Bayesian Statistics

The Theory of Inverse Probability

Mohan Delampady and T.Krishnan

1 Introduction

The theory and methodology of Statistics is mostly concerned with inductive inference—inference from the particular to the general. We have some data and we wish to make statements—*inferences*—about one or more unknown features of the physical system which gave rise to these data. Although this sounds vague, this is the nature of the widely varied inference problems that arise in general.

A simple example of a situation asking for statistical inference is the following: An opinion poll produces data consisting of 'Yes' or 'No' opinion of a sample of people from a certain population in response to a question, say whether the government should resign in view of a recent corruption exposé. The system giving rise to these data consists of a population of individuals, a mechanism for selecting a sample from that population, and a mechanism which produces a 'Yes' or 'No' response from each individual in the sample. The answering mechanism for instance, may assume that the respondent is truthful and his answer reflects his true opinion. The sampling mechanism is a known probability scheme, say a simple random sampling scheme without replacement. The unknown feature of the system is the proportion, θ, of individuals in the population who hold the 'Yes' opinion. The object of statistical inference then is to use the observed responses in the sample to make statements about θ. Evidently, exact statements cannot be made about the value of θ, and any assessment of the value of θ is subject to probabilistic or stochastic behaviour. It is the object of statistical inference to nonetheless make useful statements about θ within this limiting framework. And here there are two rather different paradigms.

2 Classical versus Bayesian Inference

Example 1. We have seen elsewhere (e.g. *Resonance*, Vol. 1, No. 5, pp. 49-58) that if we define X to be the number of 'Yes' responses in a sample of n randomly chosen individuals, then X can be modelled as a binomial random variable with the probabilities

$$P(X = k \mid \theta) \;=\; \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad k = 0, 1, \ldots, n. \tag{1}$$

The notation $P(X = k \mid \theta)$ is employed here to make it explicit that the above probability distribution is the one that we get when the proportion (of interest) is fixed at (or conditioned on) θ. In Bayesian inference, the parameters are also regarded as random variables, and thus the above probability is regarded as conditional probability when the random variable Θ takes the value θ.

The purpose of statistical analysis is fundamentally an inversion, aiming at retrieving the 'causes' (parameters of the probabilistic data generating mechanism) from the 'effects' (observations). Because of this perspective, at the time of Bayes and Laplace (late eighteenth century), Statistics was known as 'Inverse Probability'. When observing a random phenomenon driven by a parameter θ, statistical methods allow us to deduce from these observations an inference about θ, while probabilistic modelling of the random phenomenon characterizes the behaviour of the future (or unseen) observations conditional on θ. This 'inverting' aspect of Statistics is clear in the notion of the *likelihood* function: Formally, it is just the sample density rewritten as a function of θ for the observed value of sample data x:

$$\ell(\theta|x) = f(x|\theta).$$

The important thing to note is that given x, it can be interpreted as a function showing the likelihood of different values of the parameter θ. One of the most popular statistical techniques for estimating parameters, then naturally

appears as that of maximizing this likelihood function as a function of θ for the observed sample data x. The interpretation of such an estimate is that it is the parameter of that model which is most likely to have produced the sample data x.

In the example above, the likelihood function of the proportion θ is simply

$$\ell(\theta|x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \tag{2}$$

with $\log \ell(\theta|x) = \text{constant} + x\log\theta + (n-x)\log(1-\theta)$ so that the maximum likelihood estimate of θ, which maximizes ℓ or $\log\ell$ with respect to θ is the familiar estimate $\hat{\theta} = x/n$, the sample proportion.

