Bayesian Statistical Methods

Jeff Gill

Department of Political Science, University of Florida

234 Anderson Hall, PO Box 117325, Gainesville, FL 32611-7325

Voice: 352-392-0262x272, Fax: 352-392-8127, Email: jgill@polisci.ufl.edu

Word Count: 2339.

Bayes' Law

The Bayesian statistical approach is based on updating information using a probability theorem from the famous 1763 essay by the Reverend Thomas Bayes. He was an amateur mathematician whose work was found and published two years after his death by his friend Richard Price. The enduring association of an important branch of statistics with his name is due to the fundamental importance of this probability statement now called *Bayes' Law* or *Bayes' Theorem* which relates conditional probabilities.

Start with two events of interest X and Y, which are not independent. We know from the basic axioms of probability that the conditional probability of X given that Y has occurred is obtained by: $p(X|Y) = \frac{p(X,Y)}{p(Y)}$, where $p(X,Y)$ is the "the probability that both X and Y occur," and $p(Y)$ is the unconditional probability that Y occurs. The conditional expression thus gives the probability of X after some event Y occurs.

We can also define a different conditional probability in which X occurs first, $p(Y|X) = \frac{p(Y,X)}{p(X)}$. Since the joint probability that X and Y occur is the same as the joint probability that Y and X occur ($p(X,Y) = p(Y,X)$), then it is possible to make

the following simple rearrangement:

$$p(X,Y) = p(X|Y)p(Y), \qquad p(Y,X) = p(Y|X)p(X)$$

$$p(X|Y)p(Y) = p(Y|X)p(X)$$

$$\therefore p(X|Y) = \frac{p(X)}{p(Y)}p(Y|X) \tag{1}$$

(for more details, see Bernardo and Smith [1994, Ch.3]). The last line is the famous Bayes' law and this is really a device for "inverting" conditional probabilities. It is clear that one could just as easily produce $p(Y|X)$ in the last line above by moving the unconditional probabilities to the left-hand side in the last equality. Bayes' Law is useful because we often know $p(X|Y)$ and would like to know $p(Y|X)$, or vice-versa. The fact that $p(X|Y)$ is never equal to $p(Y|X)$ (that is, the probability that $\frac{p(X)}{p(Y)} = 1$ is zero) is often called the *inverse probability problem*.

It turns out that this simple result is the foundation for the general paradigm of Bayesian statistics.

Description of Bayesian Inference

Bayesian inference is based on fundamentally different assumptions about data and parameters than classical methods (Box and Tiao 1973, Ch.1). In the Bayesian world all quantities are divided into two groups: observed and unobserved. Observed quantities are typically the data and any known relevant constants. Unobserved quantities include parameters of interest to be estimated, missing data, and parameters of lesser interest that simply need to be accounted for. All observed quantities are fixed and are conditioned on. All unobserved quantities are assumed to possess distributional qualities and therefore are treated as random variables. Thus parameters are

now no longer treated as fixed unmoving in the total population, and all statements are made in probabilistic terms.

The inference process starts with assigning *prior distributions* for the unknown parameters. These range from very informative descriptions of previous research in the field to deliberately vague and diffuse forms that reflect relatively high levels of ignorance. The prior distribution is not an inconvenience imposed by the treatment of unknown quantities, it is instead an opportunity to systematically include qualitative, narrative, and intuitive knowledge into the statistical model. The next step is to stipulate a likelihood function in the conventional manner by assigning a parametric form for the data and inserting the observed quantities. The final step is to produce a *posterior distribution* by multiplying the prior distribution and the likelihood function. Thus the likelihood function uses the data to *update* the prior knowledge conditionally.

This process, as described, can be summarized by the simple mnemonic:

$$\text{Posterior Probability} \propto \text{Prior Probability} \times \text{Likelihood Function}.$$

This is just a form of Bayes' Law where the denominator on the right-hand-side has been ignored by using proportionality. The symbol "$\propto$" stands for "proportional to", which means that constants have been left out that make the posterior sum or integrate to one as is required of probability mass functions and probability density functions. Renormalizing to a proper form can always be done later, plus using proportionality is more intuitive and usually reduces the calculation burden. What this "formula" above shows is that the posterior distribution is a compromise between the prior distribution, reflecting research beliefs, and the likelihood function, which is the contribution of the data at hand.

To summarize, the Bayesian inference process is given in three general steps:

I. Specify a probability model that includes some prior knowledge about the parameters if available for unknown parameter values.

II. Update knowledge about the unknown parameters by conditioning this probability model on observed data.

III. Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

Nowhere in this process is there an artificial decision based on the assumption that some null hypothesis of no effect is true. Therefore unlike the seriously flawed null hypothesis significance test, evidence is presented by simply summarizing this posterior distribution. This is usually done with quantiles and probability statements such as the probability that the parameter of interest is less than/greater than some interesting constant, or the probability that this parameter occupies some region. Also note that if the posterior distribution is now treated as a new prior distribution, it too can be updated if new data are observed. Thus knowledge about the parameters of interest is updated and accumulated over time (Robert 2001).

