Bayesian Reference Analysis

A Postgraduate Tutorial Course

José M. Bernardo

Universitat de València
Valencia, Spain

1998

Departament d'Estadística i I. O, Facultat de Matemàtiques
46100–Burjassot, València, Spain
Tel. and Fax +34.96.364.3560
E-mail: jose.m.bernardo@uv.es, Web: http://www.uv.es/˜bernardo

Summary

This monograph offers an introduction to *Bayesian Reference Analysis*, often regarded as the more successful method to produce non-subjective, model-based, posterior distributions, the key to objective Bayesian methods in scientific research.

It has been produced as an update, with corrections and additions, of the material included in selected sections of *Bayesian Theory*, by J. M. Bernardo and A. F. M. Smith (Wiley, 1994), to be used as a set of lecture notes for postgraduate courses on *Objective Bayesian Inference*.

Chapter 1 contains an introduction to the Bayesian paradigm and introduces the necessary notation. Chapter 2 develops the necessary results in Bayesian asymptotics. Chapter 3 describes reference analysis; this is the heart of this monograph. Chapter 4 contains further discussion on the issues involved. An appendix summarizes basic formulae. Signposts are provided throughout to the huge related literature.

Keywords: AMOUNT OF INFORMATION; BAYESIAN ASYMPTOTICS; BAYESIAN INFERENCE; DEFAULT PRIORS; FISHER MATRIX; NON-INFORMATIVE PRIORS; REFERENCE PRIORS.

Contents

1. The Bayesian Paradigm

1.1. THE ROLE OF BAYES' THEOREM

The foundational issues which arise when we aspire to formal quantitative coherence in the context of decision making in situations of uncertainty, in combination with an operational approach to the basic concepts, leads to view the problem of statistical modelling as that of identifying or selecting particular forms of representation of beliefs about observables.

For example, in the case of a sequence $x_1, x_2, \ldots,$ of 0 – 1 random quantities for which beliefs correspond to a judgement of infinite exchangeability, de Finetti's theorem identifies the representation of the joint mass function for $x_1, \ldots, x_n$ as having the form

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}\, dQ(\theta),$$

for some choice of distribution Q over the interval $[0, 1]$.

More generally, for sequences of random quantities, $x_1, x_2, \ldots,$ it is known (see *e.g.*, Bernardo and Smith, 1994, Chap. 4) that beliefs which combine judgements of exchangeability with some form of further structure (either in terms of invariance or sufficient statistics), often lead us to work with representations of the form

$$p(x_1, \ldots, x_n) = \int_{\Re^k} \prod_{i=1}^n p(x_i \,|\, \boldsymbol{\theta})\, dQ(\boldsymbol{\theta}),$$

where $p(x \,|\, \boldsymbol{\theta})$ denotes a specified form of labelled family of probability distributions and Q is some choice of distribution over $\Re^k$.

Such representations exhibit the various ways in which the element of primary significance from the subjectivist, operationalist standpoint, namely the *predictive model* of beliefs about observables, can be thought of *as if* constructed from a *parametric model* together with a *prior distribution* for the labelling parameter.

Our primary concern in this monograph will be with the way in which the updating of beliefs in the light of new information takes place within the framework of such representations, when no prior *subjective* information exists, or —if it does— it is *not* desired to take such information into account.

In its simplest form, within the formal framework of predictive model belief distributions derived from quantitative coherence considerations, the problem corresponds to identifying the joint conditional density of

$$p(x_{n+1}, \ldots, x_{n+m} \,|\, x_1, \ldots, x_n)$$

for any $m \geq 1$, given, for any $n \geq 1$, the form of representation of the joint density $p(x_1, \ldots, x_n)$.

In general, of course, this simply reduces to calculating

$$p(x_{n+1}, \ldots, x_{n+m} \,|\, x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_{n+m})}{p(x_1, \ldots, x_n)}$$

and, in the absence of further structure, there is little more that can be said. However, when the predictive model admits a representation in terms of parametric models and prior distributions, the learning process can be essentially identified, in conventional terminology, with the standard parametric form of Bayes' theorem.