Now we have solved one problem. However, we are faced with another question. We have only an estimate of the true quantity, not the true quantity itself. How precise is our estimate? What is the estimation error? Many statisticians and users will also want interval estimates or confidence intervals for θ. These issues cannot be resolved by looking at the likelihood function. One possible approach is to consider the sampling distribution of the estimate: Imagine that we sample again and again and obtain a whole collection of these estimated $\hat{\theta}$ values; from these we can hope to construct a probability histogram. In our example, this can be achieved theoretically by utilizing the normal approximation to the binomial distribution (see *Resonance*, Vol. 2, No. 6, pp. 15-24) to claim that $\hat{\theta} \pm 2\sqrt{\hat{\theta}(1-\hat{\theta})/n}$ is an approximate 95% confidence interval for large values of the sample size n. What does this mean? Simply that, if we sample again and again a very large number of times, then in about 19 cases out of 20, this (random) interval will contain the actual unknown value of θ. However, if we pick a particular sample and construct the interval $\hat{\theta} \pm 2\sqrt{\hat{\theta}(1-\hat{\theta})/n}$, we don't know what to say about this particular interval. This illustrates the main difference between the frequentist and the Bayesian approach to statistics. The frequentist approach, as indicated above in the construction of a confidence set, relies on the long-run behaviour of a statistical procedure, whereas the Bayesian approach insists on using the

likelihood function of the parameter in question for the actual observed data for inference. This is the idea of conditioning on the observed data employed by the Bayesian approach as explained below.

Example 2. A blood test is to be conducted to help indicate whether or not a person has a particular disease. The result of the test is either 'positive' ($x = 1$) or 'negative' ($x = 0$). Let θ_1 denote 'disease is present' and θ_2 'disease not present'. Suppose the probability distribution of X under the different θ's is:

	$x = 0$	$x = 1$
θ_1	0.2	0.8
θ_2	0.7	0.3

Now, suppose that the result comes out to be 'positive' ($x = 1$). What should the doctor who suggested the blood test conclude? The probabilities given above do not make the blood test a foolproof method either way and whatever course of action the doctor takes is subject to error. Now, can he improve the assessment of these probabilities by using additional information on the disease prevalence? Suppose that in the community concerned, this disease is present in 5% of the cases. i.e., $P(\theta = \theta_1) = 0.05$. These probabilities of 0.05 and 0.95 for θ_1 and θ_2 respectively are called 'prior probabilities'. Now one can proceed as follows: For $i = 1, 2$,

$$\begin{aligned} P(\theta = \theta_i \mid X = x) &= \frac{P(\theta=\theta_i \text{ and } X=x)}{P(X=x)} \\ &= \frac{P(X=x|\theta=\theta_i)P(\theta=\theta_i)}{P(X=x|\theta=\theta_1)P(\theta=\theta_1)+P(X=x|\theta=\theta_2)P(\theta=\theta_2)}. \end{aligned} \tag{3}$$

From (3), it follows easily that

$$P(\theta = \theta_1 \mid X = 1) = 0.123, \quad \text{and} \quad P(\theta = \theta_2 \mid X = 1) = 0.877.$$

These are called 'posterior probabilities'. This means that a positive blood test indicates only a 12.3% chance of the disease being present in a random

member of this community. Notice that if the prior probability of $\theta_1 = 0.35$, then the posterior probability for 'disease present' increases to 0.577. Although this approach does not help diagnose a particular case with any more accuracy, it results in minimizing the overall probability of misdiagnosis when applied to members of this community. If misdiagnosis of 'disease present' and 'disease absent' are not equally serious, then these differences (if evaluated) can be used to further modify this diagnostic strategy. Anyway, in a situation like this, the doctor would most probably want further diagnostic measures!

The formula (3) which shows how to convert ('invert') the given conditional probabilities, $P(X = x \mid \theta)$ into the conditional probabilities of interest, $P(\theta \mid X = x)$ is an instance of the Bayes Theorem, and hence the *Theory of Inverse Probability* is known these days as *Bayesian inference*.

It is now clear what the basic ingredients of Bayesian inference are. In addition to the likelihood function (of the unknown parameters) which usually comes out of the model for the data, we also need a (prior) probability distribution $\pi(\theta)$ on the unknown parameters. Once we have both of these, the Bayes theorem can readily provide us a post-data (post-experimental or simply posterior) distribution of the unknown parameters conditional on the observed sample data. What this means, in particular, is that we actually end up having a probability model for the unknown parameters for the purpose of inference and future predictions. Often, a suitable way of summarizing this probability model is to be worked out. Indeed, specification of the prior probability distribution is hardly an easy exercise as is evident from the situation of Example 2.

