Bayesian Statistics

Contribution to the International Encyclopedia of Social and Behavioral Science (2nd ed.)

Hal S. Stern
Department of Statistics
University of California, Irvine
2216 Bren Hall
Irvine, CA 92697

Phone: 949-824-1568
Email: sternh@uci.edu

Keywords: probability, prior distribution, posterior distribution, inference, Monte Carlo, simulation, model checking, model assessment, subjective, objective

Abstract:
Bayesian statistics is an approach to statistical inference that is characterized by the use of probability distributions to describe the state of knowledge about unknown quantities and the use of Bayes' theorem to update the state of knowledge to account for observed information. The basic ideas date to the late 18th century but broad application of the approach did not occur until computational developments in the late 20th century allowed for simulation-based inference in a wide range of scientific applications. This survey describes the key steps in carrying out a Bayesian analysis: model specification, calculation of the posterior distribution, and model assessment.

Bayesian Statistics

Bayesian statistics refers to an approach to statistical inference that is characterized by the use of probability distributions to describe the state of knowledge or belief about unknown quantities (e.g., parameters in a statistical model) and the updating of such distributions based on observed information (e.g., data). The approach takes its name from the use of Bayes' theorem (Bayes and Price, 1763), a mathematical result relating conditional probabilities, to perform the updating of probability distributions. As commonly practiced there are three key steps in a Bayesian analysis: specification of a probability model for the quantities of interest, calculation of the posterior distribution of unknown quantities conditional on observed quantities, and assessment of the appropriateness of the model for the questions at hand. These are discussed in detail below. The last of these topics, model assessment, is not always incorporated in introductory discussions of Bayesian statistics but we argue below that it is an essential part of good statistical practice. Naturally the discussion of the Bayesian approach here is brief. Additional details can be found in a wide range of modern statistical treatments including Carlin and Louis (2008), Christensen et al. (2010) and Gelman et al. (2013).

We begin with a brief history of the development of Bayesian ideas. Early work on probability focused on what might be called "pre-data" probability calculations. Given assumptions about a random process it is possible to assign or calculate the probability of various outcomes. For example, one can calculate the probability of obtaining an even number in a roll of a die or the probability of obtaining ten or more heads in twenty tosses of a coin. Bayes and Laplace receive independent credit for "inverting" probability statements in the late 18th century to make probability statements about parameter values given observed data values (Stigler, 1986). Thus, for example, having observed a given number of successes in a series of repeated trials it is possible to make probability statements about the unknown underlying probability of success in the trials. Modern Bayesian inference developed in the period around and after World War II. The name Bayesian inference replaces "inverse probability" only in this era. Savage (1954) is an influential early book using decision theory to justify Bayesian methods. de Finetti (1974) contributed crucial work concerning the role of exchangeability which plays a role analogous to that of independent identically distributed observations in the traditional frequentist approach to inference. Raiffa and Schlaifer (1960), Lindley (1972, 1990) and Box and Tiao (1973) contributed greatly to the popularization of the Bayesian approach. Computational advances in the late 20th century, especially the development of fast, ubiquitous computers and the discovery (and rediscovery) of computational algorithms, made it possible to apply the Bayesian approach to a wide array of scientific fields and scientific problems. Indeed one can now find Bayesian treatments for a variety of different application areas including the social sciences (Jackman, 2009), econometrics (Koop et al., 2007), marketing (Rossi et al., 2007), and biostatistics (Lesaffre and Lawson, 2012).

Model Specification

The basic ideas needed to describe Bayesian statistics can be easily explained in terms of some relatively simple notation. Let *y* denote data that is to be collected, such as a series of measurements or the outcomes of a series of test questions. Take θ to denote unknown quantities associated with the collection of the data. Often θ represents the parameters in a probability

model for the data *y*. This could include the mean measurement of a population of interest or the true ability of an individual test taker. There are of course many situations that are not adequately represented by this simple setup. In some cases, like regression analysis, there may be data *x* (often called covariates) that can help explain *y* but are not of explicit interest themselves. The unknown θ may include missing values (data that we intended to collect but could not) or latent variables (additional unknowns introduced in developing a probability model) in addition to standard parameters. Here we use the simple notation *y* and θ throughout but it is straightforward to accommodate more complex situations like those described above. For details see the references provided earlier.

