Biostats VHM 801/802 Courses
Fall 2005, Atlantic Veterinary College, PEI
Henrik Stryhn

Introduction to Bayesian Analysis [1]

Little known outside the statistical science, there exist two different approaches to statistical inference, which have different concepts and philosophical bases and will in general lead to different results. The rivalry between the two schools has persisted over decades, without neither of them emerging as the clear winner. Many statisticians hold the intermediate viewpoint that each of the two approaches has its weaknesses and strengths which make them attractive in particular situations. However, most (introductory) statistics courses are taught within the non-Bayesian (classical, likelihood-based, frequentist) framework with no reference to the Bayesian view.

Bayesian analysis is gaining in popularity in recent years, and have for example been applied to complex problems in veterinary epidemiology like risk assessment or comparison of diagnostic tests without a gold standard. The scope of practical Bayesian inference has been increased widely by the invention and recent advances of a simulation-based tool for statistical inference: Markov chain Monte Carlo (MCMC) estimation.

We hope the reader will bear with us for the inevitable inadequacy of a few pages' introduction to a full, new statistical approach. Our aim can be no more than giving a superficial impression of the ideas and steps involved in a Bayesian analysis. Recent textbooks on applied Bayesian analysis in the health and biological sciences (*eg* Congdon P. (2001), *Bayesian Statistical Modelling*, and Congdon P. (2003), *Applied Bayesian Modelling*) would be the proper starting point. Most Bayesian analyses require specialized software, and the standard choice is the (free) BUGS programme developed by the Medical Biostatistics Unit in Cambridge (http://www.mrc-bsu.cam.ac.uk/bugs/). BUGS is short for Bayesian analysis using Gibbs sampling, which is a particular type of MCMC analysis.

1. Bayesian paradigm

Bayesian methodology owes its name to the fundamental role of Bayes' theorem, see equation (1) below. In Bayesian reasoning, uncertainty is attributed not only to data but also to the parameters. Therefore, all parameters are modelled by distributions. Before any data are obtained, the knowledge about the parameters of a problem are expressed in the **prior** distribution of the parameters. Given actual data, the prior distribution and the data are combined into the **posterior** distribution of the parameters. The posterior distribution summarizes our knowledge about the parameters after observing the data. The major differences between classical and Bayesian inference are outlined in Table 1, and will be detailed in the following sections.

[1] Third and slightly expanded version of notes adapted (in part) from Chapter 23 in Dohoo IR, Martin SW & Stryhn H (2003), *Veterinary Epidemiologic Research*.

Concept	Classical approach	Bayesian approach
parameter	constant	distribution
prior information on parameters	none	prior distribution
base of inference	likelihood function	posterior distribution
parameter value	(ML) estimate	statistic of posterior, *eg* median, mode, mean
parameter range	confidence interval	probability range of posterior distribution
hypothesis statement	test	testing by confidence interval (Bayesian factors)

Table 1: Bayesian *vs* classical ('frequentist', likelihood-based) approaches to statistics.

Let us briefly indicate the way the prior and the data are merged, and denote by Y the data, by θ the parameter (vector), and

- $L(Y|\theta)$ — the likelihood function, giving the probability or density of the observed data Y when the parameter takes value θ,
- $f(\theta)$ — the prior distribution for θ,
- $f(\theta|Y)$ — the posterior distribution for θ,

where the $f(\cdot)$'s are either probability functions (discrete data) or probability densities (continuous data). With these definitions Bayes' theorem says,

$$f(\theta|Y) = \mathit{const}(Y)\, L(Y|\theta)\, f(\theta), \tag{1}$$

where $\mathit{const}(Y)$ is a constant depending on Y but not on θ. Thus, the posterior distribution for θ is essentially constructed by multiplying the likelihood and the prior, and is a sort of compromise between the two. In complex models, the constant depending on Y in (1) is virtually impossible to calculate; therefore, simulation methods like MCMC have had a great impact on Bayesian analysis.