Example 1. continued. Let us go back to our Example 1. Suppose that we have no special information available on the unknown proportion θ (apart from what we hope to get from sample data x). Then we may assume that θ is uniformly distributed on the interval $(0, 1)$. i.e., the prior density is $\pi(\theta) = 1$, for $0 < \theta < 1$. This is generally the way (prior) 'ignorance' is specified in a

Bayesian approach. Bayesians even use such uniform priors over an infinite range, calling them 'improper priors'. Oftentimes, Bayesian inference from such a prior coincides with classical inference; quite frequently, Bayesians use this phenomenon to justify their approach. Maybe then classical inference is the same as 'ignorant Bayesian inference'! In the Example then we readily see that the posterior density of θ given x is given by

$$\begin{aligned} \pi(\theta|x) &= \frac{\pi(\theta)f(x|\theta)}{\int \pi(u)f(x|u)\,du} \\ &= \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}, \quad 0<\theta<1. \end{aligned}$$

Notice the similarity and difference between this and the likelihood (2). The likelihood is a function of θ and is obtained from the probability distribution of the observed discrete random variable X. This posterior density is that of the parameter θ looked upon as a continuous random variable. However, as functions of θ, they are (essentially) the same and maximizing the posterior probability will give the same estimate as the maximum likelihood estimate. But this is only because in using the uniform prior, we have not really put in any information on θ. Notice that this is just the density of the Beta distribution with parameters $x+1$ and $n-x+1$. (Note that the probability density function of the Beta distribution with parameters a and b is given by $f(y) = \Gamma(\alpha+\gamma)/\{\Gamma(\alpha)\Gamma(\gamma)\}y^{\alpha-1}(1-y)^{\gamma-1}$ for $0<y<1$.) As a matter of fact, the uniform prior we used is a special case of the Beta distribution with parameters 1 and 1. If we had some knowledge of θ which can be summarized in the form of a Beta prior distribution with parameters α and γ, it is easily seen that the posterior will be also Beta with parameters $x+\alpha$ and $n-x+\gamma$. Choices of α and γ lead to a wide spectrum of distributions helping to summarize prior knowledge of θ. Such priors which with the likelihood result in posteriors from the same 'family' are called 'natural conjugate priors'. Before modern-day computational facilities were developed, Bayesians used to formulate their priors in this fashion in order to make the mathematics and computations tractable. With modern developments in statistical com-

putation techniques and with the availability of more powerful computers, this situation has changed and Bayesians dare to formulate more realistic priors. Most often these prior densities do not have a matching form with the likelihood function. Then, the posterior density will not have a standard form either, and so posterior measures of centre and spread of θ can be very difficult to compute. Recently, a lot of techniques have been developed to facilitate these computations which go by the name of Bayesian statistical computing. Basics of these were touched upon in the series on *Statistical Computing* by S Kunte in *Resonance* previously (Vol. 4, No. 10 and Vol. 5, No. 4). Some of the more advanced tools will be discussed in subsequent articles.

In this Example, the uncertainty in θ can now be described in terms of an actual probability distribution concentrated around the maximum likelihood estimate $\hat{\theta} = x/n$. However, the interpretation of $\hat{\theta}$ as an estimate of θ is quite different now. It is the most probable value of the unknown parameter θ conditional on the sample data x; it is called the 'maximum a posteriori estimate' often abbreviated as 'MAP estimate', a Bayesian analogue of the Maximum Likelihood Estimate (MLE). Indeed if the prior is a Beta distribution with parameters α and γ, the MAP estimate will be $\hat{\theta} = (x+\alpha-1)/(n+\alpha+\gamma-2)$, different from the MLE unless $\alpha = \gamma = 1$. This is a convex combination of the information in the sample and prior information, the weights depending upon the sample size and the strength of the prior information as measured by the values of α and γ. Further, since we also have a probability distribution to quantify our (post-experimental) uncertainty in θ, a measure of estimation error can be taken to be the standard deviation of this posterior distribution. In fact, we can say much more. For any interval around $\hat{\theta}$ we can compute the (posterior) probability of it containing the true parameter θ. (Note, however, that this computation will involve incomplete Beta Integrals which have to be done numerically.) The final conclusion is that all these inferences are conditional on the given data. Certainly, this makes more sense as a method of inference from the observed sample data. An important

point to note at this time is that Bayesian inference relies on the conditional probability language to revise one's knowledge. In the above example, prior to the collection of sample data one had some (vague, perhaps) information on θ. Then came the sample data. Combining the model density of this data with the prior density one gets the posterior density, the conditional density of θ given the data. From now on until further data is available, this posterior distribution of θ is the only relevant information as far as θ is concerned.