The specification of a complete probability model for the data *y* requires two probability distributions, a probability distribution for the data given (or conditional on) parameters θ and a probability distribution for the parameters θ. The distribution of the data conditional on the parameters is usually denoted $p(y \mid \theta)$ and identified as the data distribution, the data model, the sampling distribution, or the likelihood function. The distribution for the parameters θ, denoted $p(\theta)$, is known as the prior distribution. It can be thought of as describing the relative frequencies for different values of the model parameters before considering any of the data values. The notation above violates standard conventions in that the same letter p is used to represent two different probability distributions. It is useful for our purposes however in that it avoids the introduction of other letters that are not central to our discussion. The notation above is simple but is capable of accommodating the full range of Bayesian methods. For example, it is sometimes desirable to avoid specific parametric forms for one of both of the component distributions. The development of so-called non-parametric Bayesian methods has made this possible; in practice they are still of the form $p(y \mid \theta)$ and/or $p(\theta)$ but rely on high-dimensional θ to avoid being overly restrictive in the choice of the distribution p(.). See the entry on Non-parametric Bayes Methods for more details.

A simple example helps to show how the basic notation works. Suppose that a student is asked to answer a number of mathematics questions believed to measure their knowledge of algebra. Assuming there are *n* questions believed to be of roughly the same level of difficulty then it may be natural to let the number of correct responses *y* follow a binomial distribution with parameter θ measuring the probability that the student answers a question correctly (which can be taken as a measure of their knowledge of algebra). Then $p(y \mid \theta)$ is the binomial probability of getting exactly *y* correct responses, $\binom{n}{y} \theta^{y}(1-\theta)^{n-y}$. The prior distribution $p(\theta)$ should reflect any information about θ that is available prior to collecting the data. We discuss this choice further next.

For many users of statistical methods the largest question about the use of Bayesian methods concerns the choice of prior distribution. There are a number of approaches to this choice. The subjective approach to choosing a prior distribution is to take the prior distribution to be an honest assessment of prior beliefs about the values of the model parameters. Although people are sometimes hesitant to supply such subjective prior distributions it is often the case that some prior information is available. For our example, we may not know a great deal about the specific student but previous experience may suggest that typically students in the class are able to answer between 60 and 90 percent of such questions correctly. This may be enough to suggest a

prior distribution such as a beta distribution (appropriate for quantities that take values between zero and one) with parameters 3 and 1, which takes the form $p(\theta) = 3\theta^2$. This beta distribution has a mean of .75 with about half of the probability in the interval from .60 to .90. If it is possible to specify a prior distribution, like the beta distribution suggested here, then the Bayesian paradigm provides the way to update that prior information given observed data. The updating is discussed below.

Occasionally the choice of a prior distribution is made primarily because of considerations of computational convenience. Conjugate families are families of prior distributions that combine with a given data distribution to produce posterior distributions in the same family as the prior distribution. The beta distribution developed above is an example; it is the conjugate prior distribution for binomial data. Conjugate prior distributions are capable of supporting a variety of prior opinions (e.g., by making different choices within the conjugate family) but of course not all.

The desire to avoid using subjective prior information and/or arbitrary distributional forms has led to research on using vague or “non-informative" prior distributions. Formally a vague prior distribution is one that assigns roughly equal probability to a wide range of possible values. In the binomial example we might assert that we have no information at all about θ other than that it is between zero and one and thus choose a uniform distribution between zero and one. (In fact this is a particular form of the beta distribution.) The idea is that a vague prior distribution will not have a strong influence on our results and hence the data will play the central role in determining our conclusions. In this way, the use of vague or “non-informative" prior distributions represents the pursuit of a form of objective Bayesian inference. One difficulty is that, in the limit, especially for parameters that are not known to lie in a fixed range, vague prior distributions may become so vague as to no longer be proper distributions (they may not integrate to one!). It is permissible to use such improper prior distributions as long as it is verified that the resulting posterior distribution is a proper distribution. Improper prior distributions are popular because sometimes they appear to be “non-informative" in the sense that they rely only on the likelihood. It is probably best to think of improper prior distributions as approximations to real prior distributions. If the improper prior distribution leads to a proper posterior distribution and sufficiently accurate conclusions, then one might use the results from this analysis without needing to work any harder to select a more appropriate prior distribution.

