Comparative Statistical Inference

Third Edition

Vic Barnett

Professor of Environmental Statistics
University of Nottingham, UK

© 1999

JOHN WILEY & SONS, LTD
Chichester • New York • Weinheim • Brisbane • Singapore • Toronto

The notion of robustness is central to all approaches to inference—in the sense of the relative insensitivity of inferences to assumptions about the underlying probability structure (e.g. as expressed by model, prior beliefs). Conigliani et al. (1994) discuss and compare classical and Bayesian approaches to this issue.

Returning to the general topic of inter-comparison, we find that arguments sometimes turn out to be circular: conclusions on one approach (based on its own premises) contradict the *premises* of another approach and *vice versa*. This is the situation in much of the cross-criticism of different approaches to statistical analysis, as we have witnessed, for example, in the discussion of the strong likelihood principle. All this process of argument is constructive provided that attitudes are not allowed to harden: that minds remain open, and different methods are used in different situations with an honest desire to assess their value, uncluttered by inappropriate philosophical preconceptions.

CHAPTER 6

Bayesian Inference

6.1 THOMAS BAYES

The Rev. Thomas Bayes' contribution to the literature on probability theory consists of just two papers published in the *Philosophical Transactions* in 1763 and 1764 (see Bayes, 1963a and 1963b). Of these, the first, entitled 'An essay towards solving a problem in the doctrine of chances', is the one that has earned Bayes a crucial place in the development of statistical inference and in its present-day practice. Both papers were published after his death, having been communicated to the Royal Society by his friend Richard Price who added his own introduction, comment and examples on the work. The second paper was concerned with some further details on the derivation of a particular result in the 'Essay'. Bayes' principal contribution, the use of **inverse probability** was further developed later by Laplace.

There is still some disagreement over precisely what Bayes was proposing in the 'Essay'. Two ideas can be distinguished and these are described by various names in the literature. For present purposes, we refer to them as **Bayes' theorem**, and his **principle of insufficient reason**. Bayes' reasoning was informal, in the spirit of the age, and the modern expression of these two ideas has been constructed out of the mere hints he provided, and the examples given by him and Price to illustrate them. It is in connection with this current formalisation that most dispute arises, notably in two respects:

(i) whether the concept of *inverse probability* (stemming from Bayes' theorem) is presented as a general inferential procedure even when it implies, or demands, a degree-of-belief view of probability;

(ii) how universally Bayes intended the *principle of insufficient reason* to be applied, and whether it provides a description of the state of **prior ignorance**.

This dispute on Bayes' intentions need not concern us here. It has been amply discussed elsewhere: an individual view is given by Hogben (1957, pp. 110–132). Also, a more accessible, slightly edited, version of the 'Essay' and Price's 'Appendix' has been provided, with bibliographical notes by Barnard (1958), from which the reader may reach his own conclusions. What seems to be widely agreed is that Bayes' work is noteworthy in three respects; in his use of continuous rather than discrete frameworks, in pioneering the idea of inference (essentially estimation) through assessing the chances that an 'informed guess' about the practical situation will be correct, and in proposing a formal description of what is meant by prior ignorance. On this latter point, see Edwards (1978). A history of inverse probability (and of the Bayesian approach) from Bayes to Karl Pearson is given by Dale (1991).

Bayes' ideas act as the springboard for the modern approach to inference known as *Bayesian inference*. This current expression of the earlier work was a long time in appearing (nearly 200 years) and we have considered possible reasons for this elsewhere (Chapter 1). Also, many might claim that Bayes would have difficulty in recognising his own tentative offerings in the wealth of detail and interpretative sophistication of modern *Bayesian inference*. Nonetheless, the seeds of this approach were certainly sown by Bayes 200 years ago, and the dispute over the meaning of his ideas has now been transferred to the analogous concepts in their offspring. There is little point in going further into the details of the 'Essay'. We proceed instead to describe some of the principles and techniques of *Bayesian inference*, its position within the different approaches to statistical inference and decision-making, and some of the external and internal controversy concerning basic concepts and criteria.

Introductory presentations of Bayesian inference are given in the books by Lindley (1965a, 1965b), Winkler (1972b), Iverson (1984), Lee (1989), Press (1989) and Berry (1994). Box and Tiao (1973) discuss Bayesian methods with particular application to regression and analysis of variance.

More advanced treatments are offered by Lindley (1972), Hartigan (1983), Bernardo and Smith (1994) and O'Hagan (1994).

Many treatments of the subject (for example, Winkler, 1972; Bernardo and Smith, 1994; and O'Hagan, 1994) include inter-comparison of Bayesian and classical approaches.

Much of the literature on the Bayesian idiom examines its extension to decision-making through the vehicle of *decision theory*, which we discuss in Chapter 7. These include Lindley (1971b and 1985), French (1986, 1989), Bernardo and Smith (1994) and O'Hagan (1994). Many treatments of decision theory (*per se*) include coverage of Bayesian ideas: Raiffa and Schlaifer (1961), De Groot (1970)' and French (1986 and 1989).

There is widespread coverage of the use of Bayesian methods in specific fields of application such as *actuarial science* (Klugman, 1992), *biostatistics* (Berry and Stangl, 1994), *economics* (Zellner, 1971, 1985; Cyert and De Groot, 1987; and Geweke, 1996), *education and psychology* (Pollard, 1986) and *social science* (Phillips, 1973).

A comprehensive review of the literature is given by Bernardo and Smith (1994, pp. 9–11).

Publications on specific practical problems approached by Bayesian methods are legion. Some recent examples include Crome et al. (1996; birds and small mammals in rain forests), Taylor et al. (1996; classifying Spectacled Eiders), Calabria and Pulchini (1996; failure data in repairable systems), Sinsheimer et al. (1996; molecular biology) and Wakefield (1996; pharmacokinetic models). Earlier applied studies by Mosteller and Wallace (1964, 1984; on an authorship of published material), Freeman (1976; on the existence of a fundamental unit of measurement in megalithic times) and by Efron and Thisted (1976; 'how many words did Shakespeare know?') remain of interest. See also Datta and M. Ghosh (1995b) on estimating the error variance in a one-way designed experiment, and Garthwaite et al. (1995) on the Bayesian approach to capture–recapture.

6.2 THE BAYESIAN METHOD

We start with what is known as **Bayes' theorem**. At one level, this may be regarded as a result in deductive probability theory.

Bayes' theorem. *In an indeterminate practical situation, a set of events* A_1, $A_2, \ldots, A_k$ *may occur. No two such events may occur simultaneously, but at least one of them must occur (i.e.* $A_1, \ldots, A_k$ *are mutually exclusive and exhaustive). Some other event, A, is of particular interest. The probabilities,* $P(A_i)$ $(i = 1, \ldots, k)$, *of each of the* A_i *are known, as are the conditional probabilities,* $P(A|A_i)$ $(i = 1, \ldots, k)$, *of A given that* A_i *has occurred. Then, the conditional ('inverse') probability of any* A_i $(i = 1, \ldots, k)$, *given that A has occurred, is given by*

$$P(A_i|A) = \frac{P(A|A_i)P(A_i)}{\sum_{j=1}^{k} P(A|A_j)P(A_j)} \quad (i = 1, \ldots, k). \qquad (6.2.1)$$

As expressed, *Bayes' theorem* finds wide application, and arouses no controversy. No difficulty arises in the philosophical interpretation of the probabilities involved in it.

However, it is also central to *Bayesian inference*, and in this role it is necessary to extend its meaning in one particular respect. Rather than considering *events* A_i, we must work in terms of a *set of hypotheses* $\mathrm{H}_1, \ldots, \mathrm{H}_k$ concerning what constitutes an appropriate model for the practical situation. One, and only one, of these must be true. The event A becomes reinterpreted as an observed outcome from the practical situation: it may be thought of as the *sample data*. Prior to the observation, the probability, $P(\mathrm{H}_i)$, that H_i is the appropriate model specification, is assumed known for all $i = 1, \ldots, k$. These probabilities are the **prior probabilities** of the different hypotheses, and constitute a secondary source of information. The probabilities, $P(A|\mathrm{H}_i)$ $(i = 1, \ldots, k)$, of observing A, when H_i

is the correct specification, are known also—these are simply the *likelihoods* of the sample data.

We can re-interpret *Bayes' theorem* as providing a means of updating, through use of the sample data, our earlier state of knowledge expressed in terms of the prior probabilities, $P(\mathrm{H}_i)$ $(i = 1, \ldots, k)$. The updated assessment is given by the **posterior probabilities**, $P(\mathrm{H}_i|A)$ $(i = 1, \ldots, k)$, of the different hypotheses being true after we have utilised the further information provided by observing that A has occurred. These *posterior probabilities (inverse probabilities)* are given by

$$P(\mathrm{H}_i|A) = \frac{P(A|\mathrm{H}_i)P(\mathrm{H}_i)}{\sum_{j=1}^{k} P(A|\mathrm{H}_j)P(\mathrm{H}_j)} \quad (i = 1, \ldots, k). \tag{6.2.2}$$

This is the essence of Bayesian inference: that the two sources of information provided by the prior probabilities and the sample data (represented in the form of likelihood) are combined to produce the posterior probability *of* H_i *given* A, *which is proportional to the product of the prior probability of* H_i *and the likelihood of* A *when* H_i *is true*. The denominator of the right-hand side of (6.2.2) is merely a normalising constant independent of i, although its determination is of some importance and can be problematical as we shall see later.

Thus, prior information about the practical situation is in this way augmented by the sample data to yield a current probabilistic description of that situation. In this respect, the Bayesian approach is *inferential*. It asserts that our current knowledge is fully described by the set of posterior probabilities, $\{P(\mathrm{H}_i|A)\}(i = 1, \ldots, k)$.

It is interesting to note an immediate consequence of (6.2.2), providing a restatement of the fundamental principle of Bayesian inference. Suppose we are interested in two particular hypotheses, H_i and H_j. The ratio of their posterior probabilities—their **posterior odds ratio**—is given by

$$\frac{P(\mathrm{H}_i|A)}{P(\mathrm{H}_j|A)} = \frac{P(A|\mathrm{H}_i)}{P(A|\mathrm{H}_j)} \cdot \frac{P(\mathrm{H}_i)}{P(\mathrm{H}_j)}, \tag{6.2.3}$$

that is, by the product of the **prior odds ratio** and the *likelihood ratio.*

Example 6.2.1 A box contains equal large numbers of two types of six-faced cubical die. The dice are perfectly symmetric with regard to their physical properties of shape, density, etc. Type I has faces numbered 1, 2, 3, 4, 5, 6; whilst type II has faces numbered 1, 1, 1, 2, 2, 3. A plausible probability model is one which prescribes equal probabilities of $\frac{1}{6}$ for the uppermost face when any die is thrown. One die is picked at random from the box, thrown twice, and the uppermost face shows a 1 and a 3 on the separate (independent) throws. Having observed this, we wish to comment on whether the die was type I or type II. We denote these alternative possibilities as hypotheses $\mathrm{H_I}$ and $\mathrm{H_{II}}$, and there seem to be reasonable grounds for assigning equal prior probabilities to each of these. Bayes' theorem then yields posterior probabilities of $\frac{1}{4}$ and $\frac{3}{4}$ for the hypotheses $\mathrm{H_I}$ and $\mathrm{H_{II}}$, respectively. This conclusion is an inferential statement about the underlying practical situation. It describes the relative chances that the die used was of type I or type II, respectively, as 3 : 1 in favour of it being type II. In Bayes' original formulation we would declare that a 'guess' that the die is type II has probability $\frac{3}{4}$ of being correct.

The change of emphasis in (6.2.2) is well illustrated by this example. It is now a statement about the plausibility of alternative models for generating the observed data; no longer a deductive probability statement.

This re-interpretation raises some problems, which centre on the nature of the probability concept involved in the prior, and posterior, probabilities. In Example 6.2.1 the situation is straightforward. Both the prior and the posterior probabilities can be interpreted in frequency terms, within the larger experiment of choosing a die at random from the box containing equal numbers of each type. Also, the numerical values of the prior probabilities are derived directly from this 'super-experiment'. Their accuracy, of course, remains dependent on our assumptions about the super-experiment; that choice of the die to be used really is random from equal numbers of each type. The evaluation of the likelihoods also rests on the assumed randomness and independence of consecutive outcomes of throwing the die that is chosen. Strictly speaking, there also remains the question of validating the assumptions about the super-experiment, but *if we accept them* there is no difficulty in interpreting the results of the inverse probability statement derived from *Bayes' theorem.*

Example 6.2.1 is a simple, but typical, illustration of real-life situations, for example in genetics. In general, however, further complications can occur. Consider two other examples, both superficially similar to Example 6.2.1, both representative of practical situations, but each subtly different.

Example 6.2.2 A box contains large numbers of two types of six-faced cubical die. The dice are perfectly symmetric with regard to their physical properties of shape, density, etc. Type I has faces numbered 1, 2, 3, 4, 5, 6; whilst type II has faces numbered 1, 1, 1, 2, 2, 3. There are more type I dice in the box than type II. One die is picked at random from the box, thrown twice, and the uppermost face shows a 1 and a 3 on the separate (independent) throws. What can we say about whether this die was type I or type II?

Example 6.2.3 A friend has one of each of the types of die described in Examples 6.2.1 and 6.2.2 and chooses one of these without revealing its type, but comments that he would always use that type in preference to the other. The friend throws the die twice and reports that the uppermost face shows a 1 and a 3 on the separate (independent) throws. What can we say about what type of die was used?

In Example 6.2.2, there still exists some credible notion of a 'super-experiment'. The probabilities $P(\mathrm{H}_i)$ and $P(\mathrm{H}_i|A)$ $(i = \mathrm{I}, \mathrm{II})$ have corresponding frequency interpretations. But our information about the super-experiment ('more type I than type II dice') is now insufficient to assign precise numerical values to the prior probabilities, $P(\mathrm{H}_\mathrm{I})$ and $P(\mathrm{H}_\mathrm{II})$. We know only that $P(\mathrm{H}_\mathrm{I}) > P(\mathrm{H}_\mathrm{II})$, i.e. $P(\mathrm{H}_\mathrm{I}) > \frac{1}{2}$. To make inferences by Bayesian methods we are compelled to substitute numerical values for $P(\mathrm{H}_\mathrm{I})$ and $P(\mathrm{H}_\mathrm{II})$. How is this to be done?

We may have even less information; not knowing whether type I or type II dice are in the majority. The problem remains! What do we use for the values of $P(\mathrm{H}_\mathrm{I})$ and $P(\mathrm{H}_\mathrm{II})$? We might describe this latter state as one of *prior ignorance*, and to use Bayesian methods we must quantify this condition. *The principle of insufficient reason* suggests that we assume that $P(\mathrm{H}_\mathrm{I}) = P(\mathrm{H}_\mathrm{II}) = \frac{1}{2}$, in that there is no evidence to favour type I or type II. But $P(\mathrm{H}_\mathrm{I})$ is well defined in frequency terms; it has a specific value, albeit unknown. In declaring that $P(\mathrm{H}_\mathrm{I}) = \frac{1}{2}$ we are not really claiming that there are equal numbers of each type of die in the box. We are making a statement about our own attitude of mind; we cannot find grounds for believing, a priori, in there being a majority of one type of die rather than the other. The probability concept involved in the statement $P(\mathrm{H}_\mathrm{I}) = \frac{1}{2}$ has become a degree-of-belief one (either *subjective*, or *logical*).