3 Penalized Likelihood?

To better understand how the Bayesian approach manipulates the likelihood function we extend Example 1 as follows.

Example 3. Suppose that k independent random variables $y_1, y_2, \ldots, y_k$ are observed, where y_i has the Binomial$(n_i,\ p_i)$ probability distribution, $1 \leq i \leq k$. y_i may be the number of laboratory animals cured of an ailment in an experiment involving n_i such animals. It is certainly possible to make inferences on each p_i separately based on the observed y_i (as discussed previously). This, however, is not really useful if we want to predict the results of a similar experiment in future. Suppose that the p_i are related to a covariate or an explanatory variable, such as dosage level in a clinical experiment. Then the natural approach is what is called regression: exploring and presenting the relationship between design (explanatory) variables and response variables, and (if needed) predictions of response at desired levels of the explanatory variables. Let t_i be the value of the covariate which corresponds to p_i, $i = 1, 2, \ldots, k$. Linking of p_i and t_i is made through a link function H: $p_i = H(\alpha + \beta t_i)$. H, here, is a known cumulative distribution function (c.d.f.) and α and β are two unknown parameters. (If H is an invertible function, this is precisely $H^{-1}(p_i) = \alpha + \beta t_i$.) If the standard normal c.d.f. is used for H, the model is called the probit model, and if the logistic c.d.f. (i.e., $H(z) = e^{-z}/(1 + e^{-z})$) is used, it is called the logit model. The

likelihood function for the unknown parameters, α and β, is then given by

$$\prod_{i=1}^{k} \binom{n_i}{y_i} H(\alpha+\beta t_i)^{y_i} \left(1-H(\alpha+\beta t_i)\right)^{n_i-y_i}.$$

Clearly, this function is largely intractable. For the given data, one can still numerically compute the maximum likelihood estimates. The Bayesian approach requires a prior distribution with density $\pi(\alpha,\beta)$, when combined with the likelihood function above yields the following posterior distribution:

$$\pi(\alpha,\beta|\text{data}) = \frac{\pi(\alpha,\beta)\prod_{i=1}^{k} H(\alpha+\beta t_i)^{y_i}\left(1-H(\alpha+\beta t_i)\right)^{n_i-y_i}}{\int \pi(a,b)\prod_{i=1}^{k} H(a+bt_i)^{y_i}\left(1-H(a+bt_i)\right)^{n_i-y_i}\, da\, db}.$$

How different is this from the likelihood function? In particular, how different is the MAP estimate of (α,β) from its MLE? To illustrate this we use the logit model, so that $p_i = e^{-(\alpha+\beta t_i)}/\{1+e^{-(\alpha+\beta t_i)}\}$, and hence $-\log(p_i/(1-p_i)) = \alpha+\beta t_i$. We will also employ a large sample approximation (i.e., that the n_i are large enough for a satisfactory Gaussian approximation of the Binomial model. Also, for convenience we shall employ the standard notation of $N(\mu,\sigma^2)$ for a normal or Gaussian distribution with mean μ and variance σ^2.) Consequently, if we let $\hat{p}_i = y_i/n_i$, then (approximately), these $\hat{p}_i$ are independent $N(p_i, p_i(1-p_i)/n_i)$ random variates. Now let $\theta_i = -\log(p_i/(1-p_i))$ and $\hat{\theta}_i = -\log(\hat{p}_i/(1-\hat{p}_i))$. It can be shown that, approximately, $(\hat{\theta}_i-\theta_i)\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}$ are independent $N(0,1)$ random variates. Then, (again approximately), the likelihood function for (α,β) is

$$\ell(\alpha,\beta|\text{data}) \;=\; \exp\left(-\frac{1}{2}\sum_{i=1}^{k} w_i(\hat{\theta}_i-(\alpha+\beta t_i))^2\right), \tag{4}$$

where $w_i = n_i\hat{p}_i(1-\hat{p}_i)$ are known weights. Now suppose that *a priori* α and β are independent $N(a_0,\tau_a^2)$ and $N(b_0,\tau_b^2)$ respectively. Then the approximate posterior density is

$$\pi(\alpha,\beta|\text{data}) \propto \exp\left(-\frac{1}{2}\left\{\sum_{i=1}^{k} w_i(\hat{\theta}_i-(\alpha+\beta t_i))^2 + (\frac{\alpha-a_0}{\tau_a})^2 + (\frac{\beta-b_0}{\tau_b})^2\right\}\right). \tag{5}$$

