Bayesian Inference

Byron Hall
STATISTICAT, LLC

Abstract

The Bayesian interpretation of probability is one of two broad categories of interpretations. Bayesian inference updates knowledge about unknowns, parameters, with information from data. The **LaplacesDemon** package in R enables Bayesian inference, and this vignette provides an introduction to the topic. This article introduces Bayes' theorem, model-based Bayesian inference, components of Bayesian inference, prior distributions, hierarchical Bayes, conjugacy, likelihood, numerical approximation, prediction, Bayes factors, model fit, posterior predictive checks, and ends by comparing advantages and disadvantages of Bayesian inference.

Keywords:˜Bayesian, Laplace's Demon, LaplacesDemon, R, STATISTICAT.

This article is an introduction to Bayesian inference for users of the **LaplacesDemon** package (Hall 2012) in R (R Development Core Team 2012), 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 (Anscombe and Aumann 1963; Goldstein 2006). Objective Bayesian inference is often associated with Bayes and Price (1763), Laplace (1814), and Jeffreys (1961). Subjective Bayesian inference is often associated with Ramsey (1926), De˜Finetti (1931), and Savage (1954). The first major event to bring about the rebirth of Bayesian inference was De˜Finetti (1937). Differences in the interpretation of probability 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. In which case, consider Gelman, Carlin, Stern, and Rubin (2004) for a more thorough and formal introduction.

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 and Price 1763). 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

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{1}$$

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

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

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

[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.

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

1.2. Bayes' Theorem, Example 2

Another way to state Bayes' theorem is

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

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

- $\Pr(A_1) = \Pr(\text{Hell})$
- $\Pr(A_2) = \Pr(\text{Heaven})$
- $\Pr(B) = \Pr(\text{Consort})$
- $\Pr(A_1|B) = \Pr(\text{Hell}|\text{Consort})$
- $\Pr(A_2|B) = \Pr(\text{Heaven}|\text{Consort})$
- $\Pr(B|A_1) = \Pr(\text{Consort}|\text{Hell})$
- $\Pr(B|A_2) = \Pr(\text{Consort}|\text{Heaven})$

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

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

Finally, the desired conditional probability $\Pr(\text{Hell}|\text{Consort})$ is calculated using Bayes' theorem

- $\Pr(\text{Hell}|\text{Consort}) = \frac{0.666(0.75)}{0.666(0.75)+0.714(0.25)}$
- $\Pr(\text{Hell}|\text{Consort}) = 0.737$

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. Here is Bayes' theorem, equation 1, again

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

Replacing B with observations $\mathbf{y}$, A with parameter set Θ, and probabilities Pr with densities p (or sometimes π or function f), results in the following

$$p(\Theta|\mathbf{y}) = \frac{p(\mathbf{y}|\Theta)p(\Theta)}{p(\mathbf{y})}$$

where $p(\mathbf{y})$ will be discussed below, p(Θ) is the set of prior distributions of parameter set Θ before $\mathbf{y}$ is observed, $p(\mathbf{y}|\Theta)$ is the likelihood of $\mathbf{y}$ under a model, and $p(\Theta|\mathbf{y})$ 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

$$p(\mathbf{y}) = \int p(\mathbf{y}|\Theta)p(\Theta)d\Theta$$

defines the "marginal likelihood" of $\mathbf{y}$, or the "prior predictive distribution" of $\mathbf{y}$, and may be set to an unknown constant $\mathbf{c}$. The prior predictive distribution[3] indicates what $\mathbf{y}$ should look like, given the model, before $\mathbf{y}$ has been observed. Only the set of prior probabilities and the model's likelihood function are used for the marginal likelihood of $\mathbf{y}$. The presence of the marginal likelihood of $\mathbf{y}$ normalizes the joint posterior distribution, $p(\Theta|\mathbf{y})$, ensuring it is a proper distribution and integrates to one.

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

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

[3]The predictive distribution was introduced by Jeffreys (1961).

By removing **c** from the equation, the relationship changes from 'equals' (=) to 'proportional to' ($\propto$)[4]

$$p(\Theta|\mathbf{y}) \propto p(\mathbf{y}|\Theta)p(\Theta) \tag{2}$$

This form can be stated as the unnormalized joint posterior being proportional to the likelihood times the prior. However, the goal in model-based Bayesian inference is usually not to summarize the unnormalized 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 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

$$p(\phi|\mathbf{y}) = \int p(\phi, \Lambda|\mathbf{y})d\Lambda$$

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

3. Components of Bayesian Inference

The components[5] 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. $p(\mathbf{y}|\Theta)$ is the likelihood or likelihood function, in which all variables are related in a full probability model.

