5 The Bayesian Approach to Statistical Inference

5.1 The Bayesian paradigm

In the approach to statistical inference that we have adopted so far, we have assumed that the vector of parameters θ is fixed, although unknown. We have made a point of emphasizing that no probability statements may be made about θ. In the Bayesian approach, which we now describe, a totally different viewpoint is adopted. As before, the data are observations from a family $\mathcal{F}$ of probability distributions,

$$\mathcal{F} = \{f(\mathbf{x};\theta),\ \mathbf{x} \in \mathcal{X} : \theta \in \Theta\}.$$

The experimenter's beliefs about θ are described in terms of probability statements: θ is regarded as a random vector, reflecting the uncertainty about the value of θ in the experimenter's state of mind. With that being the case, one may regard $f(\mathbf{x};\theta)$ as a conditional p.m.f./p.d.f. of $\mathbf{x}$ given θ, written as $f(\mathbf{x}|\theta)$, yielding

$$\mathcal{F} = \{f(\mathbf{x}|\theta),\ \mathbf{x} \in \mathcal{X} : \theta \in \Theta\}.$$

The experimenter's (or "our") subjective beliefs about θ before the data are observed are expressed in terms of a prior distribution or its p.d.f., the *prior density* $\{\pi(\theta),\ \theta \in \Theta\}$. After the data $\mathbf{x}$ have been observed, the information obtained from the data is used to compute a *posterior density* $\{\pi(\theta|\mathbf{x}),\ \theta \in \Theta\}$, which expresses our beliefs about θ as modified in the light of the data. All subsequent inferences about θ are based on this posterior distribution.

Given $\mathbf{x} \in \mathcal{X}$, the posterior density for θ is computed using Bayes' Theorem,

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta)f(\mathbf{x}|\theta)}{m(\mathbf{x})} \qquad \theta \in \Theta, \tag{1}$$

where $m(\mathbf{x})$ is the marginal p.d.f. of $\mathbf{X}$,

$$m(\mathbf{x}) = \int_\Theta \pi(\theta)f(\mathbf{x}|\theta)d\theta.$$

Defining $L(\theta;\mathbf{x}) = f(\mathbf{x}|\theta)$, in this context, $\theta \in \Theta$, Equation (1) may equivalently be expressed as

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta)L(\theta;\mathbf{x}), \qquad \theta \in \Theta, \tag{2}$$

where the constant of proportionality in Equation (2) is such as to ensure that the normalization condition

$$\int_\Theta \pi(\theta|\mathbf{x})d\theta = 1 \tag{3}$$

is satisfied. Equation (2) expresses the fact that the posterior density is proportional to the product of the prior density and the likelihood function.

It follows from Equation (2) that, in the Bayesian approach to inference, all the information about θ that is obtained from the observed data is contained in the likelihood function.

5.2 Likelihood and sufficiency in the Bayesian setting

Theorem 1 *The likelihood principle holds for Bayesian inference.*

Proof. Let $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ be such that they have proportional likelihood functions, i.e., $L(\theta; \mathbf{x}) \propto L(\theta; \mathbf{y})\ \theta \in \Theta$. Using Equation (2),

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto \pi(\theta)L(\theta; \mathbf{x}) \\ &\propto \pi(\theta)L(\theta; \mathbf{y}) \\ &\propto \pi(\theta|\mathbf{y}) \qquad \theta \in \Theta. \end{aligned}$$

Because both posterior densities must satisfy the normalization condition of Equation (3), it follows that

$$\pi(\theta|\mathbf{x}) = \pi(\theta|\mathbf{y}) \qquad \theta \in \Theta.$$

We have shown that two different sets of observations with proportional likelihood functions will result in the transformation of a given prior density into the same posterior density and hence lead to the same inferences about θ.

Theorem 2 *The sufficiency principle holds for Bayesian inference.*

Proof. The result follows immediately from Theorem 1 above and Theorem 3 of Chapter 1 that the likelihood principle implies the sufficiency principle.