Now let us note that finding the MLE and MAP estimates of (α, β) is equivalent to maximizing the logarithm of the likelihood function (4) and the logarithm of the posterior density (5) respectively. This in turn is equivalent to minimizing

$$R(\alpha, \beta) \;=\; \sum_{i=1}^{k} w_i(\hat{\theta}_i - (\alpha + \beta t_i))^2 \tag{6}$$

to find the MLE and, minimizing

$$PR(\alpha, \beta) \;=\; \sum_{i=1}^{k} w_i(\hat{\theta}_i - (\alpha + \beta t_i))^2 + (\frac{\alpha - a_0}{\tau_a})^2 + (\frac{\beta - b_0}{\tau_b})^2 \tag{7}$$

to find the MAP estimate. From (6), it is clear that MLE of (α, β) is the weighted least squares estimate of these regression coefficients. On the other hand, the MAP estimate in this case is different. $PR(\alpha, \beta)$ needs to balance the weighted residual sum of squares $R(\alpha, \beta)$ with (weighted) departures of α and β from their prior means a_0 and b_0. This will force it to penalize any candidates for the estimates which are far away from the prior means of the regression coefficients. Therefore, broadly speaking, the Bayesian approach to inference is a penalized likelihood method, where the penalty is for departing away from the prior inputs. The prior variances, τ_a^2 and τ_b^2 indicate how much weight is to be given to these departures from the means in the penalty term.

4 Bayesian Computation

As we mentioned earlier, Bayesians increasingly use rather more realistic and hence more complicated priors, leading to difficult computational problems, which seem to be increasingly solvable thanks to developments in statistical computing techniques and in the availability of more computing power. We now introduce an example of this kind, due to Casella and George given in Arnold (1993).

Example 4. Suppose we are studying the distribution of the number of defectives X in the daily production of a product. Let us model $(X \mid Y, \theta) \sim Bin(Y, \theta)$, where Y a day's production is a random variable with a Poisson distribution with known mean λ, i.e., $Y \sim Poi(\lambda)$ and θ is the probability that a product is defective. In a Bayesian formulation, let $(\theta \mid Y = y) \sim Beta(\alpha, \gamma)$, with known α and γ. The solution to the Bayesian estimation problem would require the conditional distribution $(Y, \theta \mid X)$ which is

$$\frac{P(x, y, \theta)}{P(X = x)}.$$

The numerator is easy to work out as $P(X = x \mid y, \theta) \times P(y)P(\theta)$, but the denominator, the marginal distribution of X is complicated. It is seen to be

$$\frac{e^{-\lambda}}{x!\beta(\alpha, \gamma)} \sum_{y=0}^{\infty} \frac{\lambda^y}{(y-x)!} \beta(x + \alpha, y + \gamma - x),$$

where β is the complete Beta function. This sum is very difficult to work out analytically and in contexts like these either numerical methods are used if possible, or statistical simulation techniques are used. We discuss these methods in a subsequent article.

Suggested Reading

1. S F Arnold, Gibbs sampling, In C.R.Rao (Ed): *Handbook of Statistics, Vol. 9*, Elsevier Science, Ch. 18, pp. 599–625, 1993.

2. J O Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd Edition, Springer-Verlag, New York, 1985.

3. A Gelman, J B Carlin, H S Stern and D B Rubin, *Bayesian Data Analysis*, Chapman & Hall, London, 1995.

4. H Jeffreys, *Theory of Probability*, 3rd Edition, Oxford University Press, New York, (Reprinted) 1985.

5. P M Lee, *Bayesian Statistics: An Introduction*, Oxford University Press, New York, 1989.

6. Science and Technology Columns, In praise of Bayes, *The Economist*, September 30th, p. 91-92, 2000.