Bayesian statistics

Dr. David Spiegelhalter, Cambridge University, UK

Kenneth Rice, University of Washington, Seattle, WA, USA

Bayesian statistics is a system for describing epistemological uncertainty using the mathematical language of probability. In the 'Bayesian paradigm,' degrees of belief in states of nature are specified; these are non-negative, and the total belief in all states of nature is fixed to be one. Bayesian statistical methods start with existing 'prior' beliefs, and update these using data to give 'posterior' beliefs, which may be used as the basis for inferential decisions.

Background

In 1763, Thomas Bayes published a paper on the problem of *induction*, that is, arguing from the specific to the general. In modern language and notation, Bayes wanted to use Binomial data comprising r successes out of n attempts to learn about the underlying chance θ of each attempt succeeding. Bayes' key contribution was to use a probability distribution to represent uncertainty about θ . This distribution represents 'epistemological' uncertainty, due to lack of knowledge about the world, rather than 'aleatory' probability arising from the essential unpredictability of future events, as may be familiar from games of chance.

Modern 'Bayesian statistics' is still based on formulating probability distributions to express uncertainty about unknown quantities. These can be underlying parameters of a system (induction) or future observations (prediction).

Bayes' Theorem

In its raw form, Bayes' Theorem is a result in conditional probability, stating that for two random quantities y and θ ,

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y),$$

where $p(\cdot)$ denotes a probability distribution (http://en.wikipedia.org/wiki/Probability_distribution) , and $p(\cdot|\cdot)$ a conditional distribution. When y represents data and θ represents parameters in a statistical model, Bayes Theorem provides the basis for Bayesian inference. The 'prior' distribution $p(\theta)$ (epistemological uncertainty) is combined with 'likelihood' $p(y|\theta)$ to provide a 'posterior' distribution $p(\theta|y)$ (updated epistemological uncertainty): the likelihood is derived from an aleatory sampling model $p(y|\theta)$ but considered as function of θ for fixed y .

While an innocuous theory, practical use of the Bayesian approach requires consideration of complex practical issues, including the source of the prior distribution, the choice of a likelihood function, computation and summary of the posterior distribution in high-dimensional problems, and making a convincing presentation of the analysis.

Bayes theorem can be thought of as way of *coherently* updating our uncertainty in the light of new evidence. The use of a probability distribution as a 'language' to express our uncertainty is not an arbitrary choice: it can in fact be determined from deeper principles of logical reasoning or rational behavior; see Jaynes (2003) or Lindley (1953). In particular, De Finetti (1937) showed that making a qualitative assumptions of *exchangeability* of binary observations (i.e. that their joint distribution is unaffected by label-permutation) is equivalent to assuming they are each independent conditional on some unknown parameter θ , where θ has a prior distribution and is the limiting frequency with which the events occur.

Use of Bayes' Theorem: a simple example

Suppose a hospital has around 200 beds occupied each day, and we want to know the underlying risk that a patient will be infected by MRSA (methicillin-resistant *Staphylococcus aureus*). Looking back at the first six months of the year, we count $y = 20$ infections in 40,000 bed-days. A simple estimate of the underlying risk θ would be 20/40,000 $=$ 5 infections per 10,000 bed-days. This is also the maximum-likelihood estimate, if we assume that the observation y is drawn from a Poisson distribution with mean θN where $N = 4$ is the number of bed-days/ $10,000$, so that