In Example 6.2.3 even the super-experiment structure seems to have disappeared! Again on the principle of insufficient reason, we may make the conventional assignment $P(\mathrm{H}_\mathrm{I}) = \frac{1}{2}$, possibly with a *logical* interpretation of the probability concept. Alternatively, personal opinions may enter the problem. We may feel that the friend is somewhat eccentric, and more likely to choose the 'odd' type II die. Is this feeling relevant to the need to assign a value to $P(\mathrm{H}_\mathrm{I})$? Some would claim it is, and use a value of $P(\mathrm{H}_\mathrm{I}) < \frac{1}{2}$. Again, arbitrariness enters in deciding what *precise* value $\left(<\frac{1}{2}\right)$ to give to $P(\mathrm{H}_\mathrm{I})$.

Someone else faced with the same problem may regard the friend as a rather 'conservative' person, more inclined to choose the type I die, and would choose a value for $P(\mathrm{H}_\mathrm{I}) > \frac{1}{2}$. The prior probabilities may now need to be interpreted in *personal* terms, rather than logical, in the sense of the distinction drawn in Chapter 3.

We have pinpointed two difficulties:

(i) the interpretation of the probability idea implicit in a particular Bayesian analysis,

(ii) the numerical specification of prior probabilities to be used in the analysis.

The first difficulty, (i), is at the heart of the criticisms that are made of the Bayesian approach in general, and of the internal divisions of attitude that exist. We take this up again in Section 6.8. We must recognise, however, that in the ever-expanding use of Bayesian methods for modelling or analysing practical problems, the problems of incompletely specified prior views and subjective/degree-of-belief expressions of probability (as illustrated by Examples 6.2.2 and 6.2.3) often arise and must be accommodated. One thing is clear, then; we shall not proceed far in studying the application of Bayesian methods unless we admit that a wider view of probability may be necessary than the frequency one. We shall consider some detailed examples later.

Lindley (1965b) chose to present Bayesian inference entirely in degree-of-belief terms 'for it to be easily understood'.

In a leading modern treatment of Bayesian inference, Bernardo and Smith (1994, pp. 2 and 4) remark:

> ... we shall adopt a whole hearted subjectivist position regarding the interpretation of probability. ... Bayesian statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty ...

It is not necessary to be so specific for much of the following discussion. To demonstrate principles at an elementary level, the term 'probability' may often be used in an intuitive way, leaving its interpretation dependent on the nature of any problem being studied and on the attitude of the investigator. But when trying to *interpret* ideas in Bayesian inference, this intuitive approach often will not do. It does seem (in agreement with Lindley) to be simplest to adopt a degree-of-belief attitude in such cases, and this is what has been done below. This is not to suggest, however, that there may not on occasions, be a perfectly valid and direct frequency interpretation in certain situations (but see Section 6.8.1)

The second difficulty, (ii), cannot be 'sidelined'. Its resolution forms an essential part of the practical detail of Bayesian inference. The need to process prior information to yield numerical values for prior probabilities is central to Bayesian inference and can be fraught with difficulties, which are still not fully resolved beyond the relatively rare instances of fully specified objective prior information. However, even the apparently simple case where the prior information consists of sample data from earlier observations of the same practical situation is not without its difficulties. (See Section 6.7 on **empirical Bayes' methods.**) We consider briefly what results are available on this matter in Section 6.4, 6.5 and 6.6, which distinguish between the cases of **prior ignorance, vague prior knowledge**, and **substantial prior knowledge**, respectively. For the moment, however, we will take the provisional attitude that numerical-valued prior probabilities are available, and enquire in what way they are used in Bayesian inference.

6.3 PARTICULAR TECHNIQUES

The *rationale* of Bayesian inference may be summarised as follows.

Inferences are to be made by combining the information provided by prior probabilities with that given by the sample data; this combination is achieved *by 'the repeated use of Bayes' theorem' (Lindley, 1965*b, p. xi*), and the final inferences are expressed solely by the posterior probabilities.*

To see how this works out, we revert to the parametric model used earlier, but some extra care is needed in its description.

We suppose that sample data, x, arise as an observation of a random variable, X. The distribution of X, specified by the probability model, is assumed to belong to some family, $\mathscr{P}$, indexed by a parameter θ. It is assumed that the probability (density) function of the random variable has a known form, $p_\theta(x)$, depending on θ; but that θ is unknown, except that it lies in a parameter space, Ω. For fixed x, $p_\theta(x)$ is, of course, the likelihood function of θ. Knowledge of the 'true' value of θ, i.e. the value relevant to the current practical situation, would be all that is needed to describe that situation completely. (This assumes, of course, the adequacy of the family $\mathscr{P} = \{p_\theta(x); \theta \in \Omega\}$ as a general model for the practical situation. This may need to be examined separately.)

As earlier, both θ and x may be either univariate or multivariate, discrete or continuous. In the *classical approach*, multivariate data and multi-parameter models required special treatment (see Section 5.3.4). In the Bayesian approach, however, they have possibly an even greater importance in two respects. It is almost inevitable that any practical problem of substance will require a higher-dimensional parametric specification for a Bayesian analysis than for the use of classical methods. Such a multi-parameter specification (often involving tens if not hundreds of parameters) poses in turn severe problems for determining the required posterior probability distributions. Much of the modern thrust of Bayesian methodology has been directed to such problems; we shall later consider such methods as the Gibbs sampler and Markov chain Monte Carlo procedures as part of this developmental drive (see Section 6.7).

The aim of any inferential procedure must be to 'reflect' the unknown 'true' value of θ. This cannot be done with any certainty. All we can expect is some probabilistic statement involving θ, based on the information at our disposal. In the *classical approach* of Chapter 5, we saw how this was achieved by processing the *sample data*, as the only available information, to produce point or interval estimates of θ, with associated assessments of their accuracy. In *Bayesian inference* it is assumed that we have further information available a priori, i.e. before observing the sample data. To incorporate this, a wider view is taken of the nature of the parameter θ. It is assumed that any knowledge we have of the 'true' value of θ, at any stage, can be expressed by a probability distribution, or some 'weight function', over Ω. *The parameter θ is now essentially regarded as a random variable* in the sense that different values are possible with different probabilities, degrees-of-belief or weights! O'Hagan (1994, p. 11) is explicit on this issue:

> . . . in Bayesian inference the parameters are random variable.

(We consider the implications of this in more detail in Section 6.8.1.)

The prior knowledge of θ is expressed as a *prior probability distribution*, having probability (density) function, $\pi(\theta)$. Sampling increases this knowledge, and the combined information about θ is described by its *posterior distribution*. If the posterior probability (density) function of θ is denoted by $\pi(\theta|x)$, we have, from Bayes' theorem,

$$\pi(\theta|x) = p_\theta(x)\pi(\theta) \Big/ \int_\Omega p_\theta(x)\pi(\theta). \qquad (6.3.1)$$

The posterior distribution, $\pi(\theta|x)$, is regarded as a complete description of what is known about θ from the prior information and the sample data; it describes how we now assess the chances of the true value of θ lying in different parts of the parameter space Ω.

Leaving aside the question of the interpretation of the probability concept involved in $\pi(\theta)$ and $\pi(\theta|x)$, we now appear to have a more direct form of inference than in the classical approach. We can answer directly such questions as 'what is the probability that $0.49 < \lambda < 0.51$?' in the radioactivity example of Section 1.3. This facility is not available in the classical approach, without reference to some larger framework than the current situation and its associated information. It was there necessary to consider sampling distributions defined in terms of a sequence of independent repetitions of the current situation under what were assumed to be identical circumstances. The effect of this, in providing only aggregate measures of accuracy, has been considered in some detail in Chapter 5.

In contrast, inferences in the Bayesian approach contain their own internal measure of accuracy and relate only to the current situation. This distinction is crucial. It is the difference between *initial precision* and *final precision* (see Section 5.7.1). Classical statistics assesses initial precision; Bayesian inference, final precision. We accept this distinction for the moment at it's face value and will consider some examples of it. Later (Section 6.8.1), however, we will need to re-examine it in the light of a closer study of the '*random*' nature of θ.

Whilst the posterior distribution, $\pi(\theta|x)$, constitutes the complete inferential statement about θ, there are situations where such a full description is not needed. Certain *summary measures* of $\pi(\theta|x)$ may suffice. For example, it may be enough to know what value of θ is *most likely*, or in what region θ is *highly likely to fall*, or even whether some specific statement about θ is *credible*. These needs lead to concepts in Bayesian inference parallel with the ideas of point estimates, confidence regions and hypothesis tests in the classical approach. It is convenient to describe these by the same (or similar) names, but it must be remembered that their interpretation, and the numerical answers they provide, are likely to be different from those of the analogous quantities in classical inference.

For illustrative convenience, θ is assumed to be scalar at this stage. We will consider some special features of multi-parameter Bayesian inference later (see Section 6.7).

Bayesian Point Estimates

There may be situations where it is convenient to choose a single value as an estimate (a point estimate) of θ. It seems sensible to choose that value with highest posterior probability (density). Consequently, we define as a **Bayesian point estimate** the quantity $\theta(x)$ that maximises $\pi(\theta|x)$. See Figure 6.3.1.

In itself, $\tilde{\theta}(x)$ has limited practical value, although it has a direct interpretation as the most likely value for θ *in the current situation*. There would seem to be no other useful criterion for choosing a single value to estimate θ than to use the most likely value (the *mode* of the posterior distribution), unless we incorporate further information on the consequences of incorrect choice of θ. (See Section 7.3.)

In classical inference, the situation was quite different. A variety of alternative estimators may exist for θ. These will have the general form $\tilde{\theta}(X)$, stressing that their interpretation is in terms of different sets of data that may potentially arise for the *fixed* current value of θ, rather than in terms of differing degrees-of-belief about θ for the present observed data x. Choice between alternative estimators, and assessments of their individual properties (bias, efficiency, etc.), were all derived from the behaviour of this sampling distribution of $\tilde{\theta}(X)$; that is, in relation to repeated realisations of the current situation.

Although the criterion for the choice of the Bayesian estimate $\tilde{\theta}(x)$ seems incontrovertible, one difficulty does arise. The estimate will not be invariant with respect to transformations of the parameter space. Working in terms of $\phi(\theta)$, rather than θ, it does not necessarily follow that $\tilde{\phi} = \phi(\tilde{\theta})$ for any particular x. Thus, different inferences may be drawn from the same data and prior information in alternative parameterisations of the model. We recall also that not all characteristics of estimators on the classical approach are invariant to reparameterisation, e.g. if $\tilde{\theta}(x)$ is unbiased, $\tilde{\phi}(x)$ need not be. In contrast, maximum likelihood estimators are invariant (see Section 5.3.3).

This problem is sometimes resolved in the Bayesian approach by claiming that there will usually be a 'natural' parameterisation, and that inferences must therefore relate to this 'natural' parameter. This does not seem completely satisfactory, and all summary measures of the posterior distribution remain somewhat arbitrary in this respect. We will meet this concept of a 'natural' parameter again in relation to ways of describing prior ignorance (Section 6.4).

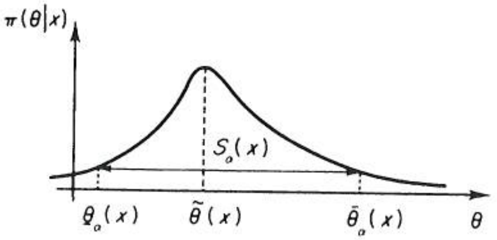

Figure 6.3.1

Bayesian Interval or Region Estimates (Credible Intervals or Regions)

A more informative summary of $\pi(\theta|x)$ for practical purposes is obtained by saying that θ lies in some region of Ω with a prescribed probability. The region $S_\alpha(x)$ is a $100(1-\alpha)$ per cent **Bayesian posterior credible region**, or **Bayesian posterior probability region**, for θ if

$$\int_{S_\alpha(x)} \pi(\theta|x) = 1 - \alpha. \qquad (6.3.2)$$

If θ is scalar and $S_\alpha(x)$ is a connected region in R^1, we refer to $S_\alpha(x)$ as a posterior credible interval, or a posterior probability interval.

If $S_\alpha(x)$ can be chosen to satisfy (6.3.2) we might call $(1-\alpha)$ the **posterior credibility coefficient**. As in the classical case, we may not be able to achieve this precisely, but can only ensure that $P[\theta \in S_\alpha(x)] \geq 1 - \alpha$. Then again, we might call $(1-\alpha)$ the **posterior credibility level**. This occurs in cases where θ has a discrete component—in the classical approach it happened if X had a discrete component.

In (6.3.2) we might use $\pi(\theta)$, the prior probability (density) function, instead of $\pi(\theta|x)$. We would then obtain corresponding **prior credible regions** or **intervals** (or **prior probability regions** or **intervals**). The idea extends also to predictive distributions of X (see Section 6.4).

To be of practical value, α will need to be chosen small and we will again typically consider 90, 95, 99 per cent, etc. credible regions, although the actual choice is arbitrary. The region $S_\alpha(x)$ will not be unique, again in parallel with the classical confidence region. There may be several regions that contain a proportion $(1-\alpha)$ of the posterior distribution, often an infinite number, and it is necessary to choose between them. In the classical approach we saw (Section 5.5) that the so-called *central confidence intervals* had certain optimality properties in special cases, and the idea of central intervals (cutting off equal tail-area probabilities) was extended to form a practical criterion for common use. This concept seems to have little relevance to the Bayesian situation, however. Any region of Ω omitted on this criterion may have small total probability, but can still contain values of θ with high probability (density) relative to some values of θ contained in $S_\alpha(x)$; we are, perhaps, excluding some θ that are more likely to be true than other θ included in $S_\alpha(x)$. Consequently, a different criterion is usually adopted: *$S_\alpha(x)$ shall not exclude any value of θ more probable than any value of θ that is included*. It seems inevitable, on the Bayesian idiom, that this criterion should always be applied if the Bayesian confidence region is to be properly interpretable.

Such regions or intervals are known as **highest probability (density)**, or **HPD, intervals** or **regions**. See Bernardo and Smith (1994, Section 5.1.5).

In situations where θ is scalar and continuous, we might find that $\pi(\theta|x)$ will be unimodal, and this principle then yields a finite interval for $S_\alpha(x)$: a **Bayesian HPD interval**. See Figure 6.3.1. Here, $S_\alpha(x)$ takes the form $(\underline{\theta}_\alpha(x), \overline{\theta}_\alpha(x))$, where

$$\int_{\underline{\theta}_\alpha(x)}^{\overline{\theta}_\alpha(x)} \pi(\theta|x) = 1 - \alpha \tag{6.3.3}$$

and

$$\pi(\theta|x) \geq \pi(\theta'|x), \quad \text{for any } \theta \in S_\alpha(x), \theta' \notin S_\alpha(x).$$

Under fairly general conditions $(\underline{\theta}_\alpha(x), \overline{\theta}_\alpha(x))$ is both unique—except for cases where several values of θ have equal probability (density)—and shortest amongst all Bayesian confidence regions of prescribed Bayesian confidence coefficient. But again, variations may occur in alternative parameterisations!

Again, these ideas extend to prior, or predictive, probability assessments.

In the classical approach, we encountered an optimality concept for confidence intervals in terms of the *uniformly most accurate* (UMA) confidence interval: a dual notion to that of a *uniformly most powerful* (UMP) hypothesis test. On the Bayesian approach, optimality is represented by the idea of **minimal size credible regions**, derived through decision theory argument for an appropriate form of loss structure (see Bernardo and Smith, 1994, Section 5.1.5, and Section 7.3.4 below).

The difference of interpretation of the classical, and Bayesian, concepts is obvious and striking. Within its framework, the Bayesian credible region has a *direct probability interpretation*,

$$P[\theta \in S_\alpha(x)] = 1 - \alpha, \tag{6.3.4}$$

unique to, and determined solely from, the current data x, and prior information. The classical confidence region is also expressed merely in terms of the current data, but its probability interpretation is as a *random* region containing the *fixed* value of θ. The assessment of its probability of actually enclosing θ is in terms of repetitions of the experimental situation. As we saw earlier (Section 5.5) there is no way of judging whether a *particular* classical confidence region does, or does not, include θ.