Thus, for example, if we consider the general parametric form of representation for an exchangeable sequence, with $dQ(\boldsymbol{\theta})$ having density representation, $p(\boldsymbol{\theta})d\boldsymbol{\theta}$, we have

$$p(x_1,\ldots,x_n) = \int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

from which it follows that

$$\begin{aligned} p(x_{n+1},\ldots,x_{n+m} \,|\, x_1,\ldots,x_n) &= \frac{\int \prod_{i=1}^{n+m} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\ &= \int \prod_{i=n+1}^{n+m} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, x_1,\ldots,x_n)\, d\boldsymbol{\theta}, \end{aligned}$$

where

$$p(\boldsymbol{\theta} \,|\, x_1,\ldots,x_n) = \frac{\prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\,.$$

This latter relationship is just *Bayes' theorem*, expressing the *posterior density* for $\boldsymbol{\theta}$, given $x_1,\ldots,x_n$, in terms of the *parametric model* for $x_1,\ldots,x_n$ given $\boldsymbol{\theta}$, and the *prior density* for $\boldsymbol{\theta}$. The (conditional, or posterior) predictive model for $x_{n+1},\ldots,x_{n+m}$, given $x_1,\ldots,x_n$ is seen to have precisely the same general form of representation as the initial predictive model, except that the corresponding parametric model component is now integrated with respect to the posterior distribution of the parameter, rather than with respect to the prior distribution. Considered as a function of $\boldsymbol{\theta}$,

$$\mathrm{lik}(\boldsymbol{\theta} \,|\, x_1,\ldots,x_n) = p(x_1,\ldots,x_n \,|\, \boldsymbol{\theta})$$

is usually referred to as the *likelihood function*. A formal definition of such a concept is, however, problematic; for details, see Bayarri *et al.* (1988) and Bayarri and DeGroot (1992b).

1.2. PREDICTIVE AND PARAMETRIC INFERENCE

Given our operationalist concern with modelling and reporting uncertainty in terms of *observables*, it is not surprising that Bayes' theorem, in its role as the key to a coherent learning process for *parameters*, simply appears as a step within the predictive process of passing from

$$p(x_1,\ldots,x_n) = \int p(x_1,\ldots,x_n \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

to

$$p(x_{n+1},\ldots,x_{n+m} \,|\, x_1,\ldots,x_n) = \int p(x_{n+1},\ldots,x_{n+m} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, x_1,\ldots,x_n)\, d\boldsymbol{\theta},$$

by means of

$$p(\boldsymbol{\theta} \,|\, x_1,\ldots,x_n) = \frac{p(x_1,\ldots,x_n \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(x_1,\ldots,x_n \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\,.$$

Writing $\boldsymbol{y} = \{y_1, \ldots, y_m\} = \{x_{n+1}, \ldots, x_{n+m}\}$ to denote future (or, as yet unobserved) quantities and $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ to denote the already observed quantities, these relations may be re-expressed more simply as

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}) = \int p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}$$

and

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) / p(\boldsymbol{x}).$$

However, if we proceed purely formally, from an operationalist standpoint it is not at all clear, at first sight, how we should interpret "beliefs about parameters", as represented by $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, or even whether such "beliefs" have any intrinsic interest. It is well known (see *e.g.*, Bernardo and Smith, 1994, Ch. 4) that, in all the forms of predictive model representations we considered, the parameters had interpretations as strong law limits of (appropriate functions of) observables. Thus, for example, in the case of the infinitely exchangeable 0 – 1 sequence beliefs about θ correspond to beliefs about what the long-run frequency of 1's would be in a future sample; in the context of a real-valued exchangeable sequence with centred spherical symmetry, beliefs about μ and σ^2, respectively, correspond to beliefs about what the large sample mean, and the large sample mean sum of squares about the sample mean would be, in a future sample.

Inference about parameters is thus seen to be a limiting form of predictive inference about observables. This means that, although the predictive form is primary, and the role of parametric inference is typically that of an intermediate structural step, parametric inference will often itself be the legitimate end-product of a statistical analysis in situations where interest focuses on quantities which could be viewed as large-sample functions of observables. Either way, parametric inference is of considerable importance for statistical analysis in the context of the models we are mainly concerned with in this volume.