$$p(y|\theta) = (\theta N)^y e^{-\theta N}/y! \ .$$

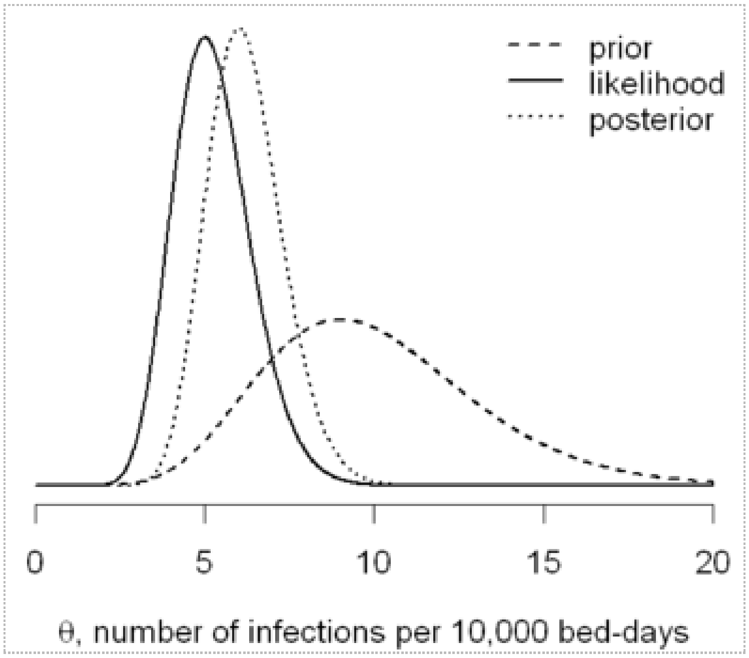

Figure 1: Prior, likelihood and posterior distributions for θ , the rate of infections per 10,000 bed-days. The posterior distribution is a formal compromise between the likelihood, summarizing the evidence in the data alone, and the prior distribution, which summarizes external evidence which suggested higher rates.

However, other evidence about the underlying risk may exist, such as the previous year's rates or rates in similar hospitals which may be included as part of a hierarchical model (see below). Suppose this other information, on its own, suggests plausible values of θ of around 10 per 10,000, with 95% of the support for θ lying between 5 and 17. This judgement about θ may be expressed as a prior probability distribution. Say, for convenience, the Gamma(a, b) family of distributions is chosen to formally describe our knowledge about θ . This family has density

$$p(\theta) = b^a \theta^{a-1} e^{-b\theta}/\Gamma(a) \ ;$$

choosing $a = 10$ and $b = 1$ gives a prior distribution with appropriate properties, as shown in Figure 1.

Figure 1 also shows a density proportional to the likelihood function, under an assumed Poisson model. Using Bayes Theorem, the posterior distribution $p(\theta|y)$ is

$$\propto \theta^y e^{-\theta N} \theta^{a-1} e^{-b\theta} \propto \theta^{y+a-1} e^{-\theta(N+b)} \ ,$$

i.e. a Gamma$(y + a, N + b)$ distribution - this closed-form posterior, within the same parametric family as the prior, is an example of a *conjugate* Bayesian analysis. Figure 1 shows that this posterior is primarily influenced by the likelihood function but is 'shrunk' towards the prior distribution to reflect that the expectation based on external evidence was of a higher rate than that actually observed. This can be thought of as an automatic adjustment for 'Regression to the mean (http://en.wikipedia.org/wiki/Regression_to_the_mean%7C) ', in that the prior distribution will tend to counteract chance highs or lows in the data.

Prior distributions

The prior distribution is central to Bayesian statistics and yet remains controversial unless there is a physical sampling mechanism to justify a choice of $p(\theta)$. One option is to seek 'objective' prior distributions that can be used in situations where judgemental input is supposed to be minimized, such as in scientific publications. While progress in Objective Bayes methods has been made for simple situations, a universal theory of priors that represent zero or minimal information has been elusive.

A complete alternative is the fully subjectivist position, which compels one to elicit priors on all parameters based on the personal judgement of appropriate individuals. A pragmatic compromise recognizes that Bayesian statistical analyses must usually be justified to external bodies and therefore the prior distribution should, as far as possible, be based on convincing external evidence or at least be guaranteed to be weakly informative: of course, exactly the same holds for the choice of functional form for the sampling distribution which will also be a subject of judgement and will need to be justified. Bayesian analysis is perhaps best seen as a process for obtaining posterior distributions or predictions based on a range of assumptions about both prior distributions and likelihoods: arguing in this way, sensitivity analysis and reasoned justification for both prior and likelihood become vital.