Example 6.3.1 Suppose a random sample of n independent observations is available from a normal distribution, $\mathbf{N}(\mu, \sigma^2)$, with unknown mean μ, and known variance σ^2. (This latter assumption is introduced for convenience, hardly in pretence of reality.) The sample mean is $\bar{x}$. Anticipating ideas in the next section, we suppose that nothing is known a priori about μ and that this is expressed by an ('improper') assignment of equal prior probability density over $(-\infty, \infty)$. It is easy to show that (6.3.1) yields for the posterior distribution of μ the normal distribution, $\mathbf{N}(\bar{x}, \sigma^2/n)$. Thus, we obtain the Bayesian point estimate, $\bar{x}$, for μ. Furthermore, the $100(1-\alpha)$ per cent Bayesian credible interval for μ has the form $(\bar{x} - z_\alpha\sigma/\sqrt{n}, \bar{x} + z_\alpha\sigma/\sqrt{n})$, where z_α is the double-tailed α-point of $\mathbf{N}(0,1)$.

We know, however, that on the classical approach $\bar{x}$ and $(\bar{x} - z_\alpha\sigma/\sqrt{n},\ \bar{x} + z_\alpha\sigma/\sqrt{n})$, are also the optimum point estimate and $100(1-\alpha)$ per cent confidence interval, respectively, for μ.

This correspondence in the explicit expressions should not be allowed to conceal the quite different meanings of the results in the two cases. Neither should it be taken as indicating any general area of unanimity. We shall obtain identical expressions only with appropriate choice of the prior distribution. It is intriguing to question whether any such agreement, for cases of prior ignorance, lends respectability to the classical approach in that it produces the same answers as the Bayesian approach, or vice versa, or whether instead it serves to justify the particular expression used to represent prior ignorance. All three cases have been argued separately in the literature. Whether any of them has any justification depends on the personal attitudes of their proponents, and can hardly be assessed objectively. (See Section 6.4)

Example 6.3.2 A random variable X has a Cauchy distribution centred at an unknown point, θ. Its probability density function is

$$f(x) = \frac{1}{\pi} \frac{1}{[1 + (x-\theta)^2]} \quad (-\infty < x < \infty).$$

Two independent observations x_1, x_2 are drawn from this distribution and constitute the sample data, x; they are to be used in drawing inferences about θ. The likelihood of the sample is

$$p_\theta(x) = \frac{1}{\pi^2} \frac{1}{[1 + (x_1-\theta)^2]} \cdot \frac{1}{[1 + (x_2-\theta)^2]}.$$

On both the classical and Bayesian approaches, some rather strange results arise when we try to draw inferences about θ. The sample mean $\bar{x}$ has intuitive appeal, but is inconsistent (in classical terms). Its sampling distribution has infinite mean and variance, and cannot lead to classical confidence intervals for θ. Maximum likelihood is no better! If x_1 and x_2 differ by more than 4, as is highly probable, $p_\theta(x)$ is symmetric about θ **but double peaked**. Typically it looks like Figure 6.3.2.

The sample mean is now at a relative **minimum** of $p_\theta(x)$. The two points $\hat{\theta}_1$ and $\hat{\theta}_2$ where $p_\theta(x)$ has relative maxima, are

$$\hat{\theta}_2, \hat{\theta}_1 = \bar{x} \pm \tfrac{1}{2}\sqrt{[(x_1 - x_2)^2 - 4]},$$

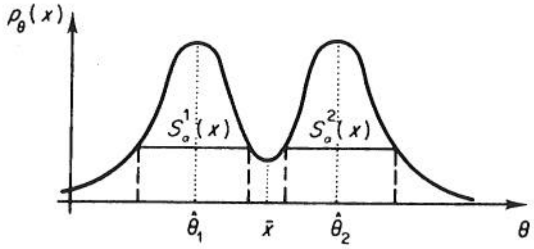

Figure 6.3.2

and $p_\theta(x)$ is equal at these points, so no unique maximum likelihood estimator exists. In terms of likelihood, there seems no good reason to distinguish between $\hat{\theta}_1$ and $\hat{\theta}_2 \cdot \bar{x}$ (or the sample median, which is the same thing here) has some intuitive appeal, but no statistical justification in view of its sampling properties.

On the Bayesian approach, again using an improper uniform distribution to express prior ignorance, the posterior distribution $\pi(\theta|x)$ is proportional to $p_\theta(x)$ so it has the same form as that shown above. No unique Bayesian point estimate exists, since $\pi(\hat{\theta}_1|x) = \pi(\hat{\theta}_2|x)$. One possibility might be to estimate θ by a value for which there is equal posterior probability of the true value being in excess of, or less than, this value. This is just the median of the posterior distribution, here $\bar{x}$. But it is surely not desirable on the Bayesian approach to estimate θ by what is one of the lowest posterior probability (density) points! Of course, we can construct a Bayesian credible region for θ, but even this is somewhat strange in that it is quite likely to have 'a hole in the middle', e.g. $S_\alpha(x) = S_\alpha^1(x) \cup S_\alpha^2(x)$ in the figure above. This illustrates a general point: that Bayesian credible regions, even for a single parameter, may consist of a set of intervals, rather than being a single interval.

This is an extreme example, of course. No one would expect to say much about the location parameter of a Cauchy distribution from two observations. But the anomalous behaviour it demonstrates is not entirely artificial. Similar difficulties can arise, in varying degrees, in quite realistic situations.

We should note that, as in many areas of Bayesian inference, the derivation of an HPD region, even for a scalar parameter, can involve extensive computation. This is compounded for the more realistic situations where θ is multi-dimensional. See Section 6.6 below for more discussion of the evaluation of posterior probabilities.

Bayesian Hypothesis Tests

A major element of the classical approach is hypothesis testing (Section 5.4). It is natural to enquire what corresponds with this in Bayesian inference. At the basic level of wishing to assess composite hypotheses about a continuous scalar parameter, the direct probability interpretation of the posterior distribution leads to a particularly simple form of **Bayesian hypothesis test**.

In a practical situation, we may need to assess whether some statement about θ lying in a particular region of Ω is reasonable. For example, a biscuit manufacturer producing packets of biscuits is required to state on the packet a weight that is at least 10 g less than the mean weight, θ, of packets of biscuits produced by the manufacturing process. The aim is to produce 200 g packets so a weight of 190 g is quoted on the packet. It is necessary to examine whether the requirement of not stating a weight in excess of $\theta - 10$ is being met. On the basis of some sample data and prior knowledge about θ, a posterior distribution $\pi(\theta|x)$ expressing current knowledge of θ is obtained.

This information must be used to test the hypothesis $\mathrm{H} : \theta < 200$ against the alternative hypothesis $\mathrm{H} : \theta \geq 200$. But on the Bayesian approach, we have a direct evaluation of the probabilities of H and $\overline{\mathrm{H}}$, in the form

$$P(\mathrm{H}|x) = \int_{\mathrm{H}} \pi(\theta|x) = 1 - P(\overline{\mathrm{H}}|x). \qquad (6.3.5)$$

If $P(\mathrm{H}|x)$ is sufficiently small we will want to reject H, and this can again be expressed in terms of '*significance levels*' if required. Thus, a 5 per cent Bayesian *test of significance* rejects H if $P(\mathrm{H}|x) \leq 0.05$; similarly for other significance levels. Alternatively, the outcome of the test may again be expressed in terms of the significance level actually attained, $P(\mathrm{H}|x)$; this is just the '*critical level*' of the test.

Note that the direct expression of the result of the test in the form $P(\mathrm{H}|x)$ eliminates the asymmetric nature of the test observed in the classical approach. In particular, there is no need to distinguish formally between the *null (working)* and the *alternative* hypotheses.

The test just described is a one-sided test of composite hypotheses in our earlier terminology. The idea of a two-sided test of a simple (or composite) hypothesis might also be considered in Bayesian inference, but opinions differ on its appropriate form, and difficulties of interpretation arise.

Much of the literature of the 1960s stressed the presence, within the Bayesian approach, of a facility that paralleled that of the classical hypothesis test. Whilst the above form of one-sided test seems straightforward, early attempts to formulate an equivalent two-sided test do not seem to have been so successful. However, it is of interest to consider some of the original proposals in this matter before examining present attitudes.

Lindley (1965b, pp. 58–62) describes one form of such a test in terms very similar to those used to construct the classical hypothesis test. To test $\mathrm{H} : \theta = \theta_0$ against $\overline{\mathrm{H}} : \theta \neq \theta_0$ at a level α he suggests that we obtain the $100(1 - \alpha)$ per cent Bayesian posterior credible interval for θ and accept H if this interval contains θ_0, otherwise reject H. This procedure is limited to cases where the prior information on θ is vague (see Section 6.5): and, in particular, where there is no prior discrete concentration of probability at $\theta = \theta_0$. We do not have any simple probability interpretation of the result of the test *per se*: we cannot talk about $P(\mathrm{H})$. There is only the inferred interpretation arising from the credible interval being an interval within which we have probability $(1 - \alpha)$ that the true θ lies—and hence only probability α that θ lies outside this interval. This is somewhat analogous to the tail-area concept of probability on the classical approach. The situation becomes more confused when one-sided tests are constructed in a similar way from *one-sided credible intervals*, which seems to have little meaning in the Bayesian approach!

Another method for testing a point hypothesis, $\mathrm{H} : \theta = \theta_0$, is described by Jeffreys (1961, Chapter 5). The idea here is that the particular value θ_0 has a different order of importance to the other values of θ in Ω. It achieves this through having been singled out for particular study, usually due to extraneous practical considerations. As a result, our prior information about θ should be in two parts; a prior discrete probability for θ_0, $P(\theta_0)$, together with some prior distribution $\pi(\theta)$ over $\theta \neq \theta_0$. The sample data, x, refine the probability that $\theta = \theta_0$, and the distribution of probability over $\theta \neq \theta_0$, to produce $P(\theta_0|x)$ and $\pi(\theta|x)$ for $\theta \neq \theta_0$. The decision on whether to accept, or reject, H is now taken on the basis of the posterior 'odds in favour of H', i.e.

$$P(\theta_0|x) \Big/ \int_{\Omega-\theta_0} \pi(\theta|x).$$

Jeffreys proposes that in the absence of any prior knowledge about θ we should divide the prior probability equally between H and $\overline{\text{H}}$, by taking $P(\theta_0) = \frac{1}{2}$ and a uniform distribution of the remaining probability mass over the values of $\theta \neq \theta_0$. He states, 'In practice there is always a limit to the possible range of these values'.

An interesting illustration of this procedure is given by Savage et al. (1962, pp. 29–33) in the context of the legend of King Heiro's crown. Pearson (1962a) offers a detailed critical re-examination of this example, with particular emphasis on its subjective elements.

Pratt (1976) declares that whilst the posterior probability that $\theta \leq \theta_0$ is a reasonable measure of the 'plausibility' of a null (working) hypothesis $\text{H} : \theta \leq \theta_0$, tested against the alternative hypothesis $\overline{\text{H}} : \theta > \theta_0$, there is no such 'natural interpretation' in the case of a simple null hypothesis $\text{H} : \theta = \theta_0$, tested against the two-sided alternative, $\overline{\text{H}} : \theta \neq \theta_0$.

Bernardo (1980) examines the basis of Bayesian tests, which he claims to be 'clear in principle'; namely, that to test whether data x are compatible with a working hypothesis H_0 it is appropriate to examine whether or not the posterior probability $P(\text{H}_0|x)$ is 'very small'. He considers the case of a 'non-informative' prior distribution (expressing prior ignorance about the parameter) and concludes that the posterior probability $P(\text{H}_0|x)$ gives a meaningful measure of the appropriateness of H_0 in the current situation only if H_0 and the alternative hypothesis H, are *both simple* or *both composite* (and of the same dimensionality). Otherwise, in the context of the Jeffreys' type of test of a simple null versus a composite alternative, he states that an interpretable unambiguous conclusion necessitates letting the prior probability $P(\theta_0)$ depend on the assumed *form* of $\pi(\theta)$ over $\theta \neq \theta_0$—possibly a rather severe constraint in terms of the practical usefulness of the procedure.

A detailed up-to-date review of hypothesis testing in the Bayesian context is given by Bernardo and Smith (1994: in particular in pp. 389–397 and 469–475). They broaden the debate to discuss choice between two alternative models, M_1, and M_2 introducing as an 'intuitive measure of pairwise comparison of plausibility' the **posterior odds ratio**

$$\frac{\pi(M_i|x)}{\pi(M_j|x)} = \frac{p(x|M_i)}{p(x|M_j)} \times \frac{\pi(M_i)}{\pi(M_j)} \quad (i \neq j = 1, 2)$$

interpretable as the product of the likelihood ratio and the prior odds ratio. This leads to the notion of the **Bayes' factor**

$$B_{ij}(x) = \frac{p(x|M_i)}{p(x|M_j)} = \left\{\frac{\pi(M_i|x)}{\pi(M_j|x)}\right\} \Big/ \left\{\frac{\pi(M_i)}{\pi(M_j)}\right\}$$

(or *posterior to prior odds ratio*) as providing for given x the relative support for M_i and M_j. This is, of course, just the *likelihood ratio*, which is central to the classical approach to hypothesis testing. Good (1950) referred to the logarithms of the various ratios above as respective 'weights of evidence', so that, for example, the *posterior weight of evidence* is the sum of the *prior weight of evidence* and *likelihood weight of evidence*.

Depending on the forms of M_i and M_j, of course, different situations are covered: point and composite hypotheses within a common family of distributions or even distinct families (e.g. geometric v. Poisson is given as an illustration).

Specific forms of **Bayes' tests** covering the various prospects of sample and composite hypotheses are exhibited based on an assumed utility structure. We examine some Bayesian decision-theoritic aspects of hypothesis testing in Chapter 7 on Decision Theory (see Section 7.3.4).

6.4 PREDICTION IN BAYESIAN INFERENCE

Consider an industrial problem in which a large batch of components contains an unknown proportion θ that are defective. The components are packaged in boxes of 50, being selected at random for this purpose from the batch.

It is of interest to be able to say something about how many defective components might be encountered in a box of 50 components, and this, of course, depends on the value of the proportion defective, θ, in the batch. Inferences about θ may be drawn in the various ways already described. We could draw a random sample of size n from the batch, observe the number of defectives, r say, and construct a classical confidence interval for θ or (more relevant to the concern of the present chapter) determine a posterior distribution for θ based on some appropriate choice of prior distribution.

If we knew the value of θ precisely, we could immediately describe the probability distribution of the number of defectives that might be present in a box of 50 components. This distribution is just binomial: $\mathbf{B}(50, \theta)$.

But θ will not be known, and yet we still have the same interest in the probabilistic behaviour of the contents of the box. We encounter a new situation here: that of *predicting the probability distribution of potential future sample data* on the basis of *inferred* knowledge of θ obtained from earlier sample data (and perhaps a prior distribution for θ).

There has been growing interest in recent years in this problem of **prediction** and the Bayesian solution takes, in principle, a particularly simple form.

In the context of the general parametric model described earlier, suppose that x represents a set of data presently available and y is a set of *potential future data*.

The problem of prediction amounts to obtaining an expression for $p(y|x)$: *the probability distribution of y conditional on the present data x and their implications in respect of the value of the parameter* θ. This distribution is known as the **predictive distribution** of y. It takes a simple form.

Specifically, we have

$$p(y|x) = \int_{\Omega} p_{\theta}(y)\pi(\theta|x), \tag{6.4.1}$$

(where $\pi(\theta|x)$ is the posterior distribution of θ, given the data x) as the **predictive probability (density) function** of y.