When a parametric form is involved simply as an intermediate step in the predictive process, we have seen that $p(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n)$, the full joint posterior density for the parameter vector $\boldsymbol{\theta}$, is all that is required. However, if we are concerned with parametric inference *per se*, we may be interested in only some subset, $\boldsymbol{\phi}$, of the components of $\boldsymbol{\theta}$, or in some transformed subvector of parameters, $\boldsymbol{g}(\boldsymbol{\theta})$. For example, in the case of a real-valued sequence we may only be interested in the large-sample mean and not in the variance; or in the case of two 0 – 1 sequences we may only be interested in the difference in the long-run frequencies.

In the case of interest in a subvector of $\boldsymbol{\theta}$, let us suppose that the full parameter vector can be partitioned into $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \boldsymbol{\lambda}\}$, where $\boldsymbol{\phi}$ is the subvector of interest, and $\boldsymbol{\lambda}$ is the complementary subvector of $\boldsymbol{\theta}$, often referred to, in this context, as the vector of *nuisance parameters*. Since

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{x})}\,,$$

the (marginal) posterior density for $\boldsymbol{\phi}$ is given by

$$p(\boldsymbol{\phi} \,|\, \boldsymbol{x}) = \int p(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\lambda} = \int p(\boldsymbol{\phi}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\lambda},$$

where

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\lambda}) p(\boldsymbol{\phi}, \boldsymbol{\lambda}) d\boldsymbol{\phi}\, d\boldsymbol{\lambda},$$

with all integrals taken over the full range of possible values of the relevant quantities.

In some situations, the prior specification $p(\boldsymbol{\phi}, \boldsymbol{\lambda})$ may be most easily arrived at through the specification of $p(\boldsymbol{\lambda} \,|\, \boldsymbol{\phi})p(\boldsymbol{\phi})$. In such cases, we note that we could first calculate the *integrated likelihood* for $\boldsymbol{\phi}$,

$$p(\boldsymbol{x} \,|\, \boldsymbol{\phi}) = \int p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\lambda})p(\boldsymbol{\lambda} \,|\, \boldsymbol{\phi})\, d\boldsymbol{\lambda},$$

and subsequently proceed without any further need to consider the nuisance parameters, since

$$p(\boldsymbol{\phi} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\phi})p(\boldsymbol{\phi})}{p(\boldsymbol{x})}\,.$$

In the case where interest is focused on a transformed parameter vector, $\boldsymbol{g}(\boldsymbol{\theta})$, we proceed using standard change-of-variable probability techniques. Suppose first that $\boldsymbol{\psi} = \boldsymbol{g}(\boldsymbol{\theta})$ is a one-to-one differentiable transformation of $\boldsymbol{\theta}$. It then follows that

$$p_\psi(\boldsymbol{\psi} \,|\, \boldsymbol{x}) = p_\theta(\boldsymbol{g}^{-1}(\boldsymbol{\psi}) \,|\, \boldsymbol{x})\, |\, \boldsymbol{J}_{g^{-1}}(\boldsymbol{\psi})\, |\,,$$

where

$$\boldsymbol{J}_{g^{-1}}(\boldsymbol{\psi}) = \frac{\partial \boldsymbol{g}^{-1}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$$

is the Jacobian of the inverse transformation $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\psi})$. Alternatively, by substituting $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\psi})$, we could write $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ as $p(\boldsymbol{x} \,|\, \boldsymbol{\psi})$, and replace $p(\boldsymbol{\theta})$ by $p_\theta(\boldsymbol{g}^{-1}(\boldsymbol{\psi}))\, |\, \boldsymbol{J}_{g^{-1}}(\boldsymbol{\psi})\, |$, to obtain $p(\boldsymbol{\psi} \,|\, \boldsymbol{x}) = p(\boldsymbol{x} \,|\, \boldsymbol{\psi})p(\boldsymbol{\psi})/p(\boldsymbol{x})$ directly.