Sets of prior distributions can themselves share unknown parameters, forming *hierarchical* models (http://en.wikipedia.org/wiki/Hierarchical_models%7C) . These feature strongly within applied Bayesian analysis and provide a powerful basis for pooling evidence from multiple sources in order to reach more precise conclusions. Essentially a compromise is reached between the two extremes of assuming the sources are estimating (a) precisely the same, or (b) totally unrelated, parameters. The degree of pooling is itself estimated from the data according to the similarity of the sources, but this does not avoid the need for careful judgement about whether the sources are indeed exchangeable, in the sense that we have no external reasons to believe that certain sources are systematically different from others.

Prediction

One of the strengths of the Bayesian paradigm is its ease in making predictions. If current uncertainty about θ is summarized by a posterior distribution $p(\theta|y)$, a *predictive distribution* for any quantity z that depends on θ through a sampling distribution $p(z|\theta)$ can be obtained as follows;

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

provided that y and z are conditionally independent given θ , which will generally hold except in time series or spatial models.

In the MRSA example above, suppose we wanted to predict the number of infections z over the next six months, or 40,000 bed-days. This prediction is given by

$$p(z|y) = \int \frac{(\theta N)^z\, e^{-\theta N}}{z!}\quad \frac{(N+b)^{y+a}\,\theta^{y+a-1}\,e^{-\theta(N+b)}}{\Gamma(y+a)}\quad d\theta = \frac{\Gamma(z+y+a)}{\Gamma(y+a)z!}\, p^{y+a}(1-p)^z\ ,$$

where $p = (N+b)/(2N+b)$. This Negative Binomial *predictive distribution* for z is shown in Figure 2.

Making Bayesian Decisions

For inference, a full report of the posterior distribution is the correct and final conclusion of a statistical analysis. However, this may be impractical, particularly when the posterior is high-dimensional. Instead, posterior summaries are commonly reported, for example the posterior mean and variance, or particular tail areas. If the analysis is performed with the goal of making a specific decision, measures of utility, or *loss functions* can be used to derive the posterior summary that is the 'best' decision, given the data.

In Decision Theory, the loss function describes how bad a particular decision would be, given a true state of nature. Given a particular posterior, the Bayes rule is the decision which minimizes the expected loss with respect to that posterior. If a rule is *admissible* (meaning that there is no rule with strictly greater utility, for at least some state of nature) it can be shown to be a Bayes rule for some proper prior and utility function.

Many intuitively-reasonable summaries of posteriors can also be motivated as Bayes rules. The posterior mean for some parameter θ is the Bayes rule when the loss function is the square of the distance from θ to the decision. As noted, for example, by Schervish (1995), quantile-based credible intervals can be justified as a Bayes rule for a bivariate decision problem, and Highest Posterior Density intervals can be justified as a Bayes rule for a set-valued decision problem.

As a specific example, suppose we had to provide a point prediction for the number of MRSA cases in the next 6 months. For every case that we over-estimate, we will lose 10 units of wasted resources, but for every case that we under-estimate we will lose 50 units through having to make emergency provision. Our selected estimate is that t which will minimise the expected total cost, given by

$$\sum_{z=0}^{t-1} 10(t-z)p(z|y) + \sum_{z=t+1}^{\infty} 50(z-t)p(z|y)$$

The optimal choice of t can be calculated to be 30, considerably more than the expected value 24, reflecting our fear of under-estimation.

