

A Closer Look at Bayesian Analysis

ARNULFO P. SUPE

Performing a statistical analysis may be done in two approaches, the Classical approach and the Bayesian approach. The term *classical* does not have the unanimous support of all statisticians, but is usually used to describe inference with the following characteristics:


1. Estimators and test procedures are evaluated in terms of their properties in repeated samples.
2. The probability of an event is defined in terms of the limit of the relative frequency of that event.
3. There is no provision for the formal inclusion of nonsample information.

In a Bayesian framework, probability is defined in terms of degree of belief and although the properties of estimators and tests in repeated samples are of some interest, they do not provide for the main basis for inference and estimator choice. One of the main features of Bayesian analysis is that uncertainty about the value of an unknown parameter can be expressed in terms of a probability distribution. In a Bayesian framework, parameters are treated as random variables, not in the sense that different outcomes of an experiment yield different realizations of a parameter, but in the sense that a parameter has associated with it a subjective probability distribution that describes our state of knowledge about that parameter.

In classical methods inferences are based on the average performance of a procedure over *all possible samples*, while the Bayesian approach give primary importance to the performance of a procedure for the *actual data* that is observed in a given experiment. Bayesians believe that it is more practical and realistic to base inference on actual data sampled rather than base it on large repetitions of sample which may or will not occur.

1. BAYES' THEOREM

When one performs Bayesian analysis, one uses prior information aside

 ARNULFO P SUPE is an Associate Professor of Mathematics and Statistics at the College of Sciences and Mathematics. He finished his PhD in Statistics at the University of the Philippines, Diliman, Quezon City.

from sample data in order to generate inferences about the parameter of the model from which the data was hypothesized to come. As opposed to classical statistics where only sample information is used to draw inferences, prior information is a necessary ingredient in a Bayesian procedure. Let

$\Theta = (\Theta_1, \Theta_2, \dots, \Theta_r)$ be the parameter of interest,
 $h(\theta)$ the prior density associated with Θ , and
 $f(x|\theta)$ the density from which the sample was taken.

Bayes' Theorem states that the *posterior density* of Θ given the sample information, denoted by $\pi(\theta|x)$, is, for a continuous Θ ,

$$\pi(\theta|x) = \frac{h(\theta)f(x|\theta)}{m(x)}, \quad (1.1)$$

where

$$m(x) = \int \dots \int f(x|\cdot) h(\cdot) d\theta$$

Since $m(x)$ does not involve θ , we may rewrite (1.1) as

$$\pi(\theta|x) \propto h(\theta)f(x|\theta),$$

where the symbol " \propto " means "is proportional to". This simplification is used in the illustration in Section 4.

2. THE PRIOR INFORMATION

Most of the objections leveled against Bayesian statistics are directed towards the treatment of θ as a random variable with a probability distribution. The use of the prior density, considered as "subjective" by classical statisticians, is to date still the most controversial aspect of Bayesian analysis. Bayesians argue, however, that adopting a frequency distribution for X and a prior distribution for θ are both quite subjective activities.

Two of the methods of determining prior distributions that have gained popular support are the noninformative priors and the use of conjugate families.

2.1. Methods of Determining Prior Distributions

A *noninformative prior* is a prior that contains no information about θ , or more crudely, a prior which favors no particular value of θ over the others. A simple example of a noninformative prior is one which assigns for a finite $\Theta = (\Theta_1, \Theta_2, \dots, \Theta_k)'$, a probability of $1/k$ for each element Θ_i of Θ .

Jeffrys (1961) developed a theory of choosing prior densities based on rules of parameter invariance where "nothing" is known about the values of the parameters. Below is a summary of the methods for choosing a noninformative prior as recommended by Jeffrys.

(i) For a location parameter θ (the density is of the form $f(x - \theta)$), use $\pi(\theta) \propto c$, where c is a constant;

(ii) For a scale parameter σ (the density is of the form

$$(1/\sigma)f(x/\sigma)), \text{ use } \pi(\theta) \propto \frac{1}{\sigma};$$

(iii) For a more general setting (univariate case), use

$$\pi(\theta) \propto [I(\theta)]^{-\frac{1}{2}},$$

where $I(\theta)$ is the expected information measure under commonly satisfied conditions;

(iv) For a vector valued $\Theta = (\Theta_1, \Theta_2, \dots, \Theta_k)'$, use

$$\pi(\theta) \propto [\det I(\theta)]^{-\frac{1}{2}}$$

where $\det[A]$ means the determinant of a square matrix A .