Some people find the need to explicitly choose a prior distribution problematic. For these individuals it is disturbing that individuals with different prior distributions will obtain different results. There are a number of important points to make here. First, it is important to recognize that situations like this occur every day in the real world when individuals with different information make different decisions from the same data. Second, it is possible to show mathematically that the prior distribution does not have a large impact on the results of our statistical analysis when we have collected a great deal of data. This again matches every day experiences where a large group of people may all come to agree on a course of action given a large amount of evidence. Perhaps the most important point to make is that prior information is not necessarily as rare as we might think. This is especially true in problems that are more sophisticated than our simple math test. Consider a situation in which the same set of algebra questions are given to a number of different individuals. Each individual has his/her own

probability of getting a correct answer. Though it may not be easy to select a prior distribution for each individual, it may be natural to treat the set of parameters as a sample from a common population. This type of model is known as a hierarchical model (reference to Hierarchical Models: Random and Fixed Effects) because the prior distribution itself may depend on additional parameters (e.g., the mean and variance of algebra knowledge parameters in the population). It is the success of Bayesian methods in multiparameter problems of this type that have led to the increased use of such methods.

Calculation of the posterior distribution

The model specification makes it clear that there are two sources of information about θ. There is information about θ in the prior distribution and there is information about θ in the data distribution. This information can be combined using the using the laws of probability, namely Bayes' theorem, to obtain the updated probability distribution on θ after observing data *y*. The latter is known as the posterior distribution of θ and denoted $p(\theta \mid y)$. Bayes' theorem is

$$p(\theta \mid y) = p(y \mid \theta)\, p(\theta) / p(y) = p(y \mid \theta)\, p(\theta) / \int p(y \mid \theta)\, p(\theta)\, d\theta.$$

The denominator of Bayes' theorem is known as the marginal probability of observing the data *y*. Bayes' theorem allows us update our opinion about the possible values of θ so that values of θ that are most consistent with the observed data (i.e., values that have large values of the likelihood) have more weight in the posterior distribution then they do in the prior distribution. The posterior distribution is fundamental to Bayesian inference for θ. The Bayesian approach makes probability statements concerning the unknown quantities after fixing the things that have been observed (i.e., the data) at their known values. Note that Bayesian methods differ from traditional methods which are justified by repeated sampling considerations -- only the sample at hand is considered relevant in the Bayesian approach.

The posterior distribution describes our current state of knowledge about θ after observing the data *y*. It is possible to draw a wide range of inferential conclusions from the posterior distribution. A point estimate is a one-number summary of the posterior distribution. Of course, a single estimate does not describe the full state of our knowledge. To obtain a point estimate it is necessary to specify a loss function that makes explicit the loss that we expect to incur because of errors in our point estimate. Given a loss function it is possible to derive the optimal point estimate as the value that minimizes the expected loss (computed from the posterior distribution) (e.g., see the entry on Bayesian Decision Theory). An interval estimate specifies a range of plausible values for a parameter. The posterior distribution allows us to identify posterior intervals that contain a given parameter with any specified probability. Such intervals are analogous to standard confidence intervals. One difference is that the posterior intervals have a natural probabilistic interpretation in that they summarize our posterior state of knowledge regarding plausible values of the parameter. By contrast, the probability statement underlying traditional confidence intervals depends on behavior in a large number of repeated samples. In traditional inference another common form of statistical inference is a hypothesis test about the value of a parameter or a set of parameters (e.g., it is common to test whether two means are equal). The information about such tests is equivalent to the information contained in posterior intervals, i.e., we can address whether two means are plausibly the same by examining the

posterior interval for the difference between the two means. Tests of some hypotheses are also possible using the Bayes factor, a ratio that measures the degree to which two competing models explain the observed data. This is discussed further below.