Theorem 3 *A statistic T is sufficient for θ if and only if, given any prior density, for every $\mathbf{x} \in \mathcal{X}$, the posterior density of θ given $\mathbf{x}$ is the same as the posterior density of θ given $t \equiv T(\mathbf{x})$.*

Proof. Let T be a sufficient statistic for θ. Given $\mathbf{x} \in \mathcal{X}$ and $t \equiv T(\mathbf{x})$, from Theorem 1 of Chapter 1 it follows that $L^*(\theta; t) \propto L(\theta; \mathbf{x})$, where $L^*(\theta; t)$ is the likelihood function based on the family of distributions of T. Hence, by an argument similar to the one used in the proof of Theorem 1,

$$\pi(\theta|\mathbf{x}) = \pi(\theta|t) \qquad \theta \in \Theta. \tag{4}$$

Conversely, suppose that Equation (4) holds for all $\mathbf{x} \in \mathcal{X}$. Rearranging Equation (1),

$$\begin{aligned} f(\mathbf{x}|\theta) &= \frac{\pi(\theta|\mathbf{x})m(\mathbf{x})}{\pi(\theta)} \\ &= \frac{\pi(\theta|T(\mathbf{x}))m(\mathbf{x})}{\pi(\theta)} \qquad \mathbf{x} \in \mathcal{X},\ \theta \in \Theta \end{aligned}$$

using Equation (4). Thus we may write

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x}); \theta)m(\mathbf{x}) \qquad \mathbf{x} \in \mathcal{X},\ \theta \in \Theta,$$

where $g(t; \theta) \equiv \pi(\theta|t)/\pi(\theta)$. Hence by the factorization criterion (Theorem 2 of Chapter 1), T is sufficient.

Theorem 3 reaffirms in the Bayesian context that we may equally well base our inferences about θ on the family of distributions of any sufficient statistic T. Theorem 3 also shows that in the Bayesian context we may adopt an alternative definition of sufficiency that is equivalent to the one used in Chapter 1.

Definition

A statistic T is said to be sufficient for θ if, given any prior distribution, for every $\mathbf{x} \in \mathcal{X}$, the posterior distribution of θ given $\mathbf{x}$ is the same as the posterior distribution of θ given $t \equiv T(\mathbf{x})$.

5.3 Conjugate priors

The Bayesian paradigm provides a reasonable way of dealing with inferences about θ. However, a basic difficulty with this approach is that we have to express our prior beliefs about θ in terms of a prior distribution, which we may not find easy to do. The question arises of how accurately we are capable of representing our beliefs about θ in the form of a probability distribution.

Given the family of distributions $\mathcal{F} = \{f(\mathbf{x}|\theta),\ \mathbf{x} \in \mathcal{X} : \theta \in \Theta\}$, it is mathematically convenient to restrict the prior distributions to a class Π that is a *conjugate family* for $\mathcal{F}$.

Definition

A class Π of prior distributions $\pi(\theta)$ is a *conjugate family* for $\mathcal{F}$ if the posterior distribution $\pi(\theta|\mathbf{x})$ is in the class Π for all $\pi(\theta) \in \Pi$ and all $\mathbf{x} \in \mathcal{X}$.

Example

Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli r.v.s with parameter $\theta,\ 0 < \theta < 1$. Recall that the statistic $T(\mathbf{X}) \equiv \sum_{i=1}^n X_i$ is sufficient for θ and has the binomial $\mathrm{Bin}(n, \theta)$ distribution with

$$f(t|\theta) = \left(\begin{array}{c} n \\ t \end{array} \right) \theta^t (1-\theta)^{n-t}, \qquad t = 0, 1, \ldots, n.$$

We may base our inferences upon the value t of T. Take as the prior distribution for θ a beta distribution with parameters a and b, where $a > 0$ and $b > 0$, i.e.,

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \qquad 0 < \theta < 1.$$

- As the values of a and b are varied, the family of beta distributions provides a considerable amount of flexibility in modelling our prior beliefs about θ. The special case $a = 1,\ b = 1$ gives a prior distribution that is uniform on $(0, 1)$.
- The mean of the beta distribution with parameters a and b is $a/(a+b)$.