It is often the case that the analysis from the use of noninformative priors yields the same result as the one done through classical methods. However, the interpretations of the two results will be different. There are many situations in which the two results differ and the classical report almost invariably suffers in comparison.

Berger (1985) argues that "noninformative prior Bayesian analysis is the single most powerful method of statistical analysis, in the sense of being the ad hoc method most likely to yield a sensible answer for a given investment of effort".

One of the operational advantages of Bayesian statistics over classical statistics is its more effective use of prior information when significant prior information is available. Classical statistics does not provide for a systematic utilization of reliable prior information, in a more

formal framework. In cases where no or very little prior information is available, Bayesians can always fall back to noninformative priors to get a sensible analysis.

2.2. Conjugate Families

There are many instances where $\pi(\theta|x)$ is not easily calculable. For this reason, a large part of the Bayesian literature was devoted to finding priors that can be easily calculated. The use of conjugate families is an answer to this problem. Below is a definition taken from Berger (1985).

Let F denote the class of density functions $f(x|\theta)$. A class H of prior distributions is said to be a *conjugate family* for F , if the posterior distribution of Θ given x , $\pi(\theta|x)$, is in the class H for all $f \in F$ and $h \in H$.

Aside from providing an easy method of finding $\pi(\theta|x)$, conjugate priors also have the intuitively appealing feature of allowing one to begin with a certain functional form of the prior and end up with a posterior of the same functional form, but with parameters updated by the sample information.

An example of a conjugate family is the normal-gamma distribution. The joint density of two random variables X_1 and X_2 is a normal-gamma with parameters μ, p, α, β if and only if

$$f(x_1, x_2) = f_1(x_1|x_2) f_2(x_2) ; -\infty < x_1 < +\infty, x_2 > 0,$$

where $f_1(x_1|x_2)$ is normally distributed with mean μ , and precision px_2 , and $f_2(x_2)$ is a gamma distribution with parameters α and β .

3. BAYESIAN INFERENCE

From a Bayesian viewpoint, all inferences about Θ are based on the posterior distribution of Θ given the sample. The idea is that, since the posterior distribution supposedly contains all the available information about Θ (both sample and prior information), any inference concerning Θ should consist solely of features of this distribution.

3.1. Point Estimation

To estimate θ , some of the classical estimation techniques can be applied to the posterior distribution. The most common classical technique is maxi-

mum likelihood estimation, which chooses, as the estimate of Θ , the value θ^* which maximizes the likelihood function. The analogous Bayesian estimate is defined below.

The *Generalized Maximum Likelihood Estimate* of Θ is the largest mode, θ^* , of $\pi(\theta|x)$, i.e., the value θ^* which maximizes $\pi(\theta|x)$, where $\pi(\theta|x)$ is considered as a function of θ .

Some Bayesians maintain that *inference should ideally consist of simply reporting the entire posterior distribution $\pi(\theta|x)$* . The reason advanced is that since the posterior distribution is an actual probability distribution for Θ , one can derive from it, any feature of interest and a visual inspection of the graph of the posterior will often provide the best insight concerning Θ .

If Θ is a vector and the corresponding posterior is thus a joint density, the marginal posterior density of any element of Θ can be found by integrating out the other elements of Θ .

3.2. Credible Set

The Bayesian analog of a classical confidence set is called a *credible set*. A $100(1 - \alpha)\%$ **credible set** for Θ is a subset C of the parameter space Ω , such that

$$1 - \alpha \leq P(C | x) = \int_C dF^{x(\theta|x)},$$

$$\text{where } \int_C dF^{x(\theta|x)} = \begin{cases} \int_C \pi(\theta|x) & (\text{for } \theta \text{ continuous}) \\ \sum_{\theta \in C} \pi(\theta|x) & (\text{for } \theta \text{ discrete}) \end{cases}$$

Since the posterior distribution is an actual probability distribution of θ , one can speak meaningfully of the probability that θ is in C . This is not the case for classical procedures where one can interpret *only* in terms of coverage probability.