2. Statistical analysis using the posterior distribution

Even if it may seem awkward to discuss the posterior before the prior distribution, let us see a simple example of a Bayesian analysis before turning to the discussion of how to choose the prior distribution. The net result of a Bayesian analysis is a **distribution**, and the analysis may therefore very conveniently be summarized by of a graph of the posterior distribution, as shown in the example below. Point estimates and confidence intervals are not truly Bayesian in spirit, but values such as the mean, median or mode, and intervals comprising a certain probability mass of the posterior (sometimes called credibility intervals) may be calculated from the posterior distribution. The two most commonly used point values are median and mode, the latter also called a MAP (maximum aposteriori) estimate.

Example: Bayesian analysis of proportions
Assume that we test 10 animals for a disease with highly variable prevalence. In one scenario 5 of the animals tested positive, in another scenario 8 animals tested positive. What information have we obtained about the disease prevalence in these two scenarios?

Recall that all Bayesian analyses involved a prior distribution, in this case for the disease prevalence p. Assume (somewhat unrealistically) that we had no particular prior information (due to the high variability of the disease) so that apriori all values of p would seem equally likely. In that case we could choose a uniform distribution on (0,1) as our prior; this is an example of a "noninformative" prior (next section). The probability density of the uniform distribution is constant (1). The likelihood function for observing the number of positive animals out of 10 are the probabilities of the binomial $(10, p)$. Therefore, if we observe Y positive animals, the posterior distribution has density

$$f(p|Y) = const(Y) \cdot p^Y(1-p)^{10-Y} \cdot 1 = const(Y)p^Y(1-p)^{10-Y}.$$

This probability density correponds to a beta-distribution with parameters $(Y+1, 10-Y+1)$; the beta-distribution is a family of distributions on (0,1) with two parameters, e.g. the beta(1,1)-distribution is the same as a uniform distribution on (0,1). Figure 1 shows the beta-distributions with parameters (6,6) and (9,3) corresponding to observed values of $Y = 5$ and $Y = 8$, respect

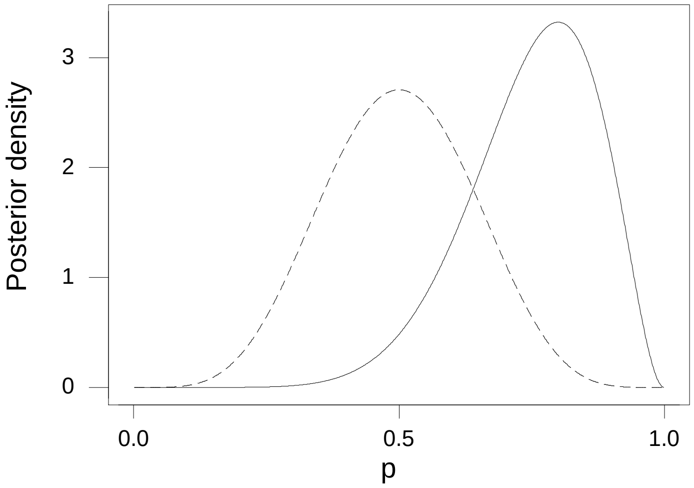

Figure 1: Posterior distributions after observing 5 animals (dashed line) and 8 animals (solid line) out of 10 tested animals positive.

If we wanted to summarize our knowledge about p after the testing into a single value, we could use the mean, median or mode of the distribution; for the two beta-distributions they equal (0.5, 0.5, 0.5) and (0.75, 0.764, 0.8), respectively. These values may be compared with the usual estimates p=0.5 and p=0.8; the agreement of mode and maximum-likelihood estimate is no coincidence! If we wanted to summarize our knowledge about p into a 95% interval, we could choose the interval with endpoints equal to the 2.5% and 97.5% percentiles of the distribution; for the two beta-distributions they are (0.234,0.736) and (0.482,0.940). These intervals may be compared with the (plus 4 corrected) binomial confidence intervals of (0.238,0.762) and (0.478,0.951); the intervals are quite similar.