If $\boldsymbol{\psi} = g(\boldsymbol{\theta})$ has dimension less than $\boldsymbol{\theta}$, we can typically define $\gamma = (\boldsymbol{\psi}, \boldsymbol{\omega}) = \boldsymbol{h}(\boldsymbol{\theta})$, for some $\boldsymbol{\omega}$ such that $\gamma = \boldsymbol{h}(\boldsymbol{\theta})$ is a one-to-one differentiable transformation, and then proceed in two steps. We first obtain

$$p(\boldsymbol{\psi}, \boldsymbol{\omega} \,|\, \boldsymbol{x}) = p_\theta(\boldsymbol{h}^{-1}(\gamma) \,|\, \boldsymbol{x})\, |\, \boldsymbol{J}_{h^{-1}}(\gamma)\, |\,,$$

where

$$\boldsymbol{J}_{h^{-1}}(\gamma) = \frac{\partial \boldsymbol{h}^{-1}(\gamma)}{\partial \gamma}\,,$$

and then marginalise to

$$p(\boldsymbol{\psi} \,|\, \boldsymbol{x}) = \int p(\boldsymbol{\psi}, \boldsymbol{\omega} \,|\, \boldsymbol{x})\, d\boldsymbol{\omega}.$$

These techniques will be used extensively in later parts of this monograph.

In order to keep the presentation of these basic manipulative techniques as simple as possible, we have avoided introducing additional notation for the ranges of possible values of the various parameters. In particular, all integrals have been assumed to be over the full ranges of the possible parameter values.

In general, this notational economy will cause no confusion and the parameter ranges will be clear from the context. However, there are situations where specific constraints on parameters are introduced and need to be made explicit in the analysis. In such cases, notation for ranges of parameter values will typically also need to be made explicit.

Consider, for example, a parametric model, $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, together with a prior specification $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, for which the posterior density, suppressing explicit use of Θ, is given by

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\,.$$

Now suppose that it is required to specify the posterior subject to the constraint $\boldsymbol{\theta} \in \Theta_0 \subset \Theta$, where $\int_{\Theta_0} p(\boldsymbol{\theta}) d\boldsymbol{\theta} > 0$.

Defining the constrained prior density by

$$p_0(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta})}{\int_{\Theta_0} p(\boldsymbol{\theta}) d(\boldsymbol{\theta})}, \quad \boldsymbol{\theta} \in \Theta_0,$$

we obtain, using Bayes' theorem,

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \boldsymbol{\theta} \in \Theta_0) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p_0(\boldsymbol{\theta})}{\int_{\Theta_0} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p_0(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \quad \boldsymbol{\theta} \in \Theta_0.$$

From this, substituting for $p_0(\boldsymbol{\theta})$ in terms of $p(\boldsymbol{\theta})$ and dividing both numerator and denominator by

$$p(\boldsymbol{x}) = \int_{\Theta} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

we obtain

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \boldsymbol{\theta} \in \Theta_0) = \frac{p(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{\int_{\Theta_0} p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \, d\boldsymbol{\theta}}, \quad \boldsymbol{\theta} \in \Theta_0,$$

expressing the constraint in terms of the unconstrained posterior (a result which could, of course, have been obtained by direct, straightforward conditioning).

Numerical methods are often necessary to analyze models with constrained parameters; see Gelfand *et al.* (1992) for the use of Gibbs sampling in this context.

1.3. SUFFICIENCY, ANCILLARITY AND STOPPING RULES

It is known (see *e.g.*, Bernardo and Smith, 1994, Ch.4) that a (minimal) sufficient statistic, $\boldsymbol{t}(\boldsymbol{x})$, for $\boldsymbol{\theta}$, in the context of a parametric model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, can be characterised by either of the conditions

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = p(\boldsymbol{\theta} \,|\, \boldsymbol{t}(\boldsymbol{x})), \qquad \text{for all } p(\boldsymbol{\theta}),$$

or

$$p(\boldsymbol{x} \,|\, \boldsymbol{t}(\boldsymbol{x}), \boldsymbol{\theta}) = p(\boldsymbol{x} \,|\, \boldsymbol{t}(\boldsymbol{x})).$$