It follows from Equation (2) that

$$\begin{array}{rcl} \pi(\theta|t) & \propto & \pi(\theta)L(\theta; t) \\ & \propto & \theta^{t+a-1}(1-\theta)^{n-t+b-1}. \end{array}$$

Thus the posterior distribution is a beta distribution with parameters $t + a$ and $n - t + b$, so that

$$\pi(\theta|t) = \frac{\Gamma(n+a+b)}{\Gamma(t+a)\Gamma(n-t+b)} \theta^{t+a-1}(1-\theta)^{n-t+b-1}, \qquad 0 < \theta < 1.$$

We conclude that the family of beta distributions is conjugate for the binomial family.

Some information on **Thomas Bayes** (1702-61) can be found at:

`http://en.wikipedia.org/wiki/Thomas_Bayes.`

5.4 Uniform and improper priors

A distribution that is uniform over the parameter space may be taken to represent the belief that all values of θ are equally likely or, possibly, a state of complete ignorance about θ. However, this latter interpretation, sometimes referred to as the *principle of indifference* or the *principle of insufficient reason*, has its difficulties, especially if the parameter space is continuous. For example, if we carry out a re-parametrization and write $\eta = w(\theta)$, say, then a uniform distribution for θ is not in general transformed into a uniform distribution for η. It follows that, although the principle of indifference may be used to suggest possible prior distributions to represent ignorance, it will not generally provide a unique answer.

Note also that if the parameter space Θ is not bounded then there does not exist a uniform distribution on Θ. If we take $\pi(\theta) = k$ for some constant $k \geq 0$ then

$$\int_{\Theta} \pi(\theta) d\theta = \begin{cases} \infty & k > 0 \\ 0 & k = 0 \end{cases}$$

Thus it is not possible to make the integral take the value 1. However, in view of Equation (2), we only have to specify $\pi(\theta)$ up to a constant of proportionality. So we can proceed, using $\pi(\theta) = k$, despite the fact that $\pi(\theta)$ does not specify a proper distribution. In such a case, $\pi(\theta)$ is known as an *improper prior.*

If we do use a uniform prior distribution, whether proper or improper, then Equation (2) simply becomes

$$\pi(\theta|\mathbf{x}) \propto L(\theta; \mathbf{x}), \qquad \theta \in \Theta, \tag{5}$$

so that the posterior density is proportional to the likelihood function.

Another example of a commonly used improper prior when $\Theta = (0, \infty)$ is

$$\pi(\theta) = \frac{1}{\theta} \qquad \theta > 0,$$

which corresponds to $\ln \theta$ having an improper uniform prior over the interval $(-\infty, \infty)$.

Generally speaking, the use of improper priors works perfectly well in mathematical terms but is open to criticism on philosophical grounds.

5.5 Estimation

In the Bayesian framework, our posterior beliefs about θ are completely described by the posterior density, but we may wish to summarize the posterior density by giving a point estimate of θ. One possibility is to use the mode of the posterior density, that is, the value of θ that maximizes $\pi(\theta|\mathbf{x})$ $\theta \in \Theta$. It follows from Equation (5) that in the special case of a uniform prior this reduces to the maximum likelihood estimate.

In the case of a single parameter, a more commonly used estimator is the *Bayes estimator* $\hat{\theta}(\mathbf{X}) \equiv E[\theta|\mathbf{X}]$, the mean of the posterior distribution, which for any given $\mathbf{x} \in \mathcal{X}$ is the value of $\hat{\theta}$ that minimizes the posterior expected mean squared error,

$$\int_\Theta (\theta - \hat{\theta})^2 \pi(\theta|\mathbf{x}) d\theta.$$

Set or interval estimation from the Bayesian viewpoint is particularly simple.