3. $p(\Theta|\mathbf{y})$ 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 single parameter of interest ϕ and the remaining parameters are considered nuisance parameters, then $p(\phi|\mathbf{y})$ is the marginal posterior distribution.

[4]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'.

[5]In Bayesian decision theory, an additional component exists (Roberts 2007, p. 53), the loss function, $\mathrm{L}(\Theta, \Delta)$.

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[6]. 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.

4.1. Informative Priors

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.

Sometimes informative prior information is not simply ready to be used, such as when it resides in another person, such as an expert. In this case, their personal beliefs about the probability of the event must be elicited into the form of a proper probability density function. This process is called prior elicitation.

4.2. Uninformative Priors

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[7]) attempt to minimize the impact of the selection of the

[6] 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).

[7] These terms are not all equivalent. For example, 'reference priors' are proposed, and are quite different from typical vague priors (Bernardo and Smith 2000). With reference priors, 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 $\mathbf{y}$ when the prior density is $p(\mathbf{y})$. In some sense, $p(\mathbf{y})$ is the 'least informative' prior about $\mathbf{y}$. Reference priors are often the objective prior of choice in multivariate problems,

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

$$\theta \sim \mathcal{U}(-\infty, \infty)$$

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 (Bayes and Price 1763), and used a flat prior for his billiard example so that all possible values of θ are equally likely *a priori* (Gelman *et˜al.* 2004, p. 34-36). Pierre-Simon Laplace (1749-1827) also used the flat prior to estimate the proportion of female births in a population, and for all estimation problems presented or justified as a reasonable expression of ignorance. 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 *et˜al.* 2004, p. 39).

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

$$\theta \sim \mathcal{N}(0, 10000)$$

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.

4.3. Proper and Improper Priors

It is important for the prior distribution to be proper. A prior distribution, $p(\theta)$, is improper[8] when

since other rules (e.g., Jeffreys' rule) may result in priors with problematic behavior.

[8]Improper priors were introduced in Jeffreys (1961).

$$\int p(\theta)d(\theta) = \infty$$

As noted previously, an unbounded uniform prior distribution is an improper prior distribution because $p(\theta) \propto 1$, for $-\infty < \theta < \infty$. An improper prior distribution can cause an improper posterior distribution. When the posterior distribution is improper, inferences are invalid, it is non-integrable, and Bayes factors cannot be used (though there are exceptions).

To determine the propriety of a posterior distribution, the marginal likelihood must be finite for all $\mathbf{y}$. Again, the marginal likelihood is

$$p(\mathbf{y}) = \int p(\mathbf{y}|\Theta)p(\Theta)d(\Theta)$$

Although improper prior distributions can be used, it is good practice to avoid them.

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. Estimating priors through hyperpriors, and from the data, is a method to elicit the optimal prior distributions. One of many natural uses for hierarchical Bayes is multilevel modeling.

Recall that the unnormalized joint posterior distribution (equation 2) is proportional to the likelihood times the prior distribution

$$p(\Theta|\mathbf{y}) \propto p(\mathbf{y}|\Theta)p(\Theta)$$

The simplest hierarchical Bayes model takes the form

$$p(\Theta, \Phi|\mathbf{y}) \propto p(\mathbf{y}|\Theta)p(\Theta|\Phi)p(\Phi)$$

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 priors $p(\Theta|\Phi)$, which in turn is used, as per usual, to estimate the likelihood $p(\mathbf{y}|\Theta)$, and finally the posterior is $p(\Theta, \Phi|\mathbf{y})$.

6. Conjugacy

When the posterior distribution $p(\Theta|\mathbf{y})$ 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[9]. For example, the Gaussian family is conjugate to itself (or self-conjugate) with respect to a Gaussian likelihood function: if the likelihood

[9]The conjugate prior approach was introduced in Raiffa and Schlaifer (1961).

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 *et˜al.* (2004) for a catalog.

Although the gamma distribution is the conjugate prior distribution for the precision of a normal distribution (Spiegelhalter, Thomas, Best, and Lunn 2003),

$$\tau \sim \mathcal{G}(0.001, 0.001),$$

better properties for scale parameters are yielded with the non-conjugate, proper, half-Cauchy[10] distribution, with a general recommendation of scale=25 for an uninformative scale parameter (Gelman 2006),

$$\begin{aligned} \sigma &\sim \mathcal{HC}(25) \\ \tau &= \sigma^{-2} \end{aligned}$$

When the half-Cauchy is unavailable, a uniform distribution is often placed on σ in hierarchical Bayes when the number of groups is, say, at least five,

$$\begin{aligned} \sigma &\sim \mathcal{U}(0, 100) \\ \tau &= \sigma^{-2} \end{aligned}$$

Conjugacy is mathematically convenient in that the posterior distribution follows a known parametric form (Gelman *et˜al.* 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 *et˜al.* 2004, p. 41).