3. Choice of prior distributions

Generally, it can be said that the strength and weakness of Bayesian methods lie in the prior distributions. In highly multidimensional and complex problems it is possible to incorporate model structure by means of prior distributions; such an approach has been fruitful for example in image analysis. The posterior of one analysis can also be taken as the prior for a subsequent study, thereby enabling successive updates of the collected and available information (empirical Bayes method). On the other hand, the choice of prior distributions might seem to open for a certain arbitrariness in Bayesian analysis, even if subjectivity in the prior does not at all contradict the Bayesian paradigm. In the past, priors have often been chosen of a particular form allowing for explicit calculation of the posterior (**conjugate** priors) but with access to MCMC methods these have decreased in importance.

A common choice of prior (in particular among less devoted Bayesian researchers) is a **non-informative** (or flat or diffuse) prior, which gives minimal preference to any particular values for θ. As an extreme case, if we take $p(\theta) \equiv 1$ in (1), the posterior distribution is just the likelihood function. So, for example, maximizing the posterior (MAP estimate) yields exactly the maximum likelihood estimate. Therefore, we would by and large expect Bayesian inference with noninformative priors to be similar to likelihood-based inference. To take $p(\theta)$ constant is not always possible, but alternatives exist; further detail is beyond these notes.

Example (ctd): Bayesian analysis of proportions
Let us expand the previous example to illustrate the impact of noninformative prior distributions. As a basis for the discussion, we introduce the **beta-distribution** with parameters (α, β) as the probability distribution on (0,1) with density given by

$$f(p|\alpha,\beta) = const \times p^{\alpha-1}\,(1-p)^{\beta-1}\,,$$

With this definition, the uniform distribution on (0,1) is in fact a beta-distribution with parameters (1,1). If we in addition to $Y \sim \mathrm{Bin}(n,p)$ assume as our prior distribution for p a beta-distribution with parameters (α, β), the posterior distribution in equation (1) takes the form

$$f(p|Y) = const(Y)\binom{n}{Y}p^{Y}\,(1-p)^{n-Y}\,const\,p^{\alpha-1}\,(1-p)^{\beta-1} = const(Y)\,p^{Y+\alpha-1}\,(1-p)^{n-Y+\beta-1}$$

This is seen to be another beta-distribution, with parameters $(Y+\alpha, n-Y+\beta)$. By this special property of the beta-distribution it is called the conjugate prior for a binomial model. We use the result to illustrate the effect of non-informative prior distributions. The graph below shows 3 prior distributions.

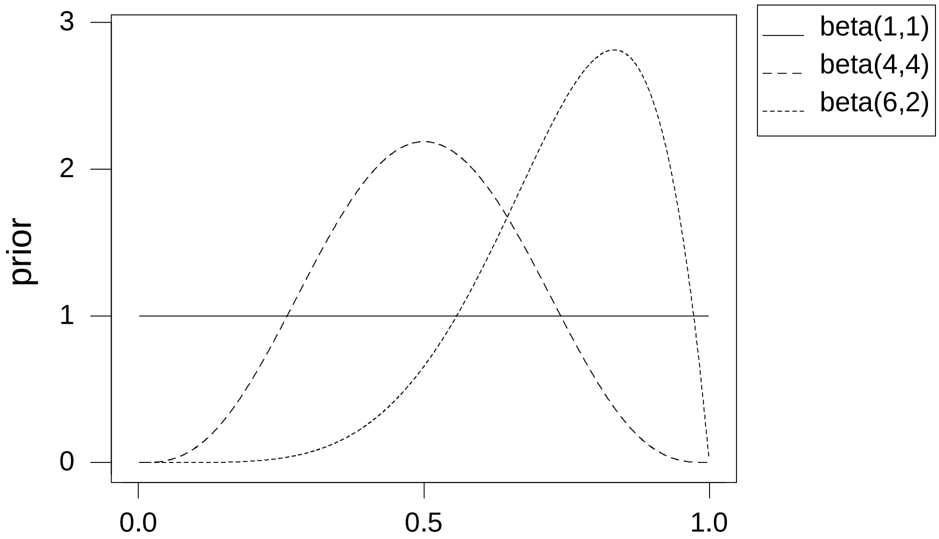

Figure 2: Three prior distributions for a proportion p.