Clearly, $\pi(\theta|x)$ has been determined on the basis of some assumed prior distribution, $\pi(\theta)$, for θ, so that (6.4.1) can be written in the form

$$p(y|x) = \left\{\int_{\Omega} p_{\theta}(y)p_{\theta}(x)\pi(\theta)\right\} \Big/ \left\{\int_{\Omega} p_{\theta}(x)\pi(\theta)\right\}. \tag{6.4.2}$$

The function $p(y|x)$ provides the complete measure of inferential import about the future data y, based on earlier data x and the prior distribution $\pi(\theta)$. However, as in other aspects of inference, it may be that some summary form of this information is adequate for practical purposes. Thus, for example, the mode of the predictive distribution might be thought of as the 'most likely' outcome for the future data set, y. Or we could obtain an interval or region with prescribed predictive probability content: a so-called **Bayesian predictive credible interval** (or **region**). For interval estimation we usually employ **HPD predictive intervals**.

Example 6.4.1 Suppose that, in the industrial problem described at the outset of this discussion of Bayesian prediction, we adopt a beta prior distribution for θ of the form $\mathscr{B}_1(l, m)$ (see Section 2.3.1 and the discussion of conjugate prior distributions in Section 6.6 below). Then the posterior distribution of θ, having observed r defectives in a sample of n components, is $\mathscr{B}_1(l + r, m + n - r)$. If y is the number of defectives in a box of $N = 50$ components, then from (6.4.1) we can immediately determine the form of its predictive distribution. We find that y has predictive probability function

$$p(y|r) = \binom{N}{y} \frac{\mathscr{B}_1(G + y, H + N - y)}{\mathscr{B}_1(G, H)},$$

where $G = l + r, H = m + n - r$. This distribution is known as the beta-binomial distribution with parameters (N, G, H).

We could go further and determine the most likely number of defectives in a box $[(G - 1)(N + 1)/(G + H - 2)$ if $G \geq 1$; 0 otherwise] or obtain a Bayesian predictive interval of some prescribed probability content (but note that only certain probability levels can be achieved in view of the discreteness of y).

We have considered only the simplest form of the prediction problem. Refinements that introduce a loss structure in relation to incorrect prediction can be incorporated. Bernardo and Smith (1994) review the range of approaches to prediction and of its potential application (for density estimation, calebration, classification, regulation, comparison and assessment of models). They warn against the naive replacement of the right-hand side of (6.4.1) with $P(y|\hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimate of θ.

This latter prospect might well be what would be done in the classical approach to the prediction problem. Alternatively, we could use *tolerance regions* or *intervals*; see Section 5.7.2. Bernardo and Smith (1994, p. 484) find such an approach 'obscure' compared with the use of HPD predictive regions or intervals.

There has been an extensive literature on prediction. Aitchison and Dunsmore (1975) and Geisser (1993) present detailed treatments of prediction analysis from a (predominantly) Bayesian standpoint, with illustrative applications and an extensive bibliography.

Various applications areas are considered; for calibration, see Dunsmore (1968), Racine-Poon (1988); for classification, see Dunsmore (1966), Bernardo (1988), Klein and Press (1992, for spatial data), Lavine and West (1992); for discrimination, see Geisser (1966), Lavine and West (1992); for regulation, see Dunsmore (1969). See also Geisser (1974, random effects models; 1975, sample reuse), Zellner (1986, decision-theoretic with asymmetric loss structure), Cifarelli and Ragazzini (1982, free of parametric models) and Bunge and Fitzpatrick (1993, capture–recapture).

The classical approach (see Section 5.7.2. above) is reviewed by Cox (1975b), Mathiasen (1979) and, for inter-comparison of HPD *predictive intervals* and *tolerance intervals*, see Guttman (1970, 1988). See also Keyes and Levy (1996) on classical and Bayesian prediction for the multivariate linear model, and Meng (1994) on a Bayesian prediction-based examination of the classical notion of '*p*-values' and 'tail-area probabilities'.

Minority approaches to prediction occur in relation to likelihood (Kalbfleisch, 1971, Butler, 1986 and Geisser, 1993)—see also Section 8.3—and in prequential analysis due to Dawid 1984, 1992; see also Section 8.4).

6.5 PRIOR INFORMATION

We must now consider in more detail the problem of the numerical specification of prior probabilities. Bayesian methods *require* quantitative expression of the prior information available about θ, whether this has a single component or is highly multi-dimensional and whether information is specific and extensive or, at the other extreme, essentially non-existent. Various distinctions might usefully be drawn: between **prior ignorance, substantial prior knowledge** and an *intermediate* category that might be termed **vague prior knowledge**. We will consider these in turn.

6.5.1 Prior Ignorance

We must recognise that practical situations may arise where we have no *tangible* (objective or subjective) prior information. To make use of Bayesian methods of inference we are, nonetheless, compelled to express our prior knowledge in quantitative terms; we need a numerical specification of the state of **prior ignorance**. The notion of prior ignorance, and how to handle it, has been one of the most contentions issues in Bayesian statistics. We start with the Bayes–Laplace *principle of insufficient reason* expressed by Jeffreys (1961) in the following way:

> If there is no reason to believe one hypothesis rather than another, the probabilities are equal. ... *to say that the probabilities are equal is a precise way of saying that we have no ground for choosing between the alternatives*. ... The rule that we should take them equal is not a statement of any belief about the actual composition of the world, nor is it an inference from previous experience; it is merely the formal way of expressing ignorance. (pp. 33–34)

Jeffreys discusses at length (1961, Chapter 3; and elsewhere) the extension of this principle from the situation of prior ignorance concerning a discrete set of hypotheses to the case of prior ignorance about a continuously varying parameter, θ, in a parameter space, Ω. The obvious extension for a one-dimensional parameter is to assign equal prior probability density to all $\theta \in \Omega$. Thus, for a location parameter, μ, where the parameter space is the whole real line $(-\infty, \infty)$ we would choose $\pi(\mu)$ to be constant. See Example 6.3.1.

This assignment of probability is *improper* in that we cannot ensure that $P(a < \mu < b) < 1$ for *all* intervals (a, b); but this presents no basic interpretative difficulty if we are prepared to adopt (as Jeffreys demands) a degree-of-belief view of the concept of probability. We will not wish just to make prior probability statements about μ, and $\pi(\mu)$ acts merely as a *weight function* operating on the likelihood $p_\mu(x)$ to produce, after normalisation, the posterior distribution $\pi(\mu|x)$. This posterior distribution is, of course, proportional to the likelihood (cf. discussion of the generalised *likelihood inference* approach in Section 8.2).

Whilst recommending this approach for the derivation of the posterior distribution of μ, we have already seen (Section 6.3; in his concept of a Bayesian hypothesis test) that Jeffreys is not wedded to such a direct extension of the principle of insufficient reason for general application. Indeed, on the topic of hypothesis tests he remarks (1961):

> The fatal objection to the universal application of the uniform distribution is that it would make any significance test impossible. (p. 118)

(By this he means that no odds, or probability, could be assigned to a point hypothesis.) Jeffreys proposes further limitations on the use of the uniform distribution to describe prior ignorance about θ in inferential problems, suggesting that the appropriate distribution depends upon the nature of Ω.

If θ is multi-dimensional, prior ignorance needs to be expressed in terms of assumptions about the independence of exchangeability of the components of θ (see Section 3.5) and choice of forms of prior distribution for each of them. Thus, in θ is the pair (μ, σ) reflecting location and scale, μ might be assigned a uniform prior an $(-\infty, \infty)$ as above, but σ (being non-negative) will need some other form of prior distribution.

Using as criteria of choice the need to avoid anomalies from the improper nature of the distribution, and a desire to maintain certain invariance properties, Jeffreys proposes the following principles.

(i) When $\Omega = (-\infty, \infty)$, we should use the prior uniform distribution for θ. He supports this choice by noting that if we are, on this basis, in a state of prior ignorance about θ, the same will be true for any linear function of θ.

(ii) When $\Omega = (0, \infty)$ we should choose a prior distribution proportional to $1/\theta$ since this implies that we are similarly in ignorance about any power, $\theta^\alpha (\alpha \neq 0)$. In this case, $\phi = \log \theta$ has parameter space $(-\infty, \infty)$ and a prior uniform distribution, in accord with (i).

For bounded parameter spaces; for example, where θ is the parameter of a binomial distribution, so that $\Omega = (0, 1)$, the original proposal of Bayes (and Laplace) was to use the uniform assignment of prior probability density. Jeffreys expresses dissatisfaction with this on intuitive grounds other than for 'a pure estimation problem' and suggests that in many problems it is more natural to assign discrete probabilities to specific values of θ and uniform probability density elsewhere (an idea originally proposed by Haldane, 1932). Haldane has introduced an alternative specification of prior ignorance for the case where $\Omega = (0, 1)$; namely, to take

$$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}. \qquad (6.5.1)$$

In spite of some attraction in terms of invariance properties, Jeffreys (1961) rejects (6.5.1) in that it gives 'too much weight to the extremes of Ω'. An opposite dissatisfaction with the Bayes–Laplace uniform distribution prompts the tentative proposal that we might compromise by taking

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} \qquad (6.5.2)$$

to express prior ignorance when $\Omega = (0, 1)$.

We have already noticed in the quality control example of Chapter 2, other reasons why (6.5.1) might constitute an appropriate specification when $\Omega = (0, 1)$, and we shall return to this in more detail when considering conjugate prior distribution in Section 6.5.3.

Jeffreys' proposals, (i) and (ii), for the doubly and singly infinite cases are still widely adopted. Any fundamental demonstration of their validity seems impossible, however, since we are once more in the area of personal judgements about the propriety or importance of any criteria (such as invariance) that are used as justification, and of the relevance of intuitive ideas of what is meant by ignorance. Inevitably, criticisms are made of the Jeffreys' proposals. Why should we use a concept of linear invariance when $\Omega = (-\infty, \infty)$, and power-law invariance when $\Omega = (0, \infty)$? To reply that $\phi = \log\theta$ is the 'natural' parameter in the latter case seems to beg the question!

See also O'Hagan (1994, Section 5.33 *et seq.*).

Again, Jeffreys insists on a degree-of-belief interpretation of prior distributions but it is easy to understand why some of his proposals might cause concern in frequency interpretable situations. In Example 6.2.2 we adopted the Bayes–Laplace idiom of assuming $P(\mathrm{H_I}) = P(\mathrm{H_{II}}) = \frac{1}{2}$. This implicitly declares that prior ignorance is equivalent to the assumption that the box contains equal numbers of each type of die. Might it not be more reasonable to describe ignorance in terms of a diffuse *meta-prior distribution* for the proportion, θ, of type I dice in the box; that is, to say that the proportion of type I dice is itself a random quantity in (0,1) about which we know nothing. This could be expressed by a prior uniform distribution for θ, or by (6.5.1) or (6.5.2), which is quite different from our earlier assumption of equality of $P(\mathrm{H_I})$ and $P(\mathrm{H_{II}})$ (amounting to the belief that $\theta = \frac{1}{2}$, a priori, with probability 1). The introduction of meta-prior structure, although inevitably increasing the dimensionality of the parameter space, is a common feature of modern Bayesian methods.

Concern for representing prior ignorance is to enable Bayesian methods to be used when we know nothing a priori about θ. But there is another interest: in using the Bayesian approach to express what the data have to say about θ, *irrespective of prior information about* θ. Thus, we set this prior information at the level of ignorance, or non-informativeness, to 'let the data speak for themselves' and to obtain what some might claim to be an *objective* approach to inference. Bernardo and Smith (1994, p. 357) regard this aim as 'misguided': Bayesian inference must, they say, be set in a subjective mould.

O'Hagan (1994, Section 5.3.2 *et seq.*) explores another aspect of objectivity—whether Bayesian methods can be applied in the context of an *objective* (rather than subjective or frequentist), *notion of probability*. He relates this to the search for an objective approach as defined in the paragraph above for describing Bayesian inference in the face of what he prefers to term *weak prior information* (rather than prior ignorance).

It is perhaps fortunate that from the practical point of view the specific method we adopt for describing prior ignorance *will seldom make any material difference to the inference we draw*. The reasons for this are outlined in the next section.

Example 6.5.1 A random sample of n observations $x_1, x_2, \ldots, x_n$, is drawn from a normal distribution with known mean μ, and unknown variance, θ. If nothing is known, a priori, about θ, the posterior distribution of nv/θ is χ^2 with n degrees of freedom; where $v = \sum_1^n (x_i - \mu)^2/n$.

Using the Jeffreys expression of prior ignorance for θ, we have $\pi(\theta)$ proportional to θ^{-1}. So

$$\pi(\theta|x) \propto \theta^{-(n/2)-1} \exp\left\{-\frac{1}{2}\frac{nv}{\theta}\right\},$$

which implies that $Y = nv/\theta$ has p.d.f.

$$f(y) = \frac{e^{-y/2} y^{(n/2)-1}}{\Gamma(n/2)2^{n/2}}.$$

That is, Y has a χ^2 distribution with n degrees of freedom.

It is interesting to note that if both μ and θ are unknown and we express prior ignorance about them by *taking* μ *and* $log\theta$ *to be independent and to both have a uniform distribution on* $(-\infty, \infty)$, *then* $(n-1)s^2/\theta$ *is* χ^2_{n-1} *and* $n^{1/2}(\mu - \bar{x})/s$ *has a t-distribution with* $(n-1)$ *degrees of freedom* $(\bar{x} = (1/n)\sum_1^n x_i,\ s^2 = \sum_1^n (x_i - \bar{x})^2/(n-1))$.

These results are similar in form, but not in interpretation, to those on the classical approach where we found that $(n-1)s^2/\theta$ is χ^2_{n-1}, and $n^{1/2}(\mu - \bar{x})/s$ is t_{n-1}. But here it is s^2 and $\bar{x}$, not θ and μ, which are the random variables. We should resist the temptation to see this similarity as adding 'respectability' to one approach from the viewpoint of the other, or indeed as justifying the particular choice of prior distributions that have been used!

In pursuing invariance considerations in relation to the expression of prior ignorance, Jeffreys (1961) produces an alternative specification for the prior distribution in the form

$$\pi(\theta) \propto \{I_s(\theta)\}^{1/2}, \tag{6.5.3}$$

where $I_s(\theta)$ is Fisher's *information* function (see Section 5.3.2.) This has been extended by others to multi-parameter problems. The use of the information function in this way, however, is not universally accepted, since its use of *sample space averaging* is anathema to many Bayesians who claim that the only legitimate expression of the data is through the likelihood of the actual realised value x.

The current attitude to prior information and how to handle it is ably reviewed by Bernardo and Smith (1994) and O'Hagan (1994). O'Hagan (1994, Section 3.2.7 *et seq.*) is concerned that 'an improper prior cannot truly represent genuine prior information' and that, although the Jeffreys' formulation may sometimes lead to acceptable results, there are 'pitfalls' to be recognised in using improper prior distributions, e.g. in that the resulting posterior distribution may also be improper. He is also concerned that such an approach is not invariant to transformations of θ. Ignorance about θ surely implies ignorance about $\theta = g(\theta)$.

The information-based prior ignorance formulation of Jeffreys, as in (6.5.3), does satisfy desirable invariance properties in that $\pi(\theta)d\theta = \pi(\phi)d\phi$ (see Bernardo and Smith, 1994, p. 358). However, the non-uniqueness of (6.5.3) in this latter respect and its dependence on the data ('in a way that violates the Likelihood Principle': O'Hagan, 1994, p. 138) render it of limited usefulness.

So how do present-day Bayesians handle prior ignorance?

As we have already remarked, *there is still widespread use* of the Jeffreys-type priors of (i) and (ii) above, with an assumption of independence for the distinct components of θ, in spite of the critics of this approach.