The important implication of the concept is that $\boldsymbol{t}(\boldsymbol{x})$ serves as a sufficient summary of the complete data $\boldsymbol{x}$ in forming any required revision of beliefs. The resulting data reduction often implies considerable simplification in modelling and analysis. In many cases, the sufficient statistic $\boldsymbol{t}(\boldsymbol{x})$ can itself be partitioned into two component statistics, $\boldsymbol{t}(\boldsymbol{x}) = [\boldsymbol{a}(\boldsymbol{x}), \boldsymbol{s}(\boldsymbol{x})]$ such that, for all $\boldsymbol{\theta}$,

$$\begin{aligned} p(\boldsymbol{t}(\boldsymbol{x}) \,|\, \boldsymbol{\theta}) &= p(\boldsymbol{s}(\boldsymbol{x}) \,|\, \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta}) \, p(\boldsymbol{a}(\boldsymbol{x}) \,|\, \boldsymbol{\theta}) \\ &= p(\boldsymbol{s}(\boldsymbol{x}) \,|\, \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta}) \, p(\boldsymbol{a}(\boldsymbol{x})). \end{aligned}$$

It then follows that, for any choice of $p(\boldsymbol{\theta})$,

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = p(\boldsymbol{\theta} \,|\, \boldsymbol{t}(\boldsymbol{x})) &\propto p(\boldsymbol{t}(\boldsymbol{x}) \,|\, \boldsymbol{\theta}) \, p(\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{s}(\boldsymbol{x}) \,|\, \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta}) \, p(\boldsymbol{\theta}), \end{aligned}$$

so that, in the prior to posterior inference process defined by Bayes' theorem, it suffices to use $p(\boldsymbol{s}(\boldsymbol{x}) \,|\, \boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta})$, rather than $p(\boldsymbol{t}(\boldsymbol{x}) \,|\, \boldsymbol{\theta})$ as the likelihood function. This further simplification motivates the following definition.

Definition 1. ***Ancillary statistic.***. *A statistic, $\boldsymbol{a}(\boldsymbol{x})$, is said to be ancillary, with respect to $\boldsymbol{\theta}$ in a parametric model $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, if $p(\boldsymbol{a}(\boldsymbol{x})\,|\,\boldsymbol{\theta}) = p(\boldsymbol{a}(\boldsymbol{x}))$ for all values of $\boldsymbol{\theta}$.*

Example 1. ***Bernoulli model***. It is well known that for the Bernoulli parametric model

$$\begin{aligned} p(x_1, \ldots, x_n\,|\,\theta) &= \prod_{i=1}^{n} p(x_i\,|\,\theta) = \theta^{rn}(1-\theta)^{n-rn} \\ &= p(n, r_n\,|\,\theta), \end{aligned}$$

where $r_n = x_1 + \cdots + x_n$, so that $\boldsymbol{t}_n = [n, r_n]$ provides a minimal sufficient statistic.

If we now write

$$p(n, r_n\,|\,\theta) = p(r_n\,|\,n, \theta)p(n\,|\,\theta),$$

and make the assumption that, for all $n \geq 1$, the mechanism by which the sample size, n, is arrived at does not depend on θ, so that $p(n\,|\,\theta) = p(n)$, $n \geq 1$, we see that n *is ancillary for* θ, in the sense of Definition 1. It follows that prior to posterior inference for θ can therefore proceed on the basis of

$$p(\theta\,|\,\boldsymbol{x}) = p(\theta\,|\,n, r_n) \propto p(r_n\,|\,n, \theta)p(\theta),$$

for any choice of $p(\theta)$, $0 \leq \theta \leq 1$. Since

$$\begin{aligned} p(r_n\,|\,n, \theta) &= \binom{n}{r_n}\theta^{rn}(1-\theta)^{n-rn}, \qquad\qquad 0 \leq r_n \leq n, \\ &= \mathrm{Bi}(r_n\,|\,\theta, n), \end{aligned}$$

inferences in this case can be made as if we had adopted a *binomial parametric model*. However, if we write

$$p(n, r_n\,|\,\theta) = p(n\,|\,r_n, \theta)p(r_n\,|\,\theta)$$

and make the assumption that, for all $r_n \geq 1$, termination of sampling is governed by a mechanism for selecting r_n, which does not depend on θ, so that $p(r_n\,|\,\theta) = p(r_n), r_n \geq 1$, we see that r_n is ancillary for θ, in the sense of Definition 1. It follows that prior to posterior inference for θ can therefore proceed on the basis of

$$p(\theta\,|\,\boldsymbol{x}) = p(\theta\,|\,n, r_n) \propto p(n\,|\,r_n, \theta)p(\theta),$$

for any choice of $p(\theta), 0 < \theta \leq 1$. It is easily verified that

$$\begin{aligned} p(n\,|\,r_n, \theta) &= \binom{n-1}{r_n-1}\theta^{rn}(1-\theta)^{n-rn}, \qquad\qquad n \geq r_n, \\ &= \mathrm{Nb}(n\,|\,\theta, r_n) \end{aligned}$$

so that inferences in this case can be made as if we had adopted a *negative-binomial parametric model*.