Nonconjugate prior distributions can make interpretations of posterior inferences less transparent and computation more difficult, though this alternative does not pose any conceptual problems. In practice, for complicated models, conjugate prior distributions may not even be possible (Gelman *et˜al.* 2004, p. 41-42).

When conjugate distributions are used, a summary statistic for a posterior distribution of θ may be represented as $t(\mathbf{y})$ and said to be a sufficient statistic (Gelman *et˜al.* 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 $t(\mathbf{y})$ is said to be a sufficient statistic for θ, because the likelihood for θ depends on the data $\mathbf{y}$ only through the value of $t(\mathbf{y})$. Sufficient statistics are useful in algebraic manipulations of likelihoods and posterior distributions.

[10] The half-t distribution is another option.

7. Likelihood

In order to complete the definition of a Bayesian model, both the prior distributions and the likelihood[11] must be approximated or fully specified. The likelihood, likelihood function, or $p(\mathbf{y}|\Theta)$, contains the available information provided by the sample. The likelihood is

$$p(\mathbf{y}|\Theta) = \prod_{i=1}^{n} p(\mathbf{y}_i|\Theta)$$

The data $\mathbf{y}$ affects the posterior distribution $p(\Theta|\mathbf{y})$ only through the likelihood function $p(\mathbf{y}|\Theta)$. In this way, Bayesian inference obeys the likelihood principle, which states that for a given sample of data, any two probability models $p(\mathbf{y}|\Theta)$ that have the same likelihood function yield the same inference for Θ. The likelihood principle is covered in more detail later in the 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 (Bayes and Price 1763). '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, $p(\mathbf{y}^{new}|\mathbf{y}, \theta)$, but to estimate this predicted probability, the probability distribution of θ must first be estimated, using coin toss data $\mathbf{y}$ to estimate the parameter θ by the likelihood function $p(\mathbf{y}|\theta)$, which contains the likelihood $p(\theta|\mathbf{y})$, where θ is estimated from the data, $\mathbf{y}$. Therefore, the data, $\mathbf{y}$, is used to estimate the most probable θ that would lead to a data-generating process for $\mathbf{y}$.

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* (Laplace 1812). 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 (Laplace 1814). In it, the Bayesian interpretation of probability was developed independently by Laplace, much

[11] Ronald A. Fisher, a prominent frequentist, introduced the term likelihood in 1921 (Fisher 1921), though the concept of likelihood was used by Bayes and Laplace. Fisher's introduction preceded a series of the most influential papers in statistics (mostly in 1922 and 1925), in which Fisher introduced numerous terms that are now common: consistency, efficiency, estimation, information, maximum likelihood estimate, optimality, parameter, statistic, sufficiency, and variance. He was the first to use Greek letters for unknown parameters and Latin letters for the estimates. Later contributions include F statistics, design of experiments, ANOVA, and many more.

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 (De˜Morgan 1837), 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 "distribution" 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 Ronald A. Fisher as a derogatory term.

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

However, $p(\mathbf{y}|\theta)$ 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 the use of maximum likelihood estimation (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 $\mathbf{y}$ affect posterior inference 'only' through the function $p(\mathbf{y}|\theta)$, which, when regarded as a function of θ, for fixed $\mathbf{y}$, 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 $p(\mathbf{y}|\theta)$ that have the same likelihood function yield the same inference for θ" (Gelman *et˜al.* 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 *et˜al.* 2004, p. 9).

> Frequentist "procedures typically violate the likelihood principle, since long-run behavior under hypothetical repetitions depends on the entire distribution $p(\mathbf{y}|\theta)$, $\mathbf{y} \in \mathbf{Y}$ 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 $p_1(\mathbf{y}_1|\theta), p_2(\mathbf{y}_2|\theta)$ 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 (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, $\hat{\theta} = f(\mathbf{y})$. Some sort of sampling experiment results in a distribution of $\mathbf{y}$ 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 *et˜al.* 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 θ, the conditional probability of $\mathbf{y}$ is the density $p(\mathbf{y}|\theta)$. With θ as a parameter, here are relationships in expressions of the likelihood function

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

where $\mathbf{y}$ is the observed outcome of an experiment, and the likelihood ($\mathscr{L}$) of θ given $\mathbf{y}$ is equal to the density $p(\mathbf{y}|\theta)$ or function $f(\mathbf{y}|\theta)$. When viewed as a function of $\mathbf{y}$ with θ fixed, it is not a likelihood function $\mathscr{L}(\theta|\mathbf{y})$, but merely a probability density function $p(\mathbf{y}|\theta)$. When viewed as a function of θ with $\mathbf{y}$ fixed, it is a likelihood function and may be denoted as $\mathscr{L}(\theta|\mathbf{y})$, $p(\mathbf{y}|\theta)$, or $f(\mathbf{y}|\theta)$[12].