Other approaches to handling prior ignorance include use of **reference priors** (Bernardo, 1979; Berger and Bernardo, 1989, 1992) and of **default priors**.

Reference priors are part of a recent thrust in Bayesian analysis directed towards the prospect that sample data dominate in import any prior information—especially when sample data are of moderate extent and prior information is 'vague' or 'non-existent'. Seeking for an expression of such a 'non-informative' prior structure leads *inter alia* to the information-based concept of a *reference prior distribution*. Bernardo and Smith (1994, Section 5.4) discuss *reference analysis* in detail—as a baseline (or default) process of examining 'the notion of a prior having a minimal effect, relative to the data, on the final inference'.

We will examine in more detail the notion of vague prior information in the next section, but the reference prior or default prior is more suitably pursued here, in our study of prior ignorance. Essentially, we start by defining a utility-based measure $I(x, \pi(\theta))$ of the *amount of information* about θ that may be expected from an experiment yielding data x, with prior distribution $\pi(\theta)$ for θ, and a measure of the expected value $I(X, \pi(\theta))$ of *perfect information* about θ.

The quantity $I(X, \pi(\theta))$ is interpretable as the *missing information* about θ, as a function of $\pi(\theta)$, and *the reference prior distribution* is that $\pi(\theta)$ which maximises $I(X, \pi(\theta))$. This is a complicated process that needs careful implementation (Bernardo and Smith, 1994, pp. 304 *et seq.*) even when θ is scalar; even more so when θ is multi-dimensional. Some examples in Bernardo and Smith (1994, Section 5.4) include the uniform prior over an appropriate range depending on data x when $X \sim U\ (\theta - 1/2, \theta + 1/2)$, the Jeffreys' prior (6.5.3), when the asymptotic posterior distribution of θ is normal, and

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

for binomial sampling (of (6.5.1) above).

Bernardo and Smith (1994, Section 5.6.2) give a detailed review of prior ignorance, referring *inter alia* to *information-theoretic* arguments of Zellner (1977, 1991) and Geisser (1979), use of *Haar measures* (e.g. Villegas, 1977), the *marginalisation paradox* of Dawid, Stone and Zidek (1973) for multi-parameter problems under which it appears that no single prior distribution may be regarded as truly 'non-informative' and the *entropy-based approach* of Jaynes (1968, 1981) (see also Csiszár, 1985).

Bernardo and Smith (1994) conclude that the entropy-based approach has merit ('for the finite case') as has Jeffreys' prior (for scalar continuous problems), but that in multi-parameter problems there is no single unique non-informative prior. They support the notion of reference priors, notwithstanding the question of Lindley (1972, p. 71).

> Why should one's knowledge, or ignorance, of a quality depend on the experiment being used to determine it?

Further discussion of the selection of prior distributions 'by formal rules' (e.g. non-informative priors) is given by Kass and Wasserman (1996). See also Sun and Ye (1995) on specific applications of reference priors.

A broader notion of default priors is discussed by Berger (1994) and includes concepts of weak information, ignorance and non-informativeness: he expresses preference for reference priors. A related topic is that of *upper and lower probabilities* under which ignorance is represented not by a single prior distribution but by a range of priors (see Walley, 1996). See Jaffray and Philippe (1997).

In the area of model comparison, improper prior distributions cause major difficulties. This prompted the development of the *fractional Bayes' factor* (O'Hagan, 1995). See also Gelfand and Dey (1994) on use of reference priors in Bayesian model choice, and the discussion of the *intrinsic Bayes' factor* of Berger and Pericchi (1995): termed *default Bayes' factors*.

Thus, we see the idea of 'ignorance' transformed to that of 'vagueness' or 'non-informativeness' or 'weak prior information'.

Quite clearly, then, there is much dissatisfaction even *within* the Bayesian approach about how we should proceed if we know nothing a priori about θ, or wish to draw inferences in relation to such a prospect. But this need not be a matter of crucial concern; it sometimes happens that the import of the data 'swamps' our prior information (however spare, or precise, this is) and the formal expression of the prior information becomes largely irrelevant, in a sense we now consider.

6.5.2 Vague Prior Knowledge

The aim of Bayesian inference is to express, through the posterior distribution of θ, the combined information provided by the prior distribution and the sample data. Inevitably, in view of the form (6.3.1) of the posterior distribution, we cannot assess what constitutes useful prior information other than in relation to the information provided by the data. Broadly speaking, the prior information increases in value the more it causes the posterior distribution to depart from the (normalised) likelihood. On this basis, we would expect prior ignorance to lead to a posterior distribution directly proportional to the likelihood. When the prior uniform distribution is used, this is certainly true. However, situations arise where it would be unreasonable to claim prior ignorance about θ but, nonetheless, the information in the sample data 'swamps' this prior information in the sense that the posterior distribution is again essentially the (normalised) likelihood. In cases of this type, we might talk of having '*vague prior knowledge*' of θ.

Bernardo and Smith (1994, Section 5.1.6) pose two questions in this context. When the information provided by the data greatly outweigh the prior information, is it reasonable to expect to be able to make a formal representation of such minimal information, or could we indeed dispense with such a representation? The first of these questions has been essentially answered in our discussion of prior ignorance above. The second is the nub of our present concern for vague (or weak, or non-informative) prior information.

An early approach was offered by Savage et al. (1962) in his formulation of the **principle of precise measurement**:

> This is the kind of measurement we have when the data are so incisive as to overwhelm the initial opinion, thus bringing a great variety of realistic initial opinions to practically the same conclusion. (p. 20)

Savage's remarks point to a major practical advantage of this principle: that conflicting opinions of the extent, and manner of expressing, prior information for a particular problem will often have little effect on the resulting conclusions.

The principle of precise measurement may be expressed in the following way (Lindley, 1965b, p. 21):

> ... if the prior [distribution of θ] ... is sensibly constant over that range of θ for which the likelihood function is appreciable, and not too large over that range of θ for which the likelihood function is small, then the posterior [distribution] ... is approximately equal to the [normalized] likelihood function ...

Elsewhere (1965b, pp. 13–14) he gives a more formal mathematical statement of this principle for the case of sampling from a normal distribution.

Savage describes this principle as one of **stable estimation**. The basic idea seems to have been recognised for a long time. Neyman (1962) attributes it to a Russian mathematician, Bernstein, and (independently) to von Mises, both during the period 1915–20.

We should note that whatever the extent of the prior knowledge this can always in principle, be outweighed by the sample data for a sufficiently large size of sample. In this sense, the principle of precise measurement leads to limiting results analogous to the classical limit laws, e.g. the *Central Limit Theorem*, although special care is needed in interpretation. The principle also enables us to interpret the (normalised) likelihood function as representing the information about θ available from the sample, *irrespective of prior knowledge* about θ.

This leads to a 'large sample' or asymptotic approach to Bayesian inference in which the posterior distribution again involves the normal distribution (Bernardo and Smith, 1994, Section 5.4) as in the classical case. We will consider this in more detail below.

This principle is easily illustrated by the example in Chapter 1 on the rate of decay, λ, of a radioactive substance (Section 1.2). We consider two different situations:

(i) 50 α-particles are observed in a period of 100 s;
(ii) 5000 α-particles are observed in a period of 10 000 s.

Suppose that, on grounds of the chemical affinity of this substance to others with known properties, we have prior reason to believe that λ is somewhere in the range $0.45 < \lambda < 0.55$. Figure 6.5.1 (a) and (b) shows, on appropriate scales, a typical prior probability density function for λ, together with the respective likelihood functions.

It is obvious that (a) and (b) represent distinctly different relationships between the prior density function and the likelihood function. The posterior distribution of λ is proportional to the product of these two functions. The case (b) provides a typical demonstration of the principle of precise measurement, where the prior distribution has negligible effect on the posterior distribution, which in turn is essentially proportional to the likelihood function. At the opposite extreme, in case (a), the major contribution to the posterior distribution comes from the prior distribution, which is little modified by the more diffuse likelihood function. Thus, we see that what appears at the outset to be tangible prior information about λ, varies in importance from one extreme to the other depending on the extent of the sample data. In simplest terms, the prior information is vague, weak or non-informative *if $\pi(\theta)$ is essentially constant over that range of values of θ for which $p_\theta(x)$ is of non-negligible value for the observed value of x.*

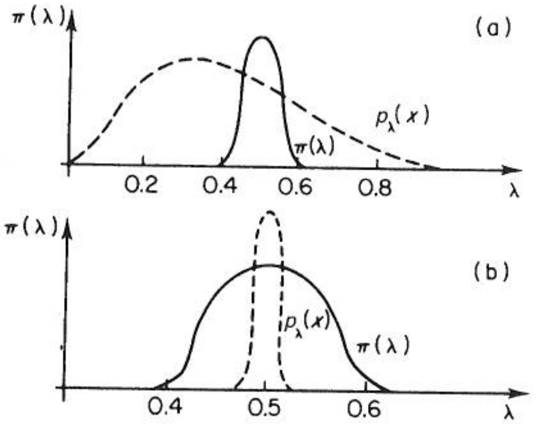

Figure 6.5.1

In more complex problems, these difficulties are even more severe. The typical practical problem needs a model involving many parameters, possibly boosted by the need for a metaprior structure (see Section 6.7). The determination of the precise form of $\pi(\theta|x)$, or of interval estimates, can now impose major computational challenges. This is one of the areas of rapid development over recent years and we will examine later the use of the *Gibbs sampler*, and of *Markov chain Monte Carlo* methods, for determination of summary measures of $\pi(\theta|x)$ in complex problems—see Section 6.6.

First, however, we need to examine a crucial concept that provides a resolution of (i) and (ii) above in appropriate circumstances. This is the notion of **conjugate families of prior distributions** that can facilitate the mathematical calculations and provide a tangible comparison of the importance of the sample data and the prior information through the idea of an '*equivalent prior sample*'.

6.5.4 Conjugate Prior Distributions

As before, we are concerned with drawing inferences about some parameter θ, which indexes the family of distributions $\mathscr{P} = \{p_\theta(x); \theta \in \Omega\}$, assumed as the model for the practical situation under study. Suppose the prior distribution of θ is a member of some parametric family of distributions, $\mathbb{P}$, with the property, in relation to $\mathscr{P}$, that the posterior distribution of θ is also a member of $\mathbb{P}$. If this is so, we say that $\mathbb{P}$ is *closed with respect to sampling* from $\mathscr{P}$, or that $\mathbb{P}$ *is a family of* **conjugate prior distributions** *relative to* $\mathscr{P}$.

More specifically, if $\mathbb{P} = \{\pi_\alpha(\theta); \alpha \in A\}$, where α is the (possibly vector) parameter of the family of prior distributions, $\mathbb{P}$, with parameter space A, then the Bayesian inference process is represented simply as a mapping of A into itself. Thus, if α_0 represents the prior information about θ, the sample data transform this to a new value α_1 representing the posterior information about θ. Symbolically, we have

$$\alpha_0 \overset{\mathscr{P}}{\rightarrow} \alpha_1 \tag{6.5.4}$$

and information at the a priori and a posteriori stages are measured by values of α in a common parameter space, A. *An initial 'amount of information', α_0, has been enhanced through sampling to a final 'amount of information', α_1.*

The potential advantages of this concept are self-evident. If only we can interpret the parameter α in terms of some properties of the sample data, we have the dual advantages of being able to define the mapping (6.5.4) in simple terms, as well as being able to measure the relative amounts of information in the prior distribution and in the sample data. For, writing α_1 as $\alpha_0 + (\alpha_1 - \alpha_0)$, we have $\alpha_1 - \alpha_0$ (expressed in terms of properties of the sample data) as a measure of the information in the sample data and can regard the prior information as that provided by an 'equivalent sample' yielding α_0.

Furthermore, explicit expressions for posterior distributions involve little calculation. The posterior distribution is in the same family $\mathbb{P}$ as the prior distribution

and all we need to do is use our knowledge of how α relates to the sample to advance the parameter from α_0 to α_1. Such a sample-oriented interpretation of α can often be achieved, as is illustrated in the following simple example.

Example 6.5.3 An electronic component has a lifetime, X, with an exponential distribution with parameter θ. That is, X has probability density function

$$f_\theta(x) = \theta e^{-\theta x}.$$

A random sample of n components have lifetimes $x_1, x_2, \ldots, x_n$. The likelihood of the sample is thus

$$p_\theta(x) = \theta^n e^{-n\theta\bar{x}}.$$

Suppose the prior distribution of θ has the form

$$\pi(\theta) = \frac{\lambda(\lambda\theta)^{r-1}e^{-\lambda\theta}}{\Gamma(r)} \tag{6.5.5}$$

then it is easy to show that the posterior distribution has precisely the same form, and is given by

$$\pi(\theta|x) = \frac{(\lambda + n\bar{x})[(\lambda + n\bar{x})\theta]^{n+r-1}e^{-\theta(\lambda+n\bar{x})}}{\Gamma(n+r)}. \tag{6.5.6}$$

We can immediately express the results of this example in the general terms above. The statistic $\bar{x}$ is sufficient for θ and essentially we are sampling at random from a gamma distribution with parameters n and $n\bar{x}$; we denote this distribution $\Gamma(n, n\bar{x})$. In this notation, the prior distribution (6.5.5) is $\Gamma(r, \lambda)$ and the posterior distribution is $\Gamma(r + n, \lambda + n\bar{x})$. Thus, the family of gamma prior distributions is closed with respect to sampling from a gamma distribution, and constitutes a family of conjugate prior distributions in this situation. The parameter α is the ordered pair (r, λ) and the rule of transformation (6.5.4) from prior to posterior distribution is

$$(r, \lambda) \rightarrow (r + n, \lambda + n\bar{x}), \tag{6.5.7}$$

which provides immediate access to the explicit form of the posterior distribution as well as an intuitive interpretation of the relative contributions of the prior distribution and the sample, to our posterior knowledge of θ.

The transformation (6.5.7) suggests that we may regard the prior information as 'equivalent to' a 'prior sample of r observations from the basic exponential distribution yielding a sample total λ'.

This concept of an **equivalent prior sample** is supported from another viewpoint. The Jeffreys' nil-prior distribution for θ (that is, his expression of prior ignorance about θ) can be represented as $\Gamma(0, 0)$. A sample of 'size' r with $n\bar{x} = \lambda$ will then, from (6.5.7), produce $\Gamma(r, \lambda)$ as the posterior distribution for θ. So to use $\Gamma(r, \lambda)$ as a prior distribution for θ, it is as if we have started from prior ignorance about θ and taken a preliminary sample (r, λ) before obtaining the real sample $(n, n\bar{x})$. In this further respect, the prior distribution (r, λ) is 'equivalent to a sample (r, λ)'. It might be tempting to feel that this argument also offers further support to Jeffreys' form of nil-prior distribution in this situation!

We must be careful not to read too much into Example 6.5.3 from the point of view of the *construction* of conjugate prior distributions. In the example, it turned out that a singly sufficient statistic existed for θ, that the prior distribution was in the same family as the sampling distribution of the sufficient statistic and that α could be expressed in terms of the value of the sufficient statistic together with the sample size. It is not true that we will always encounter such a simple structure. No singly sufficient statistic may exist for θ, and the parameter space and sample space (even reduced by sufficiency) may be quite distinct, one from another. Even when a singly sufficient statistic exists, it may not be enough merely to augment this with the sample size to construct an appropriate parameter, α, for the conjugate prior family, if we are to produce a plausible 'equivalent prior sample' concept.

The role of *sufficiency* in Bayesian inference generally is somewhat less fundamental than it is in classical inference. It is obvious that the existence of sufficient statistics will be an advantage in 'boiling down' the data and in simplifying the derivation of the posterior distribution. This aspect of the importance of sufficiency in Bayesian inference is that it acts as an 'aid to computation', rather than as an obvious prerequisite for the existence of optimal, or desirable, inferential procedures. Thus the effective advantage of a small set of sufficient statistics is that it summarises all the relevant information in the likelihood function in fewer statistics than the n sample values.