In relatively simple problems it is possible to analytically determine the form of the posterior distribution and subsequently obtain statistical inferences. For the example described earlier in which we observed a binomial random variable measuring the number of correct responses y in n questions and used a conjugate beta prior distribution (with parameters 3 and 1), it turns out that the posterior distribution is also a beta distribution. The posterior distribution is a beta distribution with parameters $y+3$ and $n-y+1$. The form of the posterior distribution in this case shows us explicitly how our prior information combines with the data. The mean of this posterior distribution is $(y+3)/(n+4)$ which is a compromise between the student's performance on the test y/n and the mean of the prior distribution 3/4 = .75. For example, a student who answered nine out of ten questions correctly would have posterior mean .86. Posterior intervals can also be easily obtained if we know the form of the posterior distribution.

In many realistic models, especially those containing a large number of parameters, it can be quite difficult to determine the form of the posterior distribution analytically. It is sometimes possible in such cases to develop a useful approximation to the posterior distribution, e.g., a normal approximation, but this too is difficult in many realistic problems. For many years the difficulty in calculating the posterior distribution kept people from applying Bayesian methods. The wide availability of high-speed computing, beginning in the late 20th century, made it possible to develop general purpose simulation algorithms that allowed researchers to approximate the posterior distribution by generating a sample of simulations from the posterior distribution (e.g., see Computation, Monte Carlo Methods and Bayesian Computation). The most common class of algorithms are known as Markov chain Monte Carlo (MCMC) algorithms; these algorithms work by building and then simulating a Markov chain which has the posterior distribution as its stationary distribution. If the Markov chain is simulated for a long time then the draws from the simulation can be treated as draws from the posterior distribution. Though there can be a number of challenges associated with Bayesian computation (e.g., developing efficient algorithms, judging when the simulation has run long enough to approximate the posterior distribution) there is little doubt that it has allowed researchers to apply Bayesian methods in diverse settings and with large data sets. Recent references describing MCMC algorithms and other relevant computational strategies include Chen et al. (2001), Liu (2008), and Brooks et al. (2011).

Simulation-based inference is a powerful tool. A set of simulations from the posterior distribution provides an empirical approximation to the posterior distribution. Any desired inference is then estimated using this empirical approximation. For example, the posterior mean is estimated by the sample mean of the simulations and a posterior interval for a parameter is obtained using the relevant percentiles of the simulations. Although the simulations are in fact providing estimates of the various quantities of interest, it is generally possible to obtain enough simulations to reduce the Monte Carlo or simulation error to acceptable levels. Simulations provide a great deal of flexibility – for example, it is quite straightforward to estimate the posterior probability that a given parameter is positive since one needs only look at the proportion of simulations in which it is positive.

Model assessment

It is important when performing statistical inference based on a probability model to examine the validity of the assumptions made in the model. For the Bayesian approach to inference the assumptions include the form of the prior distribution, the form of the data model, and any other assumptions embodied in the data model (independence, constant variance, etc.). There are a wide range of activities that can be used to assess the Bayesian model being used to analyze data. These include sensitivity analysis, model checking and model comparison. We briefly describe approaches to this important task here. The general references provided earlier give additional details.

Sensitivity analysis: The prior distribution is an important part of the Bayesian model. One can argue that there is no need to check the prior distribution in that as long as it is an honest reflection of the prior state of knowledge about unknown quantities, then the Bayesian machinery will provide the appropriate updated posterior state of knowledge. At the same time it is important to be able to describe whether the inferences are likely to change under other reasonable prior distributions. This suggests a form of sensitivity analysis where the prior distribution (or other aspect of the model) is varied and the effect on the posterior distribution is assessed. The conclusions of a Bayesian analysis are strengthened if we find that key posterior inferences (e.g., the sign of an important parameter) are unchanged under other plausible model specifications.

Model checking: In regression analysis it is common to check the form of the model by examining residuals. The residuals are functions of the data that are expected to have no particular structure if the model is correct. Evidence of structure in the residuals suggests a possible problem with the model. The same basic strategy can be applied more generally to assess the fit of a Bayesian model or analysis. A function of the data (and perhaps parameters) can be defined that is expected not to have any structure if the model is correct. Examining the posterior predictive distribution of such functions (i.e., the distribution of the function in future data sets that would be expected if the model is correct) can allow one to detect difficulties in the model. These are known as posterior predictive model checks.