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

$$\mathbf{y}_i \sim \mathcal{N}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_1 + \beta_2 \mathbf{X}_{i,1} + \beta_3 \mathbf{X}_{i,2}$$

The dependent variable $\mathbf{y}$, indexed by $i = 1, ..., n$, is stochastic, and normally-distributed according to the expectation vector μ, and variance σ^2. Expectation vector μ is an additive, linear function of a vector of regression parameters, β, and the design matrix $\mathbf{X}$.

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

$$f(\mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma} \exp[(-\frac{1}{2}\sigma^2)(\mathbf{y}_i - \mu_i)^2]; \quad \mathbf{y} \in (-\infty, \infty)$$

By considering a conditional distribution, the record-level likelihood in Bayesian notation is

$$p(\mathbf{y}_i|\Theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp[(-\frac{1}{2}\sigma^2)(\mathbf{y}_i - \mu_i)^2]; \quad \mathbf{y} \in (-\infty, \infty)$$

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

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.

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

rather than

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

As a function of θ, the unnormalized 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 unnormalized joint posterior distribution

$$p(\beta, \sigma^2|\mathbf{y}) = p(\mathbf{y}|\beta, \sigma^2)p(\beta_1)p(\beta_2)p(\beta_3)p(\sigma^2)$$

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

$$\log[p(\beta, \sigma^2|\mathbf{y})] = \log[p(\mathbf{y}|\beta, \sigma^2)] + \log[p(\beta_1)] + \log[p(\beta_2)] + \log[p(\beta_3)] + \log[p(\sigma^2)]$$

The logarithm of the unnormalized 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.

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). Since MCMC is most common, and because the R package entitled **LaplacesDemon** offers Laplace Approximation and seven MCMC algorithms (as well as ABC in these contexts), ABC, Laplace Approximation and MCMC are introduced below. For more information on MCMC algorithms in **LaplacesDemon**, see the accompanying vignette entitled "**LaplacesDemon** Tutorial".

Approximate Bayesian Computation (ABC), also called likelihood-free estimation, is a family of numerical approximation techniques in Bayesian inference. ABC is especially useful when evaluation of the likelihood, $p(\mathbf{y}|\Theta)$ is computationally prohibitive, or when suitable likelihoods are unavailable. As such, ABC algorithms estimate likelihood-free approximations. ABC is usually faster than a similar likelihood-based numerical approximation technique, because the likelihood is not evaluated directly, but replaced with an approximation that is usually easier to calculate. The approximation of a likelihood is usually estimated with a measure of distance between the observed sample, $\mathbf{y}$, and its replicate given the model, $\mathbf{y}^{rep}$, or with summary statistics of the observed and replicated samples.

Laplace Approximation dates back to Laplace (1774, 1814), and is used to approximate the posterior moments of integrals. Specifically, the posterior mode is estimated for each parameter, assumed to be unimodal and Gaussian. As a Gaussian distribution, the posterior mean is the same as the posterior mode, and the variance is estimated. Laplace Approximation is a family of deterministic algorithms that usually converge faster than MCMC, and just a little slower than Maximum Likelihood Estimation (MLE) (Azevedo-Filho and Shachter 1994). Laplace Approximation shares many limitations of MLE, including asymptotic estimation with respect to sample size.

MCMC algorithms originated in statistical physics and are now used in Bayesian inference to sample from probability distributions by constructing Markov chains. In Bayesian inference, the target distribution of each Markov chain is usually a marginal posterior distribution, such as each parameter θ. Each Markov chain begins with an initial value and the algorithm iterates, attempting to maximize the logarithm of the unnormalized joint posterior distribution and eventually arriving at each target distribution. Each iteration is considered a state. A Markov chain is a random process with a finite state-space and the Markov property, meaning that the next state depends only on the current state, not on the past. The quality of the marginal samples usually improves with the number of iterations.