In the construction of families of conjugate prior distributions, however, sufficiency has a more important role to play. The existence of a sufficient statistic of fixed dimension independent of sample size ensures that a family of conjugate prior distributions can be found in that situation. In particular, conjugate prior distributions can be derived for sampling from any distribution in the *exponential family*, but they will exist for other distributions as well (for example, for the uniform distribution on $(0, \theta)$). Raiffa and Schlaifer (1961, Chapter 2) develop a concept of 'Bayesian sufficiency' and discuss at length (in Chapter 3) its application to the construction of conjugate prior distributions. They develop in detail all the common families of conjugate prior distributions likely to be of practical value (in Chapter 3 in synoptic form, but an extended treatment is given in Chapters 7–13.)

Bernardo and Smith (1994, Section 5.2) discuss in detail the role of conjugate families of prior distributions (including the importance of sufficient statistics and of the exponential family) stressing their value in yielding a tractable form of Bayesian inference.

Before leaving the topic of conjugate prior distributions it is useful to consider some further examples.

Example 6.5.4 If x represents a random sample of size n from a normal distribution with unknown mean, θ, and known variance σ^2 [denoted by $\mathrm{N}(\theta, \sigma^2)$], and the prior distribution of θ is $\mathrm{N}(\mu_0, \sigma_0^2)$, then the posterior distribution of θ is $\mathrm{N}(\mu_1, \sigma_1^2)$ where

$$\mu_1 = \frac{n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}, \quad \sigma_1^2 = (n/\sigma^2 + 1/\sigma_0^2)^{-1}.$$

So for sampling from a normal distribution with known variance the family of conjugate prior distributions is the normal family. But some care is needed in developing a concept of an 'equivalent prior sample' here. We must redefine the parameters by putting σ_0^2 equal to σ^2/n_0. Then (6.5.4) has the form

$$(n_0, \mu_0) \to \left(n_0 + n, \frac{n_0\mu_0 + n\bar{x}}{n_0 + n}\right) \tag{6.5.8}$$

and we can now think of the prior information as equivalent to a sample of 'size' $n_0 (= \sigma^2/\sigma_0^2)$ from $\mathrm{N}(\theta, \sigma^2)$ yielding a sample of mean, μ_0. Combining this equivalent sample with the actual sample produces a composite sample of size $n_0 + n$ with sample mean $(n_0\mu_0 + n\bar{x})/(n_0 + n)$, in accord with (6.5.8). Starting from prior ignorance [that is (0, 0), which is the improper uniform distribution] the 'equivalent sample' produces a posterior distribution $\mathrm{N}(\mu_0, \sigma_0^2)$; the composite sample produces a posterior distribution

$$\mathrm{N}\left[\frac{n_0\mu_0 + n\bar{x}}{n_0 + n}, \sigma^2/(n_0 + n)\right],$$

which agrees with the results in Example 6.5.4. Note how the domain of the parameter n has had to be extended from the integers, to the half-line $(0, \infty)$. On other occasions, it may be necessary to introduce *extra* parameters to facilitate the interpretation of the conjugate prior distribution. This is true of sampling from a normal distribution when *both the mean and variance are unknown.* (See Raiffa and Schlaifer, 1961, pp. 51–52.)

Another point is brought out by the following example.

Example 6.5.5 Binomial Sampling. Suppose r successes are observed in n independent trials, where the probability of success is θ.

If the prior distribution of θ is a beta distribution with parameters (α, β), so that

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \ (\alpha > 0, \beta > 0),$$

then the posterior distribution of θ is also a beta distribution, having parameters $(\alpha + r, \beta + n - r)$.

Thus, for binomial sampling, the family of conjugate prior distributions is the beta family. We have already observed this informally in Chapter 2. Interpreting the prior information as equivalent to a sample of 'size' $\alpha + \beta$, yielding α 'successes', seems to support Haldane's proposal (6.5.1) for an expression of prior ignorance in this situation. But the case of beta prior distributions raises certain anomalies for the concept of 'equivalent prior samples', concerned with the effects of transformations of the parameter space. (See Raiffa and Schlaifer, 1961, pp. 63–66.)

The dual advantages of mathematical tractability and ease of interpretation in the use of conjugate prior distributions are self-evident and make this concept one of major technical importance in Bayesian inference. But these potential advantages are of negligible value without an essential prerequisite.

The function of the prior distribution is to express in accurate terms the actual prior information that is available. No prior distribution, however tractable or interpretable, is of any value if it misrepresents the true situation. We must rely here on the *richness* of the family of conjugate prior distributions; that is, on the wide range of different expressions of prior belief that they are able to represent.

When prior information consists of sparse factual measures augmented by subjective impressions, as is so often the case, a variety of different specific prior distributions may have the appropriate summary characteristics to encompass the limited information that is available. In such cases, it should be quite straightforward to choose an appropriate prior distribution from the family of conjugate prior distributions. We have seen an example of this in the quality control problem discussed in Chapter 2.

The notion of imprecise prior probabilities (and expert opinions) is discussed by Coolen and Newby (1994) as an 'extension' of the standard Bayesian approach.

On rare occasions, however, the objective prior information may be so extensive as to essentially yield a detailed prior *frequency distribution* for θ. If in the current situation it is reasonable to assume that θ has arisen at random from the limiting form of this frequency distribution, then it is this frequency distribution that constitutes our best practical choice of prior distribution. Whether or not we may now take advantage of the desirable properties of conjugate prior distributions depends entirely on whether a member of this family *happens* to echo characteristics of the prior frequency distribution.

6.5.5 Quantifying Subjective Prior Information

Often, prior information, whilst substantial, is *subjective* in form and it is necessary to express it in *quantitative* terms. We have already considered different aspects of this problem: with a literature review in Section 2.3.2; some discussion in Section 3.5.2 of how individuals might attempt to quantify (via a hypothetical betting situations) their personal probability assessments and examination; in Section 4.5.2, how to construct personal utility functions.

Beyond informal attempts to put a numerical value on individual subjective probabilities and utilities, there is (rightly) growing concern for, and interest in, the interaction between the statistician and 'client' in practical Bayesian studies—in the former's efforts to 'elicit' the subjective views of the latter ('the expert') on major issues such as complete prior probability distributions, possibly in highly multi-parametric situations. Often, what are sought are prior summary measures such as means, variances and covariances that can be fed into assumed families of distributions, but then again elicitation might seek to express quantitatively *ab initio* complete prior distributions.

This distinction is well illustrated in the detailed case study examples of O'Hagan (1998) drawn from the fields of domestic water distribution and sewage disposal, and of underground storage of nuclear waste, respectively.

In a companian paper, Kadane and Wolfson (1998) adopt a broader stance and review the 'psychology of elicitation' and the wide range of currently available elicitation methodologies (where the latter may be general in nature or specific to a particular application). They warn of the 'pitfalls' of unreasonable reliance on, or attention to, *availability, adjustment and anchoring, overconfidence*, and *hindsight bias*, and propose means of minimising dangers of false representation arising from these effects. Their consensus view of elicitation is that truly expert opinion is of greatest value, with assessment only of observable quantities, in terms of moments of a distribution—with regular feedback and discussion of practical data-based implications.

Kadane and Wolfson (1998) illustrate these principles in terms of *general* and *applications-specific methods* applied to practical problems.

O'Hagan (1998) and Kadane and Wolfson (1998) provide an extensive review of what O'Hagan refers to as the 'relatively little attention in the Bayesian literature' to this topic, amounting nonetheless to a combined list of 90 references in the two papers, which provides a rich background for the further study of this interesting topic.

6.6 COMPUTING POSTERIOR DISTRIBUTIONS

We have already noted that a major problem in Bayesian inference, particularly for multi-parameter systems, is that of calculating the explicit form of the posterior distribution. Great strides have been made in this matter through simulation techniques and especially in the use of the *Gibbs* and *Metropolis–Hastings* (and other) *sampling algorithms* for the **Markov chain Monte Carlo** (MCMC) approach.

Simulation and Monte Carlo methods have long been used to evaluate integrals, usually by re-expressing a deterministic problem in probabilistic terms (e.g. as in use of Buffon's needle to calculate π).

In Bayesian inference, we typically need to integrate products of prior densities and likelihoods, and such traditional methods are widely used for this purpose—see, for example, Bernardo and Smith (1994, Sections 5.5.1 to 5.5.4) and O'Hagan (1994, Sections 8.1 to 8.4.2).

More recently, powerful methods have been developed under the title of **Markov chain Monte Carlo**. The basic idea is easily explained, although implementations can be subtle.

Traditional methods of numerical integration can involve use of approximations, of reparameterisation, of quadrature and of iterative methods (of quadrature and scaling).

Suppose we are interested, in general, in $\int f(\theta)\mathrm{d}\theta$. Quadrature methods involve evaluating $f(\theta)$ at deterministically chosen points $\theta_1, \theta_2, \ldots, \theta_k$. Monte Carlo methods, in contrast, take *random values* $\theta_1, \theta_2 \ldots$ and combine the function values $f(\theta_1), f(\theta_2) \ldots$.

In particular, if we were to choose the θ 'values' from a distribution with density $g(\theta)$, then we could write:

$$F = \int f(\theta)\mathrm{d}\theta = \int \frac{f(\theta)}{g(\theta)} g(\theta)\mathrm{d}\theta = \mathrm{E}[f(\theta)/g(\theta)]. \tag{6.6.1}$$

Thus, if we take a random sample $\theta_1, \theta_2, \ldots, \theta_n$ from $g(\theta)$, the sample mean

$$\overline{F} = n^{-1}\sum_1^n \{f(\theta_i)/g(\theta)\} \tag{6.6.2}$$

will typically be unbiased for F (the integral we wish to evaluate) and consistent (in the terms of classical inference). Furthermore, as n increases, F approaches a normal distribution with variance, which can be approximated by

$$\mathrm{Var}\{f(\theta)/\pi(\theta)\} = n^{-2}\sum_1^n f^2(\theta_i)/g^2(\theta_i) - n^{-1}\overline{F}^2. \tag{6.6.3}$$

So via (6.6.2) and (6.6.3) we can, in principle, choose n to approximate F as accurately as we wish.

In **importance sampling** (see Bernardo and O'Hagan, 1994, Section 5.5.3), we seek to choose a form of $g(\theta)$ to make the (approximate) variance (6.6.3) as small as possible. In fact, this needs $f(\theta)$ and $g(\theta)$ to be close in form. Problems with implementing this approach are particularly severe in multi-parameter problems, but much work has been done on this.

In the **Markov chain Monte Carlo** approach, we operate rather differently and may obtain greater power and facility for multi-parameter problems. Typically, we are concerned with evaluating the posterior distribution $\pi(\theta|x)$ (which, of course, still involves an integral for normalisation) at least in terms of the marginal components $\pi(\theta_1|x), \pi(\theta_2|x), \ldots, \pi(\theta_k|x)$. We seek to construct a Markov chain with state space Ω, which is easy to simulate from and which has these components as its equilibrium distribution.

Then, if we sample from this chain after it has been running for some time, we are effectively sampling at random from the posterior distribution, $\pi(\theta|x)$, of interest (in view of the ergodicity properties of the Markov chain). It sounds too simple! Can we really construct a Markov chain with equilibrium distribution $\pi(\theta|x)$ when we do not even fully know the explicit form of $\pi(\theta|x)$? Surprisingly, we can, and much work has been done in refining techniques to a stage where impressive analyses are possible for highly complex multi-parameter systems. See Bernardo and Smith (1994, Section 5.5.5) for fuller details and references.

There are two specific forms of Markov chain process that have yielded conspicuous successes in approximating important characteristics of $\pi(\theta|x)$.

Gibbs Sampling Algorithm

Suppose $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, and $\pi(\theta|x) = \pi(\theta_1, \theta_2, \ldots, \theta_k|x)$. Consider the conditional densities $\pi(\theta_i|x; \theta_j, j \neq i)$ for each of the marginal components θ_i of θ when the $\theta_j (j \neq i)$ are specified in value. These are usually readily determinable. Now consider an iterative sampling scheme that starts with some chosen values $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}$ and cycles through the θ_i updating at each stage. So we draw $\theta_1^{(1)}$ from $\pi(\theta|x; \theta_2^{(0)}, \ldots, \theta_n^{(0)})$, $\theta_2^{(1)}$ from $\pi(\theta_2|x; \theta_1^{(1)}, \theta_3^{(0)}, \theta_4^{(0)}, \ldots, \theta_k^{(0)})$, $\theta_3^{(1)}$ from $\pi(\theta_3|x; \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \ldots, \theta_k^{(0)})$, and so on. If this process is continued through many (say N) iterations, and repeated m times so that we have m replicates of $\theta^{(N)} = (\theta_1^{(N)}, \theta_2^{(N)}, \ldots, \theta_k^{(N)})$, then for large N the replicated values $\theta_{ij}^{(N)}$ $(j = 1, 2, \ldots, n)$ effectively constitute a random sample from $\pi(\theta_i|x)$ $(i = 1, 2, \ldots, k)$ the single component marginal posterior distributions.

The process of moving from $\theta^{(0)}$ to $\theta^{(1)}$ and on to $\theta^{(N)}$ is a Markov chain process since we have fixed transition probabilities from $\theta^{(l)}$ to $\theta^{(l+1)}$. This will typically have equilibrium distribution $\{\pi(\theta_1|x), \pi(\theta_2|x), \ldots, \pi(\theta_k|x)\}$ and we will thus be able to approximate (again, as accurately as we wish) the marginal posterior distributions of θ.

Appropriate computer-based simulation methods are needed for effective operation and these have been widely studied.

Hastings–Metropolis Algorithm

In this, a different iterative scheme is used, but a further randomisation also takes place and the next iterate is either accepted or rejected according to a prescribed probability mechanism.

Dellaportas and Stephens (1995) use Markov chain Monte Carlo methods in a Bayesian analysis of the errors-in-variables regression problem. Van der Merwe and Botha (1993) use the Gibbs sampling algorithm in studying mixed linear models.

Nicholls (1998) uses Markov chain Monte Carlo methods in Bayesian image analysis.

6.7 EMPIRICAL BAYES' METHODS: META-PRIOR DISTRIBUTIONS

There are other ways (beyond those discussed in Sections 6.5 and 6.6) in which prior information may be manifest. Two such possibilities are the following. The prior information may consist of *limited sample data* from situations similar to the current one but of insufficient extent to build up a frequency distribution of accurate estimates of previous θ values. Alternatively, we may sometimes have available a sample of the previous *true θ values* for some similar situations. We shall consider these possibilities briefly in this section to illustrate the use of **empirical Bayes' methods**, and **meta-prior distributions**, respectively.

6.7.1 Empirical Bayes' Methods

Prominent amongst workers in this area are Robbins (1955, 1964), who pioneered the idea, and Maritz, who in his book on *empirical Bayes' methods* (Maritz, 1970, and the later version of Maritz and Lwin, 1989) gathers together a wide range of results, many of which derive from his own research efforts. In the Preface, Maritz broadly defines the *empirical Bayes' approach* as follows.

> [It] may be regarded as part of the development towards more effective utilisation of all relevant data in statistical analysis. Its field of application is expected to lie where there is no conceptual difficulty in postulating the existence of a prior distribution that is capable of a frequency interpretation, and where *data suitable for estimation of the prior distribution* may be accumulated. (p. vii, italics inserted)

Maritz describes the approach as a 'hybrid' one, in that whilst concerned with Bayesian inference it often employs classical methods of estimation for finding estimates of, for example, the prior distribution, based on the 'prior' sample data. O'Hagan (1994, Sections 5.25 to 5.27) stresses this conceptually mixed nature of the empirical Bayes approach and concludes that 'Empirical Bayes is not Bayesian' because it does not admit a distribution for all the parameters. We will examine this point more fully below. Bernardo and Smith (1994, Section 5.6.4) give only brief coverage to such 'short cut approximations to a fully Bayesian analysis of hierarchical models'. See also Deely and Lindley (1981) and Berger (1986).