Assume as in the previous example that $Y = 5$ out of $n = 10$ positives are observed. Then the posterior distributions corresponding to the 3 priors above, are beta-distributions with densities shown below.

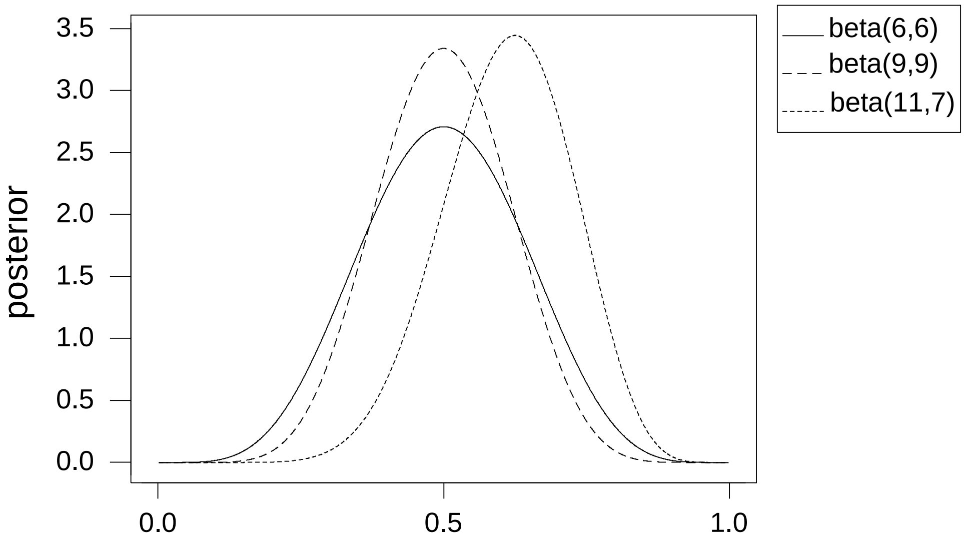

Figure 3: Posterior distributions after observing 5 animals out of 10 tested animals positive, using the three prior distributions of Figure 2.

Comparing the posterior for the noninformative and symmetric prior, we note a similar shape but a more concentrated posterior distribution for the informative prior. Comparing these with the asymmetric prior, we see that the data and prior have been combined into a quite symmetric distribution which is only slightly shifted to the right of the two other distributions. Generally, with a larger dataset the data gets stronger weight and the effect of prior distributions on the posterior diminishes. Conversely, with a more peaked prior distribution the balance tips towards the prior.

4. Summary

The Bayesian approach to statistics is an alternative method for all types of statistical analyses, contrasting the classical approach on such a fundamental point as the meaning of a parameter. Specifically, in Bayesian statistics parameters are distributions whereas they in classical statistics are unknown constants. As the two approaches *are* fundamentally different, they will generally give different results and different types of results. However, it is usually possible to make certain choices within the Bayesian framework to obtain results close and akin to those in classical statistics (though to a lesser extent so for statistical tests). Strengths of the Bayesian approach are the intuitive appeal of the posterior distribution as the base of inference, the possibility to build models for complex structures and successive updates of information by means of the prior distribution, and with the recent advances in MCMC simulation methods also the accessibility of analyses in models too complex to manage using classical methods. Problems with the Bayesian approach are (in view of the author of these notes) the subjectivity entering the analysis by the choice of a prior distribution and the added complexity for “simple” statistical situations by specification of prior distributions.