A Monte Carlo method is an algorithm that relies on repeated pseudo-random sampling for computation, and is therefore stochastic (as opposed to deterministic). Monte Carlo methods are often used for simulation. The union of Markov chains and Monte Carlo methods is called MCMC. The revival of Bayesian inference since the 1980s is due to MCMC algorithms and increased computing power. The most prevalent MCMC algorithms may be the simplest: random-walk Metropolis and Gibbs sampling. There are a large number of MCMC algorithms, and further details on MCMC are best explored outside of this article.

9. Prediction

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

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

or

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

If $\mathbf{y}$ has missing values, then the missing $\mathbf{y}$s can be estimated with the posterior predictive distribution[13] as $\mathbf{y}^{new}$ from within the model. For the linear regression example, the integral for prediction is

$$p(\mathbf{y}^{new}|\mathbf{y}) = \int p(\mathbf{y}^{new}|\beta, \sigma^2)p(\beta, \sigma^2|\mathbf{y})d\beta d\sigma^2$$

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

$$\mathbf{y}^{new} \sim \mathcal{N}(\mu, \sigma^2)$$

where $\mu = \mathbf{X}\beta$, and μ is the conditional mean, while σ^2 is the residual variance.

10. Bayes Factors

Introduced by Harold Jeffreys, 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 (Jeffreys 1961). Bayes factors are notoriously difficult to compute, and the Bayes factor is only defined when the marginal density of $\mathbf{y}$ under each model is proper. However, Bayes factors are easy to approximate with the Laplace-Metropolis Estimator (Kass and Raftery 1995; Lewis and Raftery 1997)[14].

Hypothesis testing with Bayes factors is more robust than frequentist hypothesis testing, since the Bayesian form avoids model selection bias, evaluates evidence in favor the null hypothesis, includes model uncertainty, and allows non-nested models to be compared (though of course the model must have the same dependent variable). Also, frequentist significance tests become biased in favor of rejecting the null hypothesis with sufficiently large sample size.

The Bayes factor for comparing two models may be approximated as the ratio of the marginal likelihood of the data in model 1 and model 2. Formally, the Bayes factor in this case is

$$B = \frac{p(\mathbf{y}|\mathcal{M}_1)}{p(\mathbf{y}|\mathcal{M}_2)} = \frac{\int p(\mathbf{y}|\Theta_1, \mathcal{M}_1)p(\Theta_1|\mathcal{M}_1)d\Theta_1}{\int p(\mathbf{y}|\Theta_2, \mathcal{M}_2)p(\Theta_2|\mathcal{M}_2)d\Theta_2}$$

where $p(\mathbf{y}|\mathcal{M}_1)$ is the marginal likelihood of the data in model 1.

The Bayes factor, B, is the posterior odds in favor of the hypothesis divided by the prior odds in favor of the hypothesis, where the hypothesis is usually $\mathcal{M}_1 > \mathcal{M}_2$. Put another way,

$$\text{(Posterior model odds)} = \text{(Bayes factor) x (prior model odds)}$$

[13] The predictive distribution was introduced by Jeffreys (1961).

[14] A Bayes factor may be estimated with the `BayesFactor` function in **LaplacesDemon** to compare multiple models that were fit with the `LaplaceApproximation` or `LaplacesDemon` functions. See the `BayesFactor` function for the interpretation of a Bayes factor regarding strength of evidence.

For example, when $B = 2$, the data favor $\mathcal{M}_1$ over $\mathcal{M}_2$ with 2:1 odds. Table **??** shows a variation of Jeffreys' interpretation of the Bayes factor (Jeffreys 1961).

In a non-hierarchical model, the marginal likelihood may easily be approximated with the Laplace-Metropolis Estimator for model m as

$$p(\mathbf{y}|m) = (2\pi)^{d_m/2}|\Sigma_m|^{1/2}p(\mathbf{y}|\Theta_m, m)p(\Theta_m|m)$$

where d is the number of parameters and Σ is the inverse of the negative of the Hessian matrix of second derivatives. Lewis and Raftery (1997) introduce the Laplace-Metropolis method of approximating the marginal likelihood in MCMC, though it naturally works with Laplace Approximation as well. For a hierarchical model that involves both fixed and random effects, the Compound Laplace-Metropolis Estimator must be used.

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 *et˜al.* 2004, p. 180). However, Kass and Raftery (1995) explain that "the logarithm of the marginal probability of the data may also be viewed as a predictive score. This is of interest, because it leads to an interpretation of the Bayes factor that does not depend on viewing one of the models as 'true"'.

Two of many possible alternatives are to use

1. pseudo Bayes factors (PsBF) based on a ratio of pseudo marginal likelihoods (PsMLs)
2. Deviance Information Criterion (DIC)

DIC is the most popular method of assessing model fit and comparing models, though Bayes Factor are better, when appropriate, because they take more into account.