The method works in the following way. Suppose the prior distribution is $\pi(\theta|\phi)$, implying a hierarchical structure where the parameter(s) θ depends on a hyperparameter(s) ϕ. Data x are linked to the hyperparameter ϕ by means of a form of 'likelihood': $p_\phi(x) = \int p_\theta(x)\pi(\theta|\phi)\mathrm{d}\theta$. An empirical Bayes' approach substitutes for Φ an appropriate (often classical) estimator $\hat{\phi}$ and the Bayesian analysis proceeds with $\pi(\theta|\hat{\phi})$ used as the prior distribution of θ.

The following example incorporates the notion of hierarchical structure but with the hyperparametric representation of $\pi(\theta|\phi)$ essentially dealt with in empirical non-parametric terms. It concerns estimation of the mean of a Poisson distribution.

For Bayesian point estimation of a parameter θ we have so far suggested only the mode of the posterior distribution of θ, given the sample data, x. Choice of the mode rests on it being the value of θ having greatest posterior probability (density). But other summary measures of the posterior distribution might also constitute sensible point estimates of θ. In the next chapter, we shall see that from the decision theory viewpoint, with a *quadratic loss structure*, the optimal point estimator of θ is the *mean* of the posterior distribution. That is, we would estimate θ by

$$\tilde{\theta}_\pi(x) = \int_\Omega \theta p_\theta(x)\pi(\theta) \bigg/ \int_\Omega p_\theta(x)\pi(\theta). \tag{6.7.1}$$

Suppose we apply this where θ is the mean of a Poisson distribution. Then, for a single observation, x, $p_\theta(x) = \mathrm{e}^{-\theta}\theta^x/x!$ and

$$\tilde{\theta}_\pi(x) = (x+1)\phi_\pi(x+1)/\phi_\pi(x), \tag{6.7.2}$$

where

$$\phi_\pi(x) = \int_\Omega p_\theta(x)\pi(\theta) = \frac{1}{x!}\int_\Omega \theta^x \mathrm{e}^{-\theta}\pi(\theta), \tag{6.7.3}$$

i.e. the likelihood function smoothed by the prior distribution of θ.

So if the prior distribution were known then we would have in (6.7.2) a reasonable estimator of θ. For example, if θ has a prior gamma distribution, $\Gamma(r, \lambda)$, we find that $\tilde{\theta}_\pi(x)$ is $(r+x)/(1+\lambda)$.

But π is unlikely to be known. In the typical empirical Bayes' situation, we might assume that we have, in addition to the current observation x when the parameter value is θ, a set of 'previous' observations $x_1, x_2, \ldots, x_n$ obtained when the parameter values were $\theta_1, \theta_2, \ldots, \theta_n$, say (these θ values being unknown). It is assumed that the $\theta_i (i = 1, 2, \ldots, n)$ arise as a random sample from the prior distribution, $\pi(\theta)$, and that the $x_i (i = 1, 2, \ldots, n)$ are independent sample observations arising under these values of θ. The previous observations 'reflect' the prior distribution, $\pi(\theta)$, and in the general empirical Bayes' approach are used to estimate $\pi(\theta)$ for use in the Bayesian analysis.

In some cases, direct estimation of $\pi(\theta)$ is unnecessary and may be by-passed. This is so in the present example of estimating the mean, θ, of a Poisson distribution. Suppose that amongst our previous data the observation i occurs $f_n(i)$ times $(i = 0, 1, \ldots)$. The x_i may be regarded as a random sample from the smoothed likelihood function (6.7.3) since the θ_i are assumed to arise at random from $\pi(\theta)$. Thus, a simple classical estimate of $\phi_\pi(i)$ is given by $f_n(i)/(n+1)$ for $i \neq x$, or $[1 + f_n(x)]/(n+1)$ for $i = x$ (including the current observation, x).

The Bayes' point estimate (6.7.2) is then estimated by

$$\tilde{\theta}_\pi(n, x) = (x+1)f_n(x+1)/[1 + f_n(x)]. \tag{6.7.4}$$

This approach is due to Robbins (1955).

Any questions of the efficiency or optimality of such an empirical Bayes' procedure are complicated. They must take account of the possible variations in the parameter value itself that have arisen in the current situation, as well as sampling fluctuations in $\tilde{\theta}_\pi(n, x)$ arising from the different sets of previous data that might be encountered. Appropriate concepts of efficiency or optimality need to be defined and quantified, as do means of comparing empirical Bayes' estimators with alternative classical ones. Some of the progress and thinking on these matters is described by Maritz and Lwin (1989).

There has been a steady (if not voluminous) flow of published material on the empirical Bayes' method in the 40 years or so since the early proposals of Robbins (1955). Applications include hypothesis testing, interval estimation, estimating a distribution function and the parameters in the binomial distribution, in the finite Poisson process and in multilinear regression. Recent applications include Reckhow (1996) who describes an empirical Bayes' method for measuring an index of 'biotic integrity' in the context of environmental study of river-water quality, and Samaniego and Neath (1996) who claim to show (cf. O'Hagan above) that the statistician using empirical Bayes' methods in appropriate cases is a 'better Bayesian' in being able profitably to combine empirical and subjective information. Greenland and Poole (1994) consider empirical Bayes' and 'semi-Bayes' methods for environmental-hazard surveillance. Efron (1996) examines empirical Bayes' methods for combining likelihoods.

It seems likely that we shall hear much more of this approach, though assessments of its value and interpretations of its basic nature vary from one commentator to another. Neyman (1962) regarded it as a major 'breakthrough' in statistical principle; Lindley (1971c, Section 12.1) declared that its procedures are seldom *Bayesian* in principle and represent no new point of philosophy.

6.7.2 Meta-prior Distributions

Another type of situation that Maritz regards as being within the sphere of empirical Bayes' methods is that where *previous true values of* θ *are available* relating to situations similar to the current one. He remarks that such problems have not received much attention. It is worth considering an example of such a situation to demonstrate a further extension of Bayesian methods.

Suppose some manufactured product is made in batches; for example, on different machines or with different sources of a component. The quality of a product is measured by the value, x, of some performance characteristic: the corresponding quality of the batch by the mean value, θ, of this characteristic for the products in the batch. The parameter θ varies from batch to batch according to a distribution $\pi(\theta)$, which may be regarded as the prior distribution of the value relating to the batch being currently produced.

We want to draw inferences about this current θ on the basis of a random sample of n products in the current batch having performance characteristics x_1, $x_2, \ldots, x_n$. There are circumstances in which our prior information may consist of exact values of θ, for previous batches, regarded as arising at random from $\pi(\theta)$. This would be so, for instance, if part of the final inspection of the products before distribution involved measuring x for each one, *and hence* θ *for each complete batch*. Suppose these previous values were $\theta_1, \theta_2, \ldots, \theta_s$. How are we to use this information in a Bayesian analysis?

To be more specific, suppose we are prepared to accept that, within a batch, the x_i arise at random from a normal distribution, $\mathbf{N}(\theta, \sigma^2)$, *where* σ^2 *is known*, and that $\pi(\theta)$ is also normal, $\mathbf{N}(\theta_0, v)$, but where θ_0 and v are unknown.

If we knew θ_0 and v, then the posterior distribution of θ for the current batch, given $x_1, x_2, \ldots, x_n$, would be

$$\mathrm{N}[(n\overline{x}/\sigma^2 + \theta_0/v)/(n/\sigma^2 + 1/v), (n/\sigma^2 + 1/v)^{-1}], \quad \text{where } \overline{x} = \sum_1^n x_i/n :$$

see Example 6.5.4. But θ_0 and v are not known; we have merely the random sample $\theta_1, \theta_2, \ldots, \theta_s$ from $\pi(\theta)$ from which to 'estimate' them. One intuitively appealing possibility would be to use $\overline{\theta}$ and s_θ^2, the sample mean and variance of the previous θ values, to estimate $\overline{\theta}_0$ and v. This yields

$$\mathrm{N}[(n\overline{x}/\sigma^2 - \overline{\theta}/s_\theta^2)/(n/\sigma^2 + 1/s_\theta^2), (n/\sigma^2 + 1/s_\theta^2)^{-1}] \tag{6.7.5}$$

as an 'estimate' of the posterior distribution of θ.

But this cannot be entirely satisfactory! Sampling fluctuations will cause '$\overline{\theta}$' and s_θ^2 to depart from θ_0 and v, and such departures will be more serious the smaller the size, s, of the θ-sample. The estimate (6.7.5) takes no account of this; it is irrelevant whether s is 2 or 20 000!

A more satisfactory approach might be to introduce a further preliminary stage into the inferential procedure. We can do this by declaring that $\theta_1, \theta_2, \ldots, \theta_s$ arise at random from a normal distribution, $\mathrm{N}(\theta_0, v)$, where θ_0, v are *meta-prior parameters* having some *meta-prior distribution*, $\pi(\theta_0, v)$ (or as termed in Section 6.7.1 above, *hyperparameters*). We can then form the posterior distribution of θ_0 and v, given $\theta_1, \theta_2, \ldots, \theta_s$, denoted $\pi(\theta_0, v|\boldsymbol{\theta})$ and use this as *'post-prior'*, distribution for the current situation, updating it by the sample data $x_1, x_2, \ldots, x_n$ to form the posterior distribution of $\theta : \pi(\theta|\boldsymbol{\theta}, x)$.

Let us illustrate this for the present problem. Since our only prior information about θ is the normality assumption and the set of previous θ-values we can say nothing tangible about the meta-prior parameters θ_0 and v. The customary expression of this ignorance is to take

$$\pi(\theta_0, v) \propto 1/v. \tag{6.7.6}$$

(See Section 6.4.)

Using (6.7.6) we find that θ has a post-prior distribution that is $\mathbf{N}(\theta_0, v)$ where (θ_0, v) have a joint probability density function proportional to

$$v^{-(k/2)-1}\exp\{-[(s-1)s_\theta^2 + s(\overline{\theta} - \theta_0)^2]/2v\}. \tag{6.7.7}$$

Modifying this prior distribution of θ by the sample data $x_1, x_2, \ldots, x_n$, which involves averaging over (6.7.7) for θ_0 and v, finally yields the posterior density of θ as

$$\pi(\theta|\boldsymbol{\theta}, x) \propto \phi(\bar{x}, \sigma^2/n)\left\{1 + \frac{s(\theta - \bar{\theta})^2}{(s^2 - 1)s_\theta^2}\right\}^{-k/2}, \qquad (6.7.8)$$

where $\phi(\bar{x}, \sigma^2/n)$ is the density function of $\mathrm{N}(\bar{x}, \sigma^2/n)$.

This shows just how the sampling fluctuations of θ and s_θ^2 affect the situation. As $s \to \infty$, it is easy to show that (6.7.8) tends to the density function corresponding to (6.7.5). For finite s, (6.7.8) is quite different in form from the distribution (6.7.5), although just how important this difference is from a practical point of view needs further study for special cases. (See Barnett, 1973.)

The Bayesian analysis of the linear model by Lindley and Smith (1972) utilises a somewhat similar meta-parametric structure, in a more fundamental manner. Prior ignorance in the multi-parameter situation is represented by assuming the parameters *exchangeable*, with prior distributions constructed hierarchically in terms of further 'hyperparameters', which in turn have their own prior distributions, and so on. Such a hierarchical structure features widely in modern Bayesian treatment of multi-parameter problems and details can be found in Bernardo and Smith (1994) and O'Hagan (1994).

6.8 COMMENT AND CONTROVERSY

In concluding this brief survey of basic methods of Bayesian inference, there are one or two further matters that need elaboration. We shall consider briefly the questions of the interpretation of the prior and posterior distributions, the roles of sufficiency and the likelihood function and the nature of the criticisms made of the Bayesian approach.

6.8.1 Interpretation of Prior and Posterior Distributions

In introducing the Bayesian approach in Section 6.3 for parametric models, we suggested that the parameter value θ in the current situation may be thought of as *a value* (chosen perhaps 'by nature') *of a random variable* with probability (density) function $\pi(\theta)$, the so-called prior distribution of θ. This prior expression of our knowledge of θ is augmented by sample data, x, from the current situation, through the application of Bayes' theorem, to yield the posterior distribution of θ, $\pi(\theta|x)$, as the complete expression of our total knowledge of θ from both sources of information.

Such a simple expression of the principle of Bayesian inference was sufficient for the development of the specific techniques and concepts discussed throughout this chapter so far. However, we have deliberately 'glossed over' the interpretation of the probability concept inherent in such an approach to inference and must now return to this matter. It is convenient to consider the prior and posterior distributions separately.

Prior Distribution. We have seen from examples how it is quite possible to encounter situations where the prior distribution both admits, and is naturally described by, a *frequency-based* probability concept. The quality control example of Chapter 2 is typical; the machinery of empirical Bayes' procedures often presupposes such an interpretation. We are able to think of $\pi(\theta)$ as representing the relative frequency of occurence of the value θ in the 'super-experiment' of which the current situation is but a realisation. But it is equally apparent that such a frequency interpretation will not always suffice. If θ is a measure on products made by some prototype machine, or the total rainfall during the coming month, it is difficult conceptually to define the 'super-experiment' (or 'collective' in von Mises' terms), enclosing the current situation, in order to obtain a frequency interpretation of $\pi(\theta)$. For example, suppose θ is a measure of the physical stature of Shakespeare and we wish to draw inferences about θ from 'data' derived from his allusions to men's stature in his writings. In any practical sense, it cannot be right to attribute a frequency interpretation to $\pi(\theta)$. Conceptually also, the thought of an 'infinite sequence of Shakespeares, a proportion of whom had the value θ' is untenable, without a deal of 'mental juggling' (as Lindley puts it, see Section 1.6). To apply Bayesian methods in such a situation, we are essentially forced to adopt a degree-of-belief interpretation of $\pi(\theta)$; to regard $\pi(\theta)$ as measuring the extent to which we support, from our prior experience, different values of θ as being the true one. Such an approach is even more inevitable when we are dealing with subjective prior information, e.g. genuine expert knowledge and experience but present in, say, the minds of the experts rather than in quantitative form within a computer store.

But it is not only in personalistic situations that the frequency approach is untenable. We meet difficulties of interpretation even in such an apparently 'objective' problem as that of estimating the proportion, θ, of faulty items in a current batch of some product. It may be that this batch *can* be regarded as typical of a sequence of similar batches, so that there is meaning in the concept of relative frequencies of occurrence of different values of θ. Nonetheless, we cannot use this frequency distribution as a prior distribution of θ *if we do not know its form*. In an extreme case, we may have no tangible prior information on which to base our prior distribution, $\pi(\theta)$, and consequently may need to use a conventional expression of this prior ignorance. (See Section 6.4.) We are not pretending that this $\pi(\theta)$ corresponds to the relative frequency distribution of θ from batch to batch; we are merely saying that it expresses (albeit rather formally) *our prior beliefs* about the *actual* value of θ in the *current* situation. The position is essentially the same if we quantify some limited objective and subjective information to form $\pi(\theta)$ (perhaps by choice of a member of the appropriate family of conjugate prior distributions) perhaps following a process of *elicitation* (See Section 6.5.5). The interpretation of $\pi(\theta)$ is again more naturally a degree-of-belief, rather than a frequency, one. For again we are *not* claiming that $\pi(\theta)$ *coincides* with the frequency distribution of θ from batch to batch.