The Likelihood Function

Suppose collected data are treated as a fixed quantity and we know the appropriate probability mass function or probability density function for describing the data-generation process. Standard likelihood and Bayesian methods are similar in that they both start with these two suppositions and then develop estimates of the unknown parameters in the parametric model. Maximum likelihood estimation substitutes the unbounded notion of likelihood for the bounded definition of probability by starting

with Bayes' Law:

$$p(\theta|\mathbf{X}) = \frac{p(\theta)}{p(\mathbf{X})} p(\mathbf{X}|\theta) \tag{2}$$

where θ is the unknown parameter of interest and $\mathbf{X}$ is the collected data. The key is to treat $\frac{p(\theta)}{p(\mathbf{X})}$ as an unknown function of the data independent of $p(\mathbf{X}|\theta)$. This allows us to use: $L(\theta|\mathbf{X}) \propto p(\mathbf{X}|\theta)$. Since the data are fixed, then different values of the likelihood function are obtained merely by inserting different values of the unknown parameter, θ.

The likelihood function, $L(\theta|\mathbf{X})$, is similar to the desired but unavailable inverse probability, $p(\mathbf{X}|\theta)$, in that it facilitates testing alternate values of θ, to find a most probable value: $\hat{\theta}$. However, since the likelihood function is no longer bounded by zero and one, it is now important only *relative* to other likelihood functions based on differing values of θ. Note that the prior, $p(\theta)$, is essentially ignored here rather than overtly addressed. This is equivalent to assigning a uniform prior in a Bayesian context, an observation that has led some to consider classical inference to be a special case of Bayesianism: "everybody is a Bayesian; some know it."

Interest is generally in obtaining the *maximum likelihood* estimate of θ. The value of the unconstrained and unknown parameter, θ, which provides the maximum value of the likelihood function, $L(\theta|\mathbf{X})$. This value of θ, denoted $\hat{\theta}$, is the most likely to have generated the data given H_0 expressed through a specific parametric form relative to other possible values in the sample space of θ.

Bayesian Theory

The Bayesian approach addresses the inverse probability problem by making distributional assumptions about the unconditional distribution of the parameter, θ,

prior to observing the data, $\mathbf{X}$: $p(\theta)$. The prior and likelihood are joined with Bayes' Law:

$$\pi(\theta|\mathbf{X}) = \frac{p(\theta)L(\theta|\mathbf{X})}{\int_\Theta p(\theta)L(\theta|\mathbf{X})d\theta}, \tag{3}$$

where $\int_\Theta p(\theta)L(\theta|\mathbf{X})d\theta = p(\mathbf{X})$. Here the $\pi()$ notation is used to distinguish the posterior distribution for θ from the prior. The term in the denominator is generally not important in making inferences and can be recovered later by integration. This term is typically called the *normalizing constant*, the *normalizing factor*, or the *prior predictive distribution*, although it is actually just the marginal distribution of the data, and ensures that $\pi(\theta|\mathbf{X})$ integrates to one.

A more compact and useful form of (3) is developed by dropping this denominator and using proportional notation since $p(\mathbf{X})$ does not depend on θ and therefore provides no relative inferential information about more likely values of θ:

$$\pi(\theta|\mathbf{X}) \propto p(\theta)L(\theta|\mathbf{X}), \tag{4}$$

meaning that the unnormalized *posterior* (sampling) distribution of the parameter of interest is proportional to the prior distribution times the likelihood function.

The maximum likelihood estimate is equal to the Bayesian posterior mode with the appropriate uniform prior, and they are asymptotically equal given *any* prior: both are normally distributed in the limit. In many cases the choice of a prior is not especially important since as the sample size increases, the likelihood progressively dominates the prior. While the Bayesian assignment of a prior distribution for the unknown parameters can be seen as subjective (though *all* statistical models are actually subjective), there are often strong arguments for particular forms of the prior: little or vague knowledge often justifies a diffuse or even uniform prior, certain

probability models logically lead to particular forms of the prior (conjugacy), and the prior allows researchers to include additional information collected outside the current study.

Summarizing Bayesian Results

Bayesian researchers are generally not concerned with just getting a specific point estimate of the parameter of interest, θ, as a way of providing empirical evidence in probability distributions. Rather the focus is on describing the shape and characteristics of the posterior distribution of θ. Such descriptions are typically in the form of direct probability intervals (credible sets and highest posterior density regions), quantiles of the posterior, and probabilities of interest such as $p(\theta_i < 0)$.