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 *et˜al.* 2004, p. 180), such as

$$D(\mathbf{y}, \Theta) = -2\log[p(\mathbf{y}|\Theta)]$$

Just as with the likelihood, $p(\mathbf{y}|\Theta)$, 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 *et˜al.* 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 *et˜al.* 2004, p. 182)

$$\text{pV} = \text{var}(D)/2$$

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 *et˜al.* 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

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

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 joint 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 $\mathbf{y}^{new}$ and $\mathbf{y}^{holdout}$.

12. Posterior Predictive Checks

Comparing the predictive distribution $\mathbf{y}^{rep}$ to the observed data $\mathbf{y}$ 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. However, arguments have been made 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 (Meng 1994).

Gelman recommends at the most basic level to compare $\mathbf{y}^{rep}$ to $\mathbf{y}$, looking for any systematic differences, which could indicate potential failings of the model (Gelman *et˜al.* 2004, p. 159). It is often first recommended to compare graphical plots, such as the distribution of $\mathbf{y}$ and $\mathbf{y}^{rep}$. 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, Meng, and Stern 1996). Usually the minimum or maximum observed $\mathbf{y}$ is compared to the minimum or maximum $\mathbf{y}^{rep}$. A Bayesian p-value is one of several ways to report discrepancies between $\mathbf{y}$ and $\mathbf{y}^{rep}$.

Frequentist p-values have many problems, but here it will only be noted that the frequentist p-value estimates $p(\text{data}|\text{hypothesis})$, while in this case the Bayesian form estimates $p(\text{hypothesis}|\text{data})$. 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[15].

12.2. Chi-Square

(Gelman *et˜al.* 2004, p. 175) suggest an omnibus test such as the following χ^2

$$\chi_i^2 = \frac{(\mathbf{y}_i - \frac{\sum_{t=1}^{T} \mathbf{y}_{i,t}^{rep}}{T})^2}{\text{var}(\mathbf{y}_{i,1:T}^{rep})},$$

over records $i = 1, \ldots, N$ and posterior samples $t = 1, \ldots, T$. The sum of χ_i^2 over records $i = 1, \ldots, N$ is an overall goodness of fit measure on the data set. Larger values of χ_i^2 indicate a worse fit for each record.

An alternative χ^2 test is

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

$$p(\chi^{2rep}_{i,1:T} > \chi^{2obs}_{i,1:T})$$

where a worse fit is indicated as p approaches zero or one, and it is common to consider records with a poor fit to be outside the 95% probability interval. To continue

$$\chi^{2obs}_{i,1:T} = \frac{[\mathbf{y}_i - E(\mathbf{y}_i)]^2}{E(\mathbf{y}_i)}$$

and

$$\chi^{2rep}_{i,1:T} = \frac{[\mathbf{y}^{rep}_{i,1:T} - E(\mathbf{y}^{rep}_i)]^2}{E(\mathbf{y}^{rep}_i)}$$

Newer forms of χ^2 tests have been proposed in the literature, and are best explored elsewhere.

12.3. Conditional Predictive Ordinate

Although the full predictive distribution $p(\mathbf{y}^{rep}|\mathbf{y})$ 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. The leave-one-out cross-validation predictive density has been proposed (Geisser and Eddy 1979). This is also known as the Conditional Predictive Ordinate or CPO (Gelfand 1996).

The CPO is

$$p(\mathbf{y}_i|\mathbf{y}_{[i]}) = \int p(\mathbf{y}_i|\Theta)p(\Theta|\mathbf{y}_{[i]})d\Theta$$

where $\mathbf{y}_i$ is each instance of an observed $\mathbf{y}$, and $\mathbf{y}_{[i]}$ is $\mathbf{y}$ 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

$$\text{CPO}_i = \frac{1}{T^{-1}\sum_{t=1}^{T} p(\mathbf{y}_i|\Theta_t)^{-1}}$$

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 $\mathbf{y}_i$ when the model is fitted to all data except $\mathbf{y}_i$, with a larger value implying a better fit of the model to $\mathbf{y}_i$, and very low CPO values suggest that $\mathbf{y}_i$ is an outlier and an influential observation. A Monte Carlo estimate of the CPO is obtained without actually omitting $\mathbf{y}_i$ from the estimation, and is provided by the harmonic mean of the likelihood for $\mathbf{y}_i$. Specifically, the CPO_i is the inverse of the posterior mean of the inverse likelihood of $\mathbf{y}_i$.