Posterior Distribution. Here again, it is difficult to accommodate a frequency interpretation of the probability concept, let alone insist on it. Even when the prior distribution has an immediate frequency interpretation, it is not self-evident that the posterior distribution, $\pi(\theta|x)$, can also be described in frequency terms.

Consider again the quality control problem just described, and suppose that the limiting form of the relative frequencies of occurrence of different values of θ is known precisely. We can take this as the *exact* form of the prior distribution of θ. Consider the situation represented diagrammatically in Figure 6.8.1, where we have in mind an infinite sequence of similar batches of the product each with its corresponding proportion, θ, of faulty items. A current batch (observed NOW) has unknown θ but we have extracted sample data x to reflect on θ. In the past (PREVIOUSLY) indefinitely many batches have been produced, each with its own proportion of faulty items, $\theta_1, \theta_2, \ldots, \theta_n, \ldots$, and this sequence effectively determines $\pi(\theta)$, the prior distribution of θ values. The data, x, and prior distribution, $\pi(\theta)$, combine to produce the posterior distribution, $\pi(\theta|x)$, and the question is how to interpret the probability concept that is transferred to $\pi(\theta|x)$. Since θ in the current situation is a unique, if unknown, value describing this situation, it seems natural (inevitable) that we should invoke a degree-of-belief interpretation of $\pi(\theta|x)$.

We might ask, however, if there is *any* sense in which we can give $\pi(\theta|x)$ a frequency interpretation. To answer this, it helps to re-iterate a distinction drawn earlier (Section 2.3.1) on how the proper processing of *extra data* depends on whether the extra data arose in the *current* situation or in a *new* (but similar) situation.

Consider a further batch (LATER) with its corresponding parameter value θ'. *Suppose it yields sample data,* y. To draw inferences about θ', what should we use as the prior distribution of θ? Should it be $\pi(\theta)$ or $\pi(\theta|x)$ (since x constitutes prior information at this LATER stage) or perhaps something else entirely? We cannot use $\pi(\theta|x)$ for this *new* batch—this represents our views about the parameter value *in the batch that has given rise to* x (not to y). In fact, we must again use $\pi(\theta)$ since this is assumed to be the *exact* distribution of θ from batch to batch, and knowing this there is nothing more to know about the random mechanism that has produced θ' for this new batch.

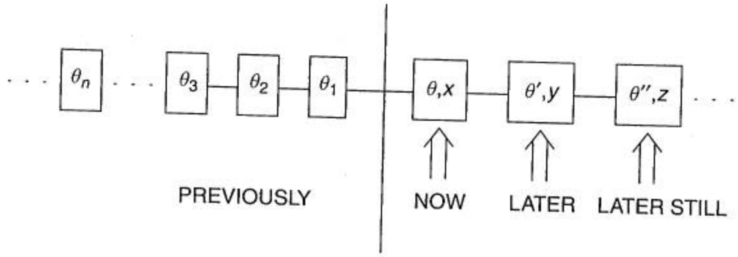

Figure 6.8.1

But the use of $\pi(\theta)$ for the new batch needs to be qualified in two respects. If y had arisen not LATER from a new batch with a new parameter value θ', but as an independent sample NOW extending the earlier data x, then it would obviously have been necessary to use $\pi(\theta|x)$, *not* $\pi(\theta)$, as the prior distribution to apply to y. This is typical of *a sequential inference* situation and we shall say more about this in the discussion of Bayesian methods in *decision theory* (Chapter 7). Secondly, had $\pi(\theta)$ not been a complete and precise expression of variation in θ from batch to batch, then the data x might have been able to provide *more* information on this variation. We should still *not* use $\pi(\theta|x)$ as the prior distribution in the LATER situation, but would want to augment the 'incomplete' $\pi(\theta)$ *appropriately* by the extra information, x, before applying it to the new data, y. What is meant by appropriately is far from clear, however, except *perhaps* in an empirical Bayes' problem. In this case, $\pi(\theta)$ has been estimated from previous data. We can go back to this previous data, extend it with x, and derive a corresponding better estimate of $\pi(\theta)$ to apply to y.

These observations give the clue to a possible *frequency* interpretation of $\pi(\theta|x)$. Consider yet further batches (LATER STILL ...), with their own parameter values, θ'', ..., and sample data, $z, \ldots$. Within this indefinite sequence, there will be a subsequence in which the data was the same as in the NOW situation; that is, x. We can regard this subsequence as a 'collective' and develop a frequency interpretation of $\pi(\theta|x)$, in which $\pi(\theta|x)$ is the limiting relative frequency of batches *with data* x for which the parameter value is θ. But such an interpretation has to be viewed on the same level as that of the *classical* confidence interval. In both cases, inferences relate to a single determined quantity, yet are interpreted in the wider framework of 'situations that may have happened'. The criticisms of the confidence interval concept (Section 5.7.1) must apply equally here, and it is likely that most Bayesians would, for similar reasons, reject such a frequency interpretation in favour of a degree-of-belief view of $\pi(\theta|x)$.

When no 'collective' can be defined in the practical manner illustrated above, *any* frequency interpretation of $\pi(\theta|x)$ becomes very contrived, and a degree-of-belief view inevitable.

6.8.2 Sufficiency, Likelihood and Unbiasedness

In Chapter 5, we discussed at length the central nature of these concepts in the classical approach to inference. For comparison, we should consider their role in Bayesian inference.

We have already discussed *sufficiency* (Section 6.5) and seen it serving essentially utilitarian needs. The existence of a small set of sufficient statistics reduces the computational effort in applying Bayesian methods generally. More specifically, it is a prerequisite for the existence of interpretable families of conjugate prior distributions. This latter service reflects a real importance of sufficiency in this approach to inference. There is little question here, as in the classical approach, of it being a precondition for the existence of optimal inferential procedures (but see Chapter 7 on *decision theory*). Bernardo and Smith (1994, Section 5.1.4) discuss the role of Sufficiency (and of ancillarity and nuisance parameters) in Bayesian inference. Dawid (1980) also examines such matters.

On the other hand, the *likelihood function* acts as the cornerstone of the Bayesian approach, whereas in the classical approach it acts more as a tool for the construction of *particular* methods of estimation and hypothesis testing. The real importance of the likelihood in Bayesian inference is in its function as the *sole* expression of the information in the sample data. So that if for two data sets, x_1 and x_2, the likelihoods $p_\theta(x_1)$ and $p^1_\theta(x_2)$ are proportional, then inferences about θ will be identical. This function is expressed through the **likelihood principle**, which has already been discussed in Section 5.6.

As we remarked earlier (Section 5.7.4), this principle is a direct consequence of Bayes' theorem. One implication of the likelihood principle is that inferences about θ will depend only on *relative variations* in the likelihood function from one value of θ to another. This leads (in the *strong* form of the likelihood principle) to the effect described as '*the irrelevance of the sampling rule*', and which constitutes one of the major philosophical distinctions between the Bayesian and classical approaches. Let us reconsider this by means of a specific example.

Consider a sequence of independent Bernoulli trials in which there is a constant probability of success, θ, on each trial. The observation of, say, 4 successes in 10 trials could arise in two ways; either by taking 10 trials yielding 4 successes, or by sampling until 4 successes occur that happens to require 10 trials. On the Bayesian approach, this distinction is irrelevant; the likelihood is proportional to $\theta^4(1-\theta)^6$ in each case and inferences about θ will be the same provided the prior distribution is the same in both cases. This is not true on the classical approach. The direct sampling procedure can produce quite different results to the inverse sampling one. For example, a 95 per cent upper confidence bound for θ is 0.697 in the first case, and 0.755 in the second case. For comparison, using the Haldane form (6.5.1) to express prior ignorance about θ, we obtain a 95 per cent upper Bayesian confidence bound of 0.749.

Reaction to this basic distinction between the classical and Bayesian approaches will again be a matter of personal attitude. The Bayesian will have no sympathy with any prescription that takes account of the method of collecting the data, in view of the centrality of the likelihood principle in the Bayesian approach. In contrast, the classicist is likely to see this as a fundamental weakness of Bayesian methods: that they cannot take account of the sampling technique. See Section 5.7.4 for a more fundamental discussion of this point and of the central role played by the concept of *coherence* in this debate.

The Bayesian view that the likelihood function conveys the total import of the data x rules out any formal consideration of the sample space $\mathscr{X}$ [except as the domain over which $p_\theta(x)$ is defined]. Inferences are conditional on the realised value x; other values that *may* have occurred are regarded as irrelevant. In particular, no consideration of the *sampling distribution* of a statistic is entertained; sample space averaging is ruled out. Thus, in particular, there can be no consideration of the *bias* of an estimation procedure and this concept is totally disregarded. A Bayesian estimator $\tilde{\theta}(x)$ relates in probability terms to the posterior distribution of θ given the particular data x; it cannot be regarded as a typical value of $\tilde{\theta}(X)$ having a probability distribution over $\mathscr{X}$. (But, again, see the decision theory applications in Chapter 7.)

6.8.3 Controversy

Attitudes to inference have become more eclectic over recent years and Bayesian and classical methods may be used depending on the nature of a problem being studied. The Bayesian approach, however, has firm advocates who cannot entertain the principles of the classical idiom and, in contrast, many statisticians feel unable to sympathise with, or adopt, the Bayesian approach to inference. This latter attitude is often supported by claims of 'lack of objectivity' and 'impracticality' in Bayesian methods, but we need to cut through the emotive nature of such criticism to appreciate the substance of the dissatisfaction. Some objections are raised to the use of the concept of 'inverse probability' as a legitimate tool for statistical inference. Fisher was particularly vehement in his rejection of this concept as we have observed earlier (Section 1.6). But most current criticism concerns the basic nature of the prior distribution, and its quantification. Dissatisfaction is expressed with the use of prior distributions where the essential form of the problem precludes a frequency interpretation. We have seen (Section 1.6) that von Mises, whilst committed to the Bayesian approach, conceived of its application only in frequency-interpretable situations; Hogben (1957, Chapters 5 and 6) likewise. Others would claim that the *whole approach* is untenable since it sometimes requires a subjective or degree-of-belief concept of probability. In this sense, it is not 'objective' and thus not appropriate for a formal scientific theory of inference!

Then again, even ignoring such philosophical complaints, the Bayesian approach meets opposition and rejection on the grounds that the prior information itself is often subjective—its quantification correspondingly 'arbitrary'. There is concern for the formality and perceived 'arbitrariness' of ways of handling prior ignorance: of the use of non-informative, or reference, priors. This leads to dissatisfaction with an approach where the conclusions are seen to depend critically on ill-formulated or imprecise information—which may vary from individual to individual, or time to time, in its formal expression!

Criticism of this type is well illustrated by the early commentary, by Pearson (1962a), which is relatively free from the emotive undertones often associated with such exprésssions of opinion:

> Let me illustrate some of my difficulties [with subjective Bayesian methods] very briefly.

> a) We are told that "if one is being consistent, there is a prior distribution". "A subjectivist feels that the prior distribution means something about the state of his mind and that he can discover it by introspection". But does this mean that if introspection fails to produce for me a stable and meaningful prior distribution which can be expressed in terms of numbers, I must give up the use of statistical method?
>
> b) Again, it is an attractive hypothesis that Bayesian probabilities "only differ between individuals because individuals are differently informed; but with common knowledge we have common Bayesian probabilities". Of course it is possible to define conceptual Bayesian probabilities and the "rational man" in this way, but how to establish that all this bears a close relation to reality?
>
> It seems to me that in many situations, if I received no more relevant knowledge in the interval and could forget the figures I had produced before, I might quote at intervals widely different Bayesian probabilities for the same set of states, simply because I should be attempting what would be for me impossible and resorting to guesswork. It is difficult to see how the matter could be put to experimental test. Of course the range of problems is very great. At one end we have the case where a prior distribution can be closely related to past observation; at the other, it has to be determined almost entirely by introspection or (because we do not trust our introspection) by the introduction of some formal mathematical function, in Jeffreys' manner, to get the model started. In the same way utility and loss functions have sometimes a clear objective foundation, but must sometimes be formulated on a purely subjectivist basis.
>
> To have a unified mathematical model of the mind's way of working in all these varied situations is certainly intellectually attractive. But is it always meaningful? I think that there is always this question at the back of my mind: can it really lead to my own clear thinking to put at the very foundation of the mathematical structure used in acquiring knowledge, functions about whose form I have often such imprecise ideas?

Such remarks encapsulate many of the elements of criticism of the Bayesian approach. It is not part of our purpose to take sides on such issues—any approach to inference involves personal judgements on the relevance and propriety of its criteria. We have seen equally forceful criticism of the classical approach. In essence, it revolved around the same issues of how well the approach meets up to its practical aims; how 'arbitrary' are its criteria, concepts and methods. In broad terms, one can find criticisms of the two approaches that are basically the same criticism. Again, see Section 5.7.4.

What cannot be denied is that there are some *fundamental* differences in the classical and Bayesian approaches. These include

- *the interpretation of the parameter* θ, as a determined quantity or a random variable
- *the nature of the probability concept* (frequentist, degree-of-belief, subjective)
- *the role of the data* x, as specific and conditioning or as representative of sample-space variability

and, stemming from the last of these and central to the whole debate,

- the distinction between *initial precision* and *final precision* (see Section 5.7.1).

There must be undoubted appeal in an approach to inference (the Bayesian) in which inferential statements relate to the specific problem in hand (*final precision*) rather than being (as in classical methods) representative of the range of prospects that might be encountered in the long run (*initial precision*). But as with all distinctions between the approaches, reactions will depend on personal attitudes.

The principle justification of the Bayesian approach advanced by its advocates is its '*inevitability*': that if we accept the ideas of *coherence* and *consistency*, then prior probabilities (and utilities) *must* exist and be employed. We shall take this up in more detail in the next chapter.

On the matter of the nature and form of prior distributions, the *principle of precise measurement* (Section 6.5.2) is crucial. It implies that in a large number of situations detailed quantitative expression of the prior information is unnecessary. The Bayesian approach in such situations is *robust*; the data 'swamps' the prior information, different observers must arrive at essentially the same conclusions and questions of 'subjectivity' or 'objectivity' become less focused.

In situations where the prior information is quantitative and relates to a prior distribution admitting a frequency interpretation, one hears little objection expressed to the principle, or practice, of Bayesian inference.

We considered in Section 6.1 above, texts at various levels, and with different emphases, on Bayesian statistics, and publications on applied themes. In conclusion, we should consider published work that is addressed to comparative issues: attempts to reconcile, or to contrast, Bayesian inference and other approaches. In this context, we should include Lindley (1958; Bayesian and fiducial interval estimation, see Section 8.1), Pearson (1962a; Bayesian, classical and decision-theoretic approaches), Thatcher (1964) and Peers (1968) on Bayesian credibility intervals and confidence intervals, Bartholomew (1965 and 1971; Bayesian versus classical), Barnard (1972; a call for unity), De Groot (1973; 'tail areas'), Cox (1978; the case for eclecticism), Bernardo (1980; hypothesis testing), Diaconis and Freedman (1983; frequency properties of Bayes' rules), Efron (1993; interval estimates), O'Hagan (1994; Sections 1.33 *et seq.*), Bernardo and Smith (1994; pp. 443–488 with literature review), Sweeting (1995a; conditional inference) and Sweeting (1995b; Bayesian and likelihood methods), Datta (1996) and Datta and Ghosh (1995a) on frequentist validity of Bayesian inference, Bernard (1996; comparing methods of estimating a Bernoulli process parameter), Conigliana et al. (1997; robustness), Cox (1997; basic distinctions) Vaurio (1992; objective prior distributions) and Zellner (1995). Case-study inter-comparisons include Biggerstaff et al. (1994; meta-analyses on passive smoking) and Tamura et al. (1994; treatment of depressive disorder).