Model comparison: In some applications there are a number of models that are contemplated. In this case researchers may want to compare or combine the models. Model comparison is frequently carried out using Bayes factors (see the entry on Model Selection and Model Averaging). It can also be carried out using predictive measures (e.g., cross validation) or information criteria. There are several important points here. If model selection is used to identify a best-fitting model, it is still important to use model checking and sensitivity analysis as described above to ensure that the selected model is appropriate for the questions of interest. Also, one can argue that there are alternatives to model comparison and model selection that make more sense in the Bayesian framework. It may be possible to embed one model into a bigger family of models and then average over the entire family. For example, a normal model might be replaced by the Student-t family of models that includes the normal distribution as a limiting case. Then the posterior distribution for the degrees of freedom in the Student-t model provides information about whether the normal distribution is an appropriate choice.

Summary

The Bayesian approach to statistics outlined here is in wide use across a wide range of application areas, including the social sciences. Historically there has been controversy about the use of Bayesian methods. This remains true to a limited extent (see, e.g., Efron, 1986). The arguments favoring the use of the Bayesian approach for data analysis follow from its use of probability distributions to describe uncertainty about unknown quantities. There is a natural probability-based interpretation for Bayesian results (e.g., interval estimates) and great flexibility in the types of inferences that can be obtained (e.g., one can easily obtain a posterior distribution on the ranks of a set of parameters). In addition, the framework is extremely powerful in its ability to accommodate increasing degrees of complexity in a data analysis. For example, accommodating missing data values is straightforward in the Bayesian framework because the values that we had hoped to measure become unknown random variables. In the same vein censoring or truncation of observed values are easily incorporated through modifications of the underlying probability model. The reliance on formal probability distributions also means that it is possible to draw valid Bayesian inferences in finite samples without relying on large sample results.

References

Bayes, T, and Price, R 1763 An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London* **53**: 370–418. doi:10.1098/rstl.1763.005

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

Brooks S, Gelman A, Jones G and Meng, X-L (editors) 2012 *Handbook of Markov Chain Monte Carlo*. CRC/Chapman and Hall, London.

Carlin B P, and Louis T A 2008 *Bayes Methods for Data Analysis* (3rd edition). CRC/Chapman and Hall, London.

Chen M-H, Shao Q-M and Ibrahim J G 2001 *Monte Carlo Methods in Bayesian Computation.* Springer, New York.

Christensen R, Johnson W O, Branscum A J and Hanson T E 2010 *Bayesian Ideas and Data Analysis*. CRC/Chapman and Hall, London.

de Finetti, B 1974 *Theory of Probability*. Wiley, New York.

Efron B 1986 Why isn't everyone a Bayesian? *The American Statistician* 40:1--11 (including discussion).

Gelman A, Carlin J B, Stern H S, Dunson, D B, Vehtari, A, and Rubin D B 2013 *Bayesian Data Analysis (3rd edition)*. CRC/Chapman and Hall, London.

Jackman, S 2009 *Bayesian Analysis for the Social Sciences.* John Wiley, New York.

Koop G, Poirier D J and Tobias J L 2007 *Bayesian Econometric.* John Wiley, New York.

Lesaffre E and Lawson A 2012 *Bayesian Biostatistics.* John Wiley, New York.

Lindley D V 1972 *Bayesian Statistics, A Review*. Society for Industrial and Applied Mathematics, Philadelphia.

Lindley D V 1990 The 1988 Wald Memorial Lectures: The present position of Bayesian statistics. *Statistical Science*, 5, 44--89 (including discussion).

Liu J S 2008 *Monte Carlo Strategies in Scientific Computing (*2nd edition). Springer, New York.

Raiffa, H, and Schlaifer, R 1961 *Applied Statistical Decision Theo*ry, Harvard Business School, Boston.

Rossi P E, Allenby G M and McCulloch R 2006 *Bayesian Statistics and Marketing.* John Wiley, New York.

Savage, L J 1954 *The Foundations of Statistics*. Dover, New York

Stigler, S M 1986 *The History of Statistics*. Harvard University Press, Cambridge