The CPO is connected with the frequentist studentized residual test for outlier detection. Data with large studentized residuals have small CPOs and will be detected as outliers.

An advantage of the CPO is that observations with high leverage will have small CPOs, 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.

Inverse-CPOs (ICPOs) larger than 40 can be considered as possible outliers, and higher than 70 as extreme values (Ntzoufras 2009, p. 376). Congdon recommends scaling CPOs by dividing each by its individual maximum (after the posterior mean) and considering observations with scaled CPOs under 0.01 to be outliers (Congdon 2005). The range in scaled CPOs is useful as an indicator of a good-fitting model.

The sum of the logged CPOs can be an estimator for the logarithm of the marginal likelihood[16], sometimes called the log pseudo marginal likelihood (LPsML). A ratio of PsMLs 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.

12.4. Predictive Concordance

Gelfand (1996) suggests that any $\mathbf{y}_i$ that is in either 2.5% tail area of $\mathbf{y}_i^{rep}$ should be considered an outlier. For each i, I am calling this the predictive quantile (PQ), which is calculated as

$$\text{PQ}_i = p(\mathbf{y}_i^{rep} > \mathbf{y}_i)$$

and is somewhat similar to the Bayesian p-value. The percentage of $\mathbf{y}_i$s that are not outliers is called the 'Predictive Concordance'. Gelfand (1996) 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 $\mathbf{y}_i^{rep}$ distribution are useful in this case with the actual $\mathbf{y}_i$ included as a vertical bar to show its position.

12.5. L-criterion

Laud and Ibrahim (1995) introduced the L-criterion as one of three posterior predictive checks for model and variable selection. The L-criterion is a posterior predictive check that is widely applicable and easy to apply. It is a sum of two components: one involves the predictive variance and the other includes the accuracy of the means of the predictive distribution. The L-criterion measures model performance with a combination of how close its predictions are to the observed data and variability of the predictions. Better models have smaller values of L. L is measured in the same units as the response variable, and measures how close the data vector $\mathbf{y}$ is to the predictive distribution. In addition to the value of L, there is a value for S_L, which is the calibration number of L, and is useful in determining how much of a decrease is necessary between models to be noteworthy. The L-criterion is

[16] Exercise extreme caution when approximating the marginal likelihood from CPOs, or use a method with better repute, such as the Laplace-Metropolis Estimator or importance sampling.

$$\mathrm{L} = \sum_{i=1}^{N} \sqrt{\mathrm{var}(\mathbf{y}_{i,1:T}^{rep}) + (\mathbf{y}_i - \frac{\sum_{t=1}^{T} \mathbf{y}_{i,t}^{rep}}{T})^2},$$

over $t = 1, \ldots, T$ posterior samples. The calibration number, S_L, is the standard deviation of L over records $i = 1, \ldots, N$.

The L-criterion has been extended, such as including weights, and is best explored elsewhere.

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 can 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 with 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 *et˜al.* 2005, p. 4).
- Bayesian inference estimates a full probability model. Frequentist inference does not. There is no frequentist probability distribution associated with parameters or hypotheses.
- Bayesian inference estimates p(hypothesis|data). In contrast, 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 axiomatic foundation (Cox 1946) that is uncontested by frequentists. Therefore, Bayesian inference is coherent to a frequentist, but frequentist inference is incoherent to a Bayesian.
- Bayesian inference has a decision theoretic foundation (Bernardo and Smith 2000; Roberts 2007). The purpose of most of statistical inference is to facilitate decision-making (Roberts 2007, p. 51). The optimal decision is the Bayesian decision.
- 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 is consistent with much of philosophy of science regarding epistemology, where knowledge cannot be built entirely through experimentation, but requires prior knowledge (Roberts 2007, p. 510). Elsewhere, it has been suggested that the best choice for philosophy of science is through Bayesian inference.
- 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 *et˜al.* 2005, p. 15).
- Bayesian inference uses observed data only. Frequentist inference uses both observed data and future data that is unobserved and hypothetical.
- 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.
- 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 with respect to sample size. Frequentist inference uses approximate estimation that relies on asymptotic theory.
- Bayesian inference with correlated predictors sometimes allows the hyperparameters to be distributed multivariate-normal, therefore including such correlation into the MCMC algorithm to improve estimation. Frequentist inference does not use prior distributions, so confidence intervals are wider and less certain with correlated predictors.

- 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 have traditionally been restricted to small sample sizes (though Bayesian models via Laplace Approximation can handle large data sets). MCMC algorithms in Laplace's Demon, however, do not loop through records, and because these algorithms are vectorized in this respect, 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, initial values for numerical approximation, 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.