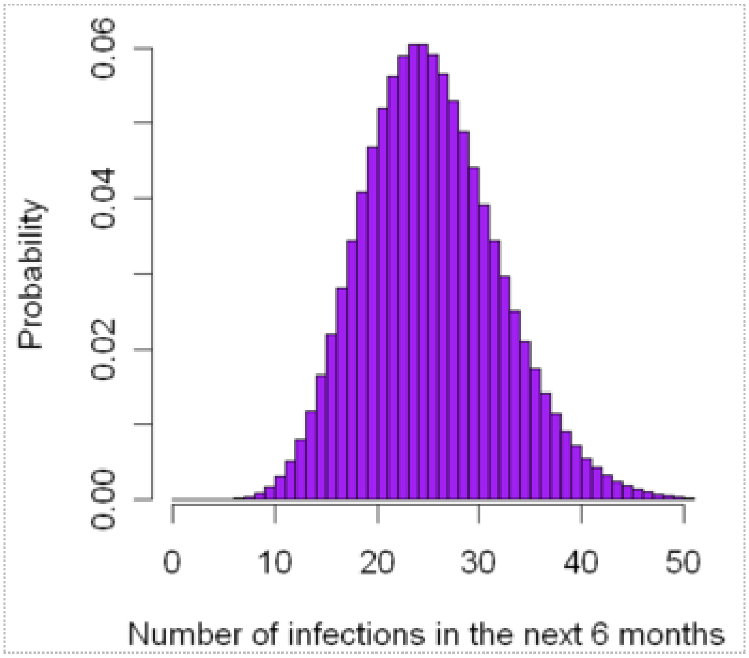

Figure 2: Predictive distribution for number of infections in the next six months, expressed as Negative Binomial$(a + y + 1, \frac{b+N}{b+2N})$ distribution with $a = 10$, $b = 1$, $y = 20$, $N = 4$. The mean is 25 and standard deviation is 6.7, and the probability that there are more than 20 infections is 73%. Essentially, more infections are predicted for the second six months, because external evidence suggests the observations were lucky in the first half of the year.

Computation for Bayesian statistics

Bayesian analysis requires evaluating expectations of functions of random quantities as a basis for inference, where these quantities may have posterior distributions which are multivariate or of complex form or often both. This meant that for many years Bayesian statistics was essentially restricted to conjugate analysis, where the mathematical form of the prior and likelihood are jointly chosen to ensure that the posterior may be evaluated with ease. Numerical integration methods based on analytic approximations or quadrature were developed in 70s and 80s with some success, but a revolutionary change occurred in the early 1990s with the adoption of indirect methods, notably Monte Carlo Markov Chain).

The Monte Carlo method

Any posterior distribution $p(\theta|y)$ may be approximated by taking a very large random sample of realizations of θ from $p(\theta|y)$; the approximate properties of $p(\theta|y)$ by the respective summaries of the realizations. For example, the posterior mean and variance of θ may be approximated by the mean and variance of a large number of realizations from $p(\theta|y)$. Similarly, quantiles of the realizations estimate quantiles of the posterior, and the mode of a smoothed histogram of the realizations may be used to estimate the posterior mode.

Samples from the posterior can be generated in several ways, without exact knowledge of $p(\theta|y)$. Direct methods include rejection sampling, which generates independent proposals for θ , and accepts them at a rate whereby those retained are proportional to the desired posterior. Importance sampling can also be used to numerically evaluate relevant integrals; by appropriately weighting independent samples from a user-chosen distribution on θ , properties of the posterior $p(\theta|y)$ can be estimated.

Markov Chain Monte Carlo (MCMC)

Realizations from the posterior used in Monte Carlo methods need not be independent, or generated directly. If the conditional distribution of each parameter is known (conditional on all other parameters), one simple way to generate a possibly-dependent sample of data points is via Gibbs Sampling. This algorithm generates one parameter at a time; as it sequentially updates each parameter, the entire parameter space is explored. It is appropriate to start from multiple starting points in order to check convergence, and in the long-run, the 'chains' of realizations produced will reflect the posterior of interest.

More general versions of the same argument include the Metropolis-Hastings algorithm; developing practical algorithms to approximate posterior distributions for complex problems remains an active area of research.

Applications of Bayesian statistical methods

Explicitly Bayesian statistical methods tend to be used in three main situations. The first is where one has no alternative but to include quantitative prior judgments, due to lack of data on some aspect of a model, or because the inadequacies of some evidence has to be acknowledged through making assumptions about the biases involved. These situations can occur when a policy decision must be made on the basis of a combination of imperfect evidence from multiple sources, an example being the encouragement of Bayesian methods by the Food and Drug Administration (FDA) division responsible for medical devices.

The second situation is with moderate-size problems with multiple sources of evidence, where hierarchical models can be constructed on the assumption of shared prior distributions whose parameters can be estimated from the data. Common application areas include meta-analysis, disease mapping, multi-centre studies, and so on. With weakly-informative prior distributions the conclusions may often be numerically similar to classic techniques, even if the interpretations may be different.