Definition

Given an observed $\mathbf{x} \in \mathcal{X}$, a $100(1-\alpha)\%$ *credible set* (or *Bayesian confidence set*) for θ is any set $\mathcal{S}(\mathbf{x}) \subseteq \Theta$ such that

$$\int_{\mathcal{S}(\mathbf{x})} \pi(\theta|\mathbf{x}) d\theta = 1 - \alpha,$$

i.e.,

$$\mathbb{P}(\theta \in \mathcal{S}(\mathbf{x})|\mathbf{x}) = 1 - \alpha. \tag{6}$$

In contrast to the classical approach to set estimation, as outlined in Section 4.6, the probability statement of Equation (6) is now a probability statement about θ, the random vector of parameters, given the set $\mathcal{S}(\mathbf{x})$. Previously, it was the vector of parameters θ that was regarded as fixed and the confidence set $\mathcal{S}(\mathbf{X})$ as random.

Given $\mathbf{x} \in \mathcal{X}$, there will be many possible $100(1-\alpha)\%$ credible sets. One possible choice is a *highest posterior density credible set* $\mathcal{S}(\mathbf{x})$, one such that $\pi(\theta_1|\mathbf{x}) \geq \pi(\theta_2|\mathbf{x})$ for all $\theta_1 \in \mathcal{S}(\mathbf{x}),\ \theta_2 \notin \mathcal{S}(\mathbf{x})$.

So if $\pi(\theta|\mathbf{x})$ is unimodal, the set is contiguous (otherwise, it may not be).

5.6 Advantages and difficulties of the Bayesian approach

It is argued by some that the Bayesian methodology provides a much more unified and logically coherent system of inference than do the classical techniques.

Hypothesis testing as a distinctive form of inference does not arise in the Bayesian approach. The posterior distribution determines the probability that any given hypothesis about the parameter is true, i.e., the posterior probability $\mathbb{P}(\theta \in \Theta_0|\mathbf{x})$ that the parameter θ belongs to any given subset Θ_0 of the parameter space.

Another specific advantage of the Bayesian approach is in the way that it can deal with nuisance parameters. If the vector of parameters is written as $\theta = (\theta_1, \theta_2)$, say, where we wish to make inferences about the vector θ_1 alone, and θ_2 is a vector of nuisance parameters, in the classical approach we obtain the likelihood function $L(\theta_1, \theta_2; \mathbf{x})$, but we cannot obtain a likelihood function $L(\theta_1; \mathbf{x})$ for θ_1 on its own. But with the Bayesian approach we find the posterior density $\pi(\theta_1, \theta_2|\mathbf{x})$ and integrate to obtain the marginal posterior density of θ_1:

$$\pi(\theta_1|\mathbf{x}) = \int \pi(\theta_1, \theta_2|\mathbf{x}) d\theta_2 \ .$$

On the negative side, some critics of the Bayesian approach maintain that a properly scientific method aims to separate objective evidence from subjective preconceptions, whereas the Bayesian approach wholeheartedly embraces subjectivity. However the ideal of scientific objectivity is itself open to question: there is a subjective element in all scientific research, so why not be open about it?

A serious difficulty of the Bayesian approach lies at the first step, which may be regarded as too simplistic. Is there not more to our beliefs and uncertainties than can be expressed in terms of mathematical probabilities? Is it necessarily possible to quantify one's subjective beliefs about a parameter θ in the form of a probability distribution? We saw, for example, that even the representation of complete ignorance is problematic, although a Bayesian rejoinder to this point might be that the concept of complete ignorance is a false one, since we shall always have some prior idea, however vague, about the possible values of a parameter.

Another issue for a purely Bayesian approach to inference is that, just as in the frequentist approach discussed previously, the theoretical framework for the data, i.e., the model, as expressed in terms of the family of distributions $\mathcal{F}$, is taken as fixed. In the Bayesian approach, given the observed data $\mathbf{x}$, our changes in belief are changes in belief about θ, expressed solely in terms of the change from the prior $\pi(\theta)$ to the posterior distribution $\pi(\theta|\mathbf{x})$. But another aspect of inference is that the observation of the data $\mathbf{x}$ may also result in our deciding that the model $\mathcal{F}$ that we had chosen to use was inappropriate, so that we are led to change the model. So inference cannot simply be reduced to the Bayesian paradigm that takes us from $\pi(\theta)$ to $\pi(\theta|\mathbf{x})$.