Hypothesis testing can also be performed in the Bayesian setup. Suppose Θ_1 and Θ_2 represent two competing hypotheses about the state of some unknown parameter, θ, and which form a partition of the sample space: $\Theta = \Theta_1 \cup \Theta_2, \Theta_1 \cap \Theta_2 = \phi$. Prior probabilities are assigned to each of the two outcomes: $\pi_1 = p(\theta \in \Theta_1)$ and $\pi_2 = p(\theta \in \Theta_2)$. This leads to competing posterior distributions from the two priors and the likelihood function: $p_1 = p(\theta \in \Theta_1|\mathbf{X})$ and $p_2 = p(\theta \in \Theta_2|\mathbf{X})$. It is common to define the prior odds, π_1/π_2, and the posterior odds, p_1/p_2, as evidence for H_1 versus H_2. A much more useful quantity, however, is $(\pi_1/\pi_2)/(p_1/p_2)$ which is called the *Bayes Factor*. The Bayes Factor is usually interpreted as odds favoring H_1 versus H_2 given the observed data. For this reason it leads naturally to the Bayesian analog of hypothesis testing.

An Example

Suppose that $x_1, x_2, \ldots, x_n$ are observed independent random variables, all produced by the same Bernoulli probability mass function with parameter θ so that their sum, $y = \sum_{i=1}^{n} x_i$, is distributed binomial(n, θ) with known n and unknown θ. Assign a beta(A,B) prior distribution for θ, $p(\theta|A, B) = \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)}\theta^{A-1}(1-\theta)^{B-1}, \quad 0 < \theta < 1, 0 < A, B$, with researcher-specified values of A and B.

The posterior distribution for θ is obtained by the use of Bayes' Law:

$$\pi(\theta|y) = \frac{p(\theta|A, B)p(y|\theta)}{p(y)}, \tag{5}$$

where $p(y|\theta)$ is the probability mass function for y. The numerator of the right-hand-side is easy to calculate now that there are assumed parametric forms:

$$\begin{aligned} p(y, \theta) &= p(y|\theta)p(\theta) \\ &= \left[\binom{n}{y}\theta^y(1-\theta)^{n-y}\right] \quad \times \quad \left[\frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)}\theta^{A-1}(1-\theta)^{B-1}\right] \\ &= \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)}\theta^{y+A-1}(1-\theta)^{n-y+B-1}, \end{aligned} \tag{6}$$

and the denominator can be obtained by integrating θ out of this joint distribution:

$$\begin{aligned} p(y) &= \int_0^1 \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)}\theta^{y+A-1}(1-\theta)^{n-y+B-1}dp \\ &= \frac{\Gamma(n+1)\Gamma(A+B)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(A)\Gamma(B)}\frac{\Gamma(y+A)\Gamma(n-y+B)}{\Gamma(n+A+B)}. \end{aligned} \tag{7}$$

For more technical details see Gill (2002, p.67). Performing arithmetic the operations in (5) provides:

$$\begin{aligned} \pi(\theta|y) &= \frac{p(\theta)p(y|\theta)}{p(y)} \\ &= \frac{\Gamma(n+A+B)}{\Gamma(y+A)\Gamma(n-y+B)}\theta^{(y+A)-1}(1-\theta)^{(n-y+B)-1}. \end{aligned} \tag{8}$$

This is a new beta distribution for θ, with parameters $A' = y + A$ and $B' = n - y + B$, based on updating the prior distribution with the data y. Since the prior and the posterior have the same distributional family form, this an interesting special case and it is termed a *conjugate model* (the beta distribution is conjugate for the binomial probability mass function). Conjugate models are simple and elegant, but it is typically the case that realistic Bayesian specifications are much more complex analytically and may even be impossible.

Markov Chain Monte Carlo

A persistent and troubling problem for those developing Bayesian models in the twentieth century was that it was often possible to get a sufficiently complicated posterior from multiplying the prior and the likelihood, that the mathematical form existed but quantities of interest such as means and quantiles could not be calculated analytically. This problem pushed Bayesian methods to the side of mainstream statistics for quite some time. What changed this unfortunate state of affairs was the publication of a review essay by Gelfand and Smith in 1990 that described how similar problems had been solved in statistical physics with Markov chain simulation techniques. Essentially these are iterative techniques where the generated values are (eventually) from the posterior of interest (Gill 2002). Thus the difficult posterior form can be described empirically using a large number of simulated values, thus performing difficult integral calculations through simulation. The result of this development was a flood of papers that solved a large number of unresolved problems in Bayesian statistics, and the resulting effect on Bayesian statistics can be easily described as revolutionary.

Markov chains are composed of successive quantities that depend probabilistically only on the value of their immediate predecessor. In general, it is possible to set up a chain to estimate multidimensional probability structures, the desired posterior distributions, by starting a Markov chain in the appropriate sample space and letting it run until it settles into the desired target distribution. Then when it runs for some time confined to this particular distribution, it is possible to collect summary statistics such as means, variances, and quantiles from the simulated values. The two most common procedures are the Metropolis-Hastings Algorithm and the Gibbs sampler, which have been shown to possess desirable theoretical properties that lead to empirical samples from the target distribution. These methods are reasonably straightforward to implement and are becoming increasingly popular in the social sciences.

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* New York: John Wiley & Sons.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* New York: Wiley & Sons.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85, 389-409.

Gill, Jeff. (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach.* New York: Chapman and Hall.

Robert, C. P. (2001). *The Bayesian Choice: A Decision Theoretic Motivation.* Second Edition. New York: Springer-Verlag.