We note, incidentally, that whereas in the binomial case it makes sense to consider $p(\theta)$ as specified over $0 \leq \theta \leq 1$, in the negative-binomial case it may only make sense to think of $p(\theta)$ as specified over $0 < \theta \leq 1$, since $p(r_n\,|\,\theta = 0) = 0$, for all $r_n \geq 1$.

So far as prior to posterior inference for θ is concerned, we note that, for any specified $p(\theta)$, and assuming that either $p(n\,|\,\theta) = p(n)$ or $p(r_n\,|\,\theta) = p(r_n)$, we obtain

$$p(\theta\,|\,x_1, \ldots, x_n) = p(\theta\,|\,n, r_n) \propto \theta^{rn}(1-\theta)^{n-rn}p(\theta)$$

since, considered as functions of θ,

$$p(r_n\,|\,n, \theta) \propto p(n\,|\,r_n, \theta) \propto \theta^{rn}(1-\theta)^{n-rn}.$$

□

The last part of the above example illustrates a general fact about the mechanism of parametric Bayesian inference which is trivially obvious; namely, *for any specified $p(\boldsymbol{\theta})$, if the likelihood functions $p_1(\boldsymbol{x}_1 \,|\, \boldsymbol{\theta}), p_2(\boldsymbol{x}_2 \,|\, \boldsymbol{\theta})$ are proportional as functions of $\boldsymbol{\theta}$, the resulting posterior densities for $\boldsymbol{\theta}$ are identical.* It turns out that many non-Bayesian inference procedures do not lead to identical inferences when applied to such proportional likelihoods. The assertion that they *should,* the so-called *Likelihood Principle,* is therefore a controversial issue among statisticians . In contrast, in the Bayesian inference context described above, this is a straightforward consequence of Bayes' theorem, rather than an imposed "principle". Note, however, that the above remarks are predicated on a specified $p(\boldsymbol{\theta})$. It may be, of course, that knowledge of the particular sampling mechanism employed has implications for the specification of $p(\boldsymbol{\theta})$, as illustrated, for example, by the comment above concerning negative-binomial sampling and the restriction to $0 < \theta \leq 1$.

Although the likelihood principle is implicit in Bayesian statistics, it was developed as a separate principle by Barnard (1949), and became a focus of interest when Birnbaum (1962) showed that it followed from the widely accepted sufficiency and conditionality principles. Berger and Wolpert (1984/1988) provide an extensive discussion of the likelihood principle and related issues. Other relevant references are Barnard *et al.* (1962), Fraser (1963), Pratt (1965), Barnard (1967), Hartigan (1967), Birnbaum (1968, 1978), Durbin (1970), Basu (1975), Dawid (1983a), Joshi (1983), Berger (1985b), Hill (1987) and Bayarri *et al.* (1988).

Example 1 illustrates the way in which ancillary statistics often arise naturally as a consequence of the way in which data are collected. In general, it is very often the case that the sample size, n, is fixed in advance and that inferences are automatically made conditional on n, without further reflection. It is, however, perhaps not obvious that inferences can be made conditional on n if the latter has arisen as a result of such familiar imperatives as "stop collecting data when you feel tired", or "when the research budget runs out". The kind of analysis given above makes it intuitively clear that such conditioning is, in fact, valid, provided that the mechanism which has led to n "does not depend on $\boldsymbol{\theta}$". This latter condition may, however, not always be immediately transparent, and the following definition provides one version of a more formal framework for considering sampling mechanisms and their dependence on model parameters.