The third area concerns where a huge joint probability model is constructed, relating possibly thousands of observations and parameters, and the only feasible way of making inferences on the unknown quantities is through taking a Bayesian approach: examples include image processing, spam filtering, signal analysis, and gene expression data. Classical model-fitting fails, and MCMC or other approximate methods become essential.

There is also extensive use of Bayesian ideas of parameter uncertainty but without explicit use of Bayes theorem. If a deterministic prediction model has been constructed, but some of the parameter inputs are uncertain, then a joint prior distribution can be placed on those parameters and the resulting uncertainty propagated through the model, often using Monte Carlo methods, to produce a predictive probability distribution. This technique is used widely in risk analysis, health economic modelling and climate projections, and is sometimes known as *probabilistic sensitivity analysis*.

Another setting where the 'updating' inherent in the Bayesian approach is suitable is in machine-learning; simple examples can be found in modern software for spam filtering, suggesting which books or movies a user might enjoy given his or her past preferences, or ranking schemes for millions of on-line gamers. Formal inference may only be approximately carried out, but the Bayesian perspective allows a flexible and adaptive response to each additional item of information.

Open Areas in Bayesian Statistics

The philosophical rationale for using Bayesian methods was largely established and settled by the pioneering work of De Finetti, Savage, Jaynes and Lindley. However, widespread concern remain over how to apply these methods in practice, where various concerns over sensitivity to assumptions can detract from the rhetorical impact of Bayesians' epistemological validity.

Hypothesis testing and model choice

Jeffreys (1939) developed a procedure for using data y to test between alternative scientific hypotheses H_0 and H_1 , by computing the *Bayes factor* $p(y|H_0)/p(y|H_1)$. He suggested thresholds for strength of evidence for or against the hypotheses. The Bayes factor can be combined with the prior odds $p(H_0)/p(H_1)$ to give posterior probabilities of each hypothesis, that can be used to weight predictions in Bayesian Model Averaging (BMA). Although BMA can be an effective pragmatic device for prediction, the use of posterior model probabilities for scientific hypothesis-testing is controversial even among the Bayesian community, for both philosophical and practical reasons: first, it may not make sense to talk of probabilities of hypotheses that we know are not strictly 'true', and second, the calculation of the Bayes factor can be extremely sensitive to apparently innocuous prior assumptions about parameters within each hypothesis. For example, the ordinate of a widely dispersed uniform prior distribution would be irrelevant for estimation within a single model, but becomes crucial when comparing models.

It has also been argued that model choice is not necessarily the same as identifying the 'true' model, particularly as in most circumstances no true model exists and so posterior model probabilities are not interpretable or useful. Instead, other criteria, such as the Akaike Information Criterion or the Deviance Information Criterion, are concerned with selecting models that are expected to make good short-term predictions.

Robustness and reporting

In the uncommon situation that the data are extensive and of simple structure, the prior assumptions will be unimportant and the assumed sampling model will be uncontroversial. More generally we would like to report that any conclusions are robust to reasonable changes in both prior and assumed model: this has been termed *inference robustness* to distinguish it from the frequentist idea of robustness of procedures when applied to different data. (Frequentist statistics uses the properties of statistical procedures over repeated applications to make inference based on the data at hand)

Bayesian statistical analysis can be complex to carry out, and explicitly includes both qualitative and quantitative judgement. This suggests the need for agreed standards for analysis and reporting, but these have not yet been developed. In particular, audiences should ideally fully understand the contribution of the prior distribution to the conclusions, the reasonableness of the prior assumptions, the robustness to alternative models and priors, and the adequacy of the computational methods.

Model criticism

In the archetypal Bayesian paradigm there is no need for testing whether a single model adequately fits the data, since we should be always comparing two competing models using hypothesis-testing methods. However there has been recent growth in techniques for testing absolute adequacy, generally involving the simulation of replicate data and checking whether specific characteristics of the observed data match those of the replicates. Procedures for model criticism in complex hierarchical models are still being developed. It is also reasonable to check there is not strong conflict between different data sources or between prior and data, and general measures of conflict in complex models is also a subject of current research.

