able to include large data sets, while Bayesian models via MCMC are restricted to small sample sizes (though Bayesian models via Laplace Approximation can handle large data sets).
- Frequentist models are usually much easier to prepare because many things do not need to be specified, such as prior distributions, starting values for MCMC chains or Laplace Approximations, and usually the likelihood function. Most frequentist methods have been standardized to "procedures" where less knowledge and programming are required, and in many cases the user can just click on a few things and not really know what they are doing. Garbage in, garbage out.
- Frequentist models have much shorter run-times than Bayesian models via MCMC (though Laplace Approximation yields run-times that are almost as fast as the frequentist MLE). Simple models with small sample sizes may be similar, but complex models with larger sample sizes may be minutes (frequentist) vs. weeks (Bayesian via MCMC).

Please send an email to statisticat@gmail.com with any questions or concerns about this article.

15 References

Barnard, G.A. (1949). "Statistical Inference". *Journal of the Royal Statistical Society*, B 11, 115-149.

Bayarri, M.J. and DeGroot, M.H. (1987). "Bayesian Analysis of Selection Models". *The Statistician*, 36, 137-146.

Bayes, T. and Price, R. (1763). "An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.". *Philosophical Transactions of the Royal Society of London*, 53, 370-418.

Berger, J. (2006). "The Case for Objective Bayesian Analysis". *Bayesian Analysis*, 1(3), p. 385-402.

Bernardo, J.M. (2008). "Comment on Article by Gelman". *Bayesian Analysis*. 3(3), p. 451-454.

Bernardo, J.M. and Smith, A.F.M. (2000). Bayesian Theory. John Wiley & Sons: West Sussex, England.

Birnbaum, A. (1962). "On the Foundations of Statistical Inference". *Journal of the American Statistical Association*, 57, 296-306.

Congdon, P. (2006). Bayesian Models for Categorical Data. John Wiley & Sons: West Sussex, England.

De Morgan, A. (1837). "Review of Laplace's Theorie Analytique des Probabilites. (3rd Edition)." *Dublin Review*, 2, 3: 338-354, 237-248.

Geisser, S. and Eddy, W.F. (1979). "A Predictive Approach to Model Selection". *Journal of the American Statistical Association*, 74, 153-160.

Gelfand, A. (1996). "Model Determination Using Sampling Based Methods". In Gilks, W., Richardson, S., Spiegehalter, D., Chapter 9 in Markov Chain Monte Carlo in Practice. Chapman & Hall: Boca Raton, FL.

Gelman, A. (2006). "Prior Distributions for Variance Parameters in Hierarchical Models". *Bayesian Analysis*, 1, 3, p. 515-533.

Gelman, A. (2008). "Scaling Regression Inputs by Dividing by Two Standard Deviations". *Statistics in Medicine*, 27, p. 2865–2873.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D. (2004). Bayesian Data Analysis, Second Edition. Chapman & Hall: Boca Raton, FL.

Gelman, A., Meng, X.L., and Stern, H. (1996). "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies". *Statistica Sinica*, 6, p. 733–807.

Goldstein, M. (2006). "Subjective Bayesian Analysis: Principles and Practice". *Bayesian Analysis*, 1(3), p. 403-420.

Jeffreys, H. (1961). Theory of Probability, Third Edition. Oxford University Press.

Laplace, P.S. (1774). "Memoire sur la probabilite des causes par les evenements". Mem. Acad. Sci. Paris, 6, 621-656. English translation in 1986 as "Memoir on the probability of the causes of events", *Statistical Science*, 1, 359-378.

Laplace, P.S. (1812). Theorie Analytique des Probabilites. Paris: Courcier. Reprinted as Oeuvres Completes de Laplace, 7, 1878-1912. Paris: Gauthier-Villars.

Meng, X.L. (1994). "Posterior Predictive P-Values". *Annals of Statistics*, 22, 1142-1160.

Ntzoufras, I. (2009). Bayesian Modeling Using WinBUGS. John Wiley & Sons: West Sussex, England.

Rossi, P.E., Allenby, G.M., and McCulloch,, R. (2005). Bayesian Statistics and Marketing. John Wiley & Sons: West Sussex, England.