3.3. Hypothesis Testing

Testing $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega_1$ is conceptually more straightforward within the Bayesian framework. One simply calculates

$$\alpha_0 = P(\Omega_0|x) \text{ and}$$

$$\alpha_1 = P(\Omega_1|x),$$

then decide logically whether to accept or reject the hypothesis based on these actual probabilities. Another method would be to accept or reject the hypothesis based on whether or not the specified value under a null hypothesis H_0 lies within a *highest posterior density credible set* with content $1 - \alpha$.

4. BAYESIAN ANALYSIS OF THE LINEAR MODEL

The Bayesian approach will now be illustrated below with the analysis of the general linear model. Recall that the least square estimate and the unbiased estimate of the parameter β is

$$\beta' = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}).$$

Consider now the general linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ where ε has a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma^2\mathbf{I}$ and

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

\mathbf{Y} is the $(n \times 1)$ vector of dependent variables, \mathbf{X} is the $(n \times p)$ matrix of independent variables of full rank p and β is the $(p \times 1)$ vector of parameters. Then the likelihood function of (β, σ^2) given the sample observations (\mathbf{X}, \mathbf{Y}) is

$$L(\beta, \sigma^2 | \mathbf{X}, \mathbf{Y}) \propto (\sigma^2)^{-n} \exp \left\{ -\frac{1}{2\sigma^2} [(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)] \right\} \quad (4.1)$$

Let the joint density of (β, σ^2) be given by the noninformative prior

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \quad (4.2)$$

Let $\delta = \frac{1}{\sigma^2}$. Combining the likelihood function and the joint prior, by Bayes' Theorem, yields

$$\pi(\beta, \delta | \mathbf{X}, \mathbf{Y}) \propto \delta^{\frac{n-1}{2}} \exp \left\{ -\frac{\delta}{2} [(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)] \right\} \quad (4.3)$$

To find the marginal density of β , we integrate δ out from (4.3) to get,

$$\begin{aligned} \pi_1(\beta | \mathbf{X}, \mathbf{Y}) &\propto \int_0^\infty \delta^{\frac{n-1}{2}} \exp \left\{ -\frac{\delta}{2} [(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)] \right\} d\delta, \\ \pi_1(\beta | \mathbf{X}, \mathbf{Y}) &\propto [(\mathbf{Y} - \mathbf{X}\beta^*)'(\mathbf{Y} - \mathbf{X}\beta^*) + (\beta - \beta^*)'\mathbf{X}'\mathbf{X}(\beta - \beta^*)]^{-\frac{n}{2}}, \end{aligned}$$

which can be rewritten as

$$\pi_1(\beta | \mathbf{X}, \mathbf{Y}) \propto \left[1 + \frac{(\mathbf{Y} - \mathbf{X}\beta^*)'(\mathbf{Y} - \mathbf{X}\beta^*)}{(n-p)} \right]^{-\frac{(n-p)+p}{2}} \quad (4.4)$$

The expression at the right hand side of (4.4) can be recognized as the kernel of a multivariate t-distribution with

degrees of freedom $(n-p)$,

mean vector $\beta^* = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$, and

precision matrix $\frac{(n-p)\mathbf{X}'\mathbf{X}}{(\mathbf{Y} - \mathbf{X}\beta^*)'(\mathbf{Y} - \mathbf{X}\beta^*)}$

A point estimate of β is $\beta^* = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$, which corresponds to the unbiased and least squares estimate of β . There are many situations where the Bayesian estimates, when using the noninformative prior, are the same as the estimates made through the classical approach. However, their interpretations will be different.

From the posterior density of β , hypothesis testing and the construction of credible sets become relatively straightforward. The predictive distribution of future independent observations is also easily derived. This illustrates the advantages of Bayesian analysis. All inference

problems are solved once one has found the posterior distributions of the parameters. Instead of learning a large variety of sampling theory techniques when inference is done by classical analysis, one only needs to learn how to apply Bayes' Theorem when doing it the Bayesian way.

References

- [1] Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer Verlag
- [2] Jeffrys, H. (1961), *Theory of Probability*, Oxford University Press