Connections and comparisons with other schools of statistical inference

At a simple level, 'classical' likelihood-based inference closely resembles Bayesian inference using a flat prior, making the posterior and likelihood proportional. However, this underestimates the deep philosophical differences between Bayesian and frequentist inference; Bayesian make statements about the relative evidence for parameter values given a dataset, while frequentists compare the relative chance of datasets given a parameter value.

The incompatibility of these two views has long been a source of contention between different schools of statisticians; there is little agreement over which is 'right', 'most appropriate' or even 'most useful'. Nevertheless, in many cases, estimates, intervals, and other decisions will be extremely similar for Bayesian and frequentist analyses. Bernstein von Mises Theorems give general results proving approximate large-sample agreement between Bayesian and frequentist methods, for large classes of standard parametric and semi-parametric models. A notable exception is in hypothesis testing, where default Bayesian and frequentist methods can give strongly discordant conclusions. Also, establishing Bayesian interpretations of non-model based frequentist analyses (such as Generalized Estimating Equations) remains an open area.

Some qualities sought in non-Bayesian inference (such as adherence to the likelihood principle (http://en.wikipedia.org/wiki/Likelihood_principle%7C) and exploitation of sufficiency) are natural consequences of following a Bayesian approach. Also, many Bayesian procedures can also, quite straightforwardly, be calibrated to have desired frequentist properties, such as intervals with 95% coverage. This can be useful when justifying Bayesian methods to external bodies such as regulatory agencies, and we might expect an increased use of 'hybrid' techniques in which a Bayesian interpretation is given to the inferences, but the long-run behaviour of the procedure is also taken into account.

References

- Thomas Bayes (1763), "An Essay towards solving a Problem in the Doctrine of Chances" Phil. Trans. Royal Society London
- B. de Finetti, La Prevision: Ses Lois Logiques, Ses Sources Subjectives (1937) Annales de l'Institut Henri Poincare, 7: 1-68. Translated as Foresight: Its Logical Laws, Its Subjective Sources, in Kyburg, H. E. and Smokler, H. E. eds., (1964). Studies in Subjective Probability. Wiley, New York, 91-158
- E.T. Jaynes Probability Theory: The Logic of Science (2003) Cambridge University Press, Cambridge, UK
- H. Jeffreys (1939) Theory of Probability Oxford, Clarendon Press
- D.V. Lindley: Statistical Inference (1953) Journal of the Royal Statistical Society, Series B, 16: 30-76
- Schervish, M. J. (1995) Theory of Statistics. Springer-Verlag, New York.

Further reading

- Bernardo and Smith (1994) Bayesian Theory, Wiley
- Berger (1993) Statistical Decision Theory and Bayesian Analysis, Springer-Verlag
- Carlin and Louis (2008) Bayesian Methods for Data Analysis (Third Edition) Chapman and Hall/CRC
- Gelman, Carlin, Stern and Rubin (2003) Bayesian Data Analysis (Second Edition) Chapman and Hall/CRC
- Gelman and Hill (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press
- Lindley (1991) Making Decisions (2nd Edition) Wiley
- Robert (2007) The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (Second Edition), Springer-Verlag

See also

Sponsored by: Eugene M. Izhikevich, Editor-in-Chief of Scholarpedia, the peer-reviewed open-access encyclopedia

Sponsored by: Dr. Michael Hardy, University of Minnesota, School of Mathematics

Reviewed by (http://www.scholarpedia.org/w/index.php?title=Bayesian_statistics&oldid=60276) : Anonymous

Reviewed by (http://www.scholarpedia.org/w/index.php?title=Bayesian_statistics&oldid=60276) : Dr. Andrew Gelman, University of Columbia, New York, NY

Accepted on: 2009-03-25 00:15:45 GMT (http://www.scholarpedia.org/w/index.php?title=Bayesian_statistics&oldid=60276)

Categories: Statistics | Multiple Curators

This page was last modified on 21 October 2011, at 04:05.

This page has been accessed 102,344 times.

Served in 0.465 secs.