5.7 Predictive distributions

Suppose that $\mathbf{x} \equiv (x_1, x_2, \ldots, x_n)$ represents a random sample of observations and suppose that another observation X_{n+1} is to be taken from the density $f(x_{n+1}|\theta)$, where f now represents the marginal p.d.f. of a single observation. The *predictive distribution* of X_{n+1} given $\mathbf{x}$ is the conditional distribution $f(x_{n+1}|\mathbf{x})$ of X_{n+1} given $\mathbf{x}$. Using the formula for a conditional p.d.f.,

$$\begin{aligned} f(x_{n+1}, \theta|\mathbf{x}) &= f(x_{n+1}|\theta, \mathbf{x})\pi(\theta|\mathbf{x}) \\ &= f(x_{n+1}|\theta)\pi(\theta|\mathbf{x}), \end{aligned} \tag{7}$$

where we are assuming that, for any given θ, X_{n+1} is independently distributed of $\mathbf{X}$. Hence, integrating out θ in Equation (7),

$$f(x_{n+1}|\mathbf{x}) = \int_{\Theta} f(x_{n+1}|\theta)\pi(\theta|\mathbf{x})d\theta.$$

Equivalently, if T is a sufficient statistic and $T(\mathbf{x}) = t$, the predictive distribution is given by

$$f(x_{n+1}|t) = \int_{\Theta} f(x_{n+1}|\theta)\pi(\theta|t)d\theta. \tag{8}$$

Example (continued)

In the case of a single Bernoulli trial, we have $f(1|\theta) = \theta$. Substituting into Equation (8), we obtain

$$\mathbb{P}(X_{n+1} = 1|t) \equiv f(1|t) = \int_{\Theta} \theta\, \pi(\theta|t) d\theta = E[\theta|t].$$

Thus, given the results of the first n in a sequence of trials, the Bayes estimate $E[\theta|t]$ is the posterior probability that the result of the next trial will be a "success".

The Bayes estimator is the mean of the beta distribution with parameters $t + a$ and $n - t + b$, i.e.,

$$E[\theta|t] = \frac{t+a}{n+a+b}.$$

In the special case of the uniform prior distribution with $a = 1$ and $b = 1$, we have

$$E[\theta|t] = \frac{t+1}{n+2}.$$

In the further special case when all n trials have yielded successes, so that $t = n$, we have

$$E[\theta|t] = \frac{n+1}{n+2}.$$

This result is sometimes referred to as *Laplace's law of succession* — the probability of a success at the next trial if all of n previous trials have been successes.

Returning to the general case with arbitrary a, b and t, it is worth noting that the Bayes estimator is a weighted average of the natural estimate t/n (the MLE and the MVUE) of θ based on the data alone and of the mean $a/(a + b)$ of the prior distribution. This is made explicit in the following identity:

$$\frac{t+a}{n+a+b} = \left(\frac{n}{n+a+b}\right)\left(\frac{t}{n}\right) + \left(\frac{a+b}{n+a+b}\right)\left(\frac{a}{a+b}\right). \tag{9}$$

We see from Equation (9) how prior beliefs, as summarized in the mean of the prior distribution, and the evidence from the data, as summarized in the maximum likelihood estimate, are combined to produce the mean of the posterior distribution.

The variance of the beta distribution with parameters a and b is

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

The larger the value of $a + b$ the more precise is the prior distribution, and $a + b$ may be regarded as a measure of the strength of the prior beliefs; but the information in the data is proportional to the sample size n. These considerations are reflected in the fact that in Equation (9) the weights given to the data and to the prior distribution are in the ratio of n to $a + b$.