References

Anscombe F, Aumann R (1963). "A Definition of Subjective Probability." *The Annals of Mathematical Statistics*, **34**(1), 199–205.

Azevedo-Filho A, Shachter R (1994). "Laplace's Method Approximations for Probabilistic Inference in Belief Networks with Continuous Variables." In R˜Mantaras, D˜Poole (eds.), *Uncertainty in Artificial Intelligence*, pp. 28–36. Morgan Kauffman, San Francisco, CA.

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

Bayarri M, DeGroot M (1987). "Bayesian Analysis of Selection Models." *The Statistician*, **36**, 136–146.

Bayes T, Price R (1763). "An Essay Towards Solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, MA. 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), 385–402.

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

Bernardo J, Smith A (2000). *Bayesian Theory*. John Wiley & Sons, West Sussex, England.

Congdon P (2005). *Bayesian Models for Categorical Data*. John Wiley & Sons, West Sussex, England.

Cox R (1946). "Probability, Frequency and Reasonable Expectation." *American Journal of Physics*, **14**(1), 1–13.

De˜Finetti B (1931). "Probabilismo." *Erkenntnis*, **31**, 169–223. English translation as "Probabilism: A Critical Essay on the Theory of Probability and on the Value of Science".

De˜Finetti B (1937). "La Prevision: ses lois logigues, ses sources subjectives." *Annales de l'Institut Henri Poincare*. English translation in H.E. Kyburg and H.E. Smokler (eds), (1964), "Foresight: Its Logical Laws, Its Subjective Sources", Studies in Subjective Probability, New York: Wiley.

De˜Morgan A (1837). "Review of Laplace's Theorie Analytique des Probabilites." *Dublin Review*, **2,3**, 338–354,237–354.

Fisher R (1921). "On the 'Probable Error' of a Coefficient of Correlation Deduced From a Small Sample." *Metron*, **1**(4), 3–32.

Geisser S, Eddy W (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 W˜Gilks, S˜Richardson, D˜Spiegelhalter (eds.), *Markov Chain Monte Carlo in Practice*, pp. 145–161. Chapman & Hall, Boca Raton, FL.

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

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

Gelman A, Carlin J, Stern H, Rubin D (2004). *Bayesian Data Analysis*. 2nd edition. Chapman & Hall, Boca Raton, FL.

Gelman A, Meng X, Stern H (1996). "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies." *Statistica Sinica*, **6**, 773–807.

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

Hall B (2012). ***LaplacesDemon**: Software for Bayesian Inference*. R package version 12.06.04, URL http://cran.r-project.org/web/packages/LaplacesDemon/index.html.

Jeffreys H (1961). *Theory of Probability*. Third edition. Oxford University Press, Oxford, England.

Kass R, Raftery A (1995). "Bayes Factors." *Journal of the American Statistical Association*, **90**(430), 773–795.

Laplace P (1774). "Memoire sur la Probabilite des Causes par les Evenements." *l'Academie Royale des Sciences*, **6**, 621–656. English translation by S.M. Stigler in 1986 as "Memoir on the Probability of the Causes of Events" in *Statistical Science*, **1**(3), 359–378.

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

Laplace P (1814). "Essai Philosophique sur les Probabilites." English translation in Truscott, F.W. and Emory, F.L. (2007) from (1902) as "A Philosophical Essay on Probabilities". ISBN 1602063281, translated from the French 6th ed. (1840).

Laud P, Ibrahim J (1995). "Posterior Model Selection." *Journal of the Royal Statistical Society*, **B 57**, 247–262.

Lewis S, Raftery A (1997). "Estimating Bayes' Factors via Posterior Simulation with the Laplace-Metropolis Estimator." *Journal of the American Statistical Association*, **92**, 648–655.

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

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

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raiffa H, Schlaifer R (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University.

Ramsey F (1926). "Truth and Probability." In R˜Braithwaite (ed.), *The Foundations of Mathematics and other Logical Essays*, pp. 156–198. Harcourt, Brace and Company, New York.

Roberts C (2007). *The Bayesian Choice*. 2nd edition. Springer, Paris, France.

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

Savage L (1954). *The Foundations of Statistics*. John Wiley & Sons, West Sussex, England.

Spiegelhalter D, Thomas A, Best N, Lunn D (2003). *WinBUGS User Manual, Version 1.4*. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK. http://www.mrc-bsu.cam.ac.uk/bugs.

Affiliation:

Byron Hall
STATISTICAT, LLC
Farmington, CT
E-mail: statisticat@gmail.com
URL: http://www.statisticat.com/laplacesdemon.html