> **Definition 2.** ***Stopping rule***. *A* ***stopping rule***, $\boldsymbol{\tau}$, *for (sequential) sampling from a sequence of observables $x_1 \in X_1, x_2 \in X_2, \ldots,$ is a sequence of functions $\tau_n : X_1 \times \cdots \times X_n \to [0,1]$, such that, if $\boldsymbol{x}_{(n)} = (x_1, \ldots, x_n)$ is observed, then sampling is terminated with probability $\tau_n(\boldsymbol{x}_{(n)})$; otherwise, the $(n+1)$th observation is made. A stopping rule is* ***proper*** *if the induced probability distribution $p_\tau(n), n = 1, 2, \ldots,$ for final sample size guarantees that the latter is finite. The rule is* ***deterministic*** *if $\tau_n(\boldsymbol{x}_{(n)}) \in \{0,1\}$ for all $(n, \boldsymbol{x}_{(n)})$; otherwise, it is a* ***randomised*** *stopping rule.*

In general, we must regard the data resulting from a sampling mechanism defined by a stopping rule $\boldsymbol{\tau}$ as consisting of $(n, \boldsymbol{x}_{(n)})$, the sample size, together with the observed quantities $x_1, \ldots, x_n$. A parametric model for these data thus involves a probability density of the form $p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{\tau}, \boldsymbol{\theta})$, conditioning both on the stopping rule (*i.e.*, sampling mechanism) and on an underlying labelling parameter $\boldsymbol{\theta}$. But, either through unawareness or misapprehension, this is typically ignored and, instead, we act as if the actual observed sample size n had been fixed in advance, in effect assuming that

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{\tau}, \boldsymbol{\theta}) = p(\boldsymbol{x}_{(n)} \,|\, n, \boldsymbol{\theta}) = p(\boldsymbol{x}_{(n)} \,|\, \boldsymbol{\theta}),$$

using the standard notation we have hitherto adopted for fixed n. The important question that now arises is the following: under what circumstances, if any, can we proceed to make inferences about $\boldsymbol{\theta}$ on the basis of this (generally erroneous!) assumption, without considering explicit conditioning on the actual form of $\boldsymbol{\tau}$? Let us first consider a simple example.

Example 2. *"Biased" stopping rule for a Bernoulli sequence*. Suppose, given θ, that $x_1, x_2, \ldots$ may be regarded as a sequence of independent Bernoulli random quantities with $p(x_i \mid \theta) = \text{Bi}(x_i \mid \theta, 1)$, $x_i = 0, 1$, and that a sequential sample is to be obtained using the deterministic stopping rule $\boldsymbol{\tau}$, defined by: $\tau_1(1) = 1$, $\tau_1(0) = 0$, $\tau_2(x_1, x_2) = 1$ for all x_1, x_2. In other words, if there is a success on the first trial, sampling is terminated (resulting in $n = 1$, $x_1 = 1$); otherwise, two observations are obtained (resulting in either $n = 2$, $x_1 = 0$, $x_2 = 0$ or $n = 2$, $x_1 = 0$, $x_2 = 1$).

At first sight, it might appear essential to take explicit account of τ in making inferences about θ, since the sampling procedure seems designed to bias us towards believing in large values of θ. Consider, however, the following detailed analysis:

$$\begin{aligned} p(n = 1, x_1 = 1 \mid \boldsymbol{\tau}, \theta) &= p(x_1 = 1 \mid n = 1, \boldsymbol{\tau}, \theta) p(n = 1 \mid \boldsymbol{\tau}, \theta) \\ &= 1 \cdot p(x_1 = 1 \mid \theta) = p(x_1 = 1 \mid \theta) \end{aligned}$$

and, for $x = 0, 1$,

$$p(n = 2, x_1 = 0, x_2 = x \mid \boldsymbol{\tau}, \theta) = p(x_1 = 0, x_2 = x \mid n = 2, \boldsymbol{\tau}, \theta) p(n = 2 \mid \boldsymbol{\tau}, \theta)$$

$$\begin{aligned} &= p(x_1 = 0 | n = 2, \boldsymbol{\tau}, \theta) p(x_2 = x \mid x_1 = 0, n = 2, \boldsymbol{\tau}, \theta) p(n = 2 \mid \boldsymbol{\tau}, \theta) \\ &= 1 \cdot p(x_2 = x \mid x_1 = 0, \theta) p(x_1 = 0 \mid \theta) \\ &= p(x_2 = x, x_1 = 0 \mid \theta). \end{aligned}$$

Thus, for all $(n, \boldsymbol{x}_{(n)})$ having non-zero probability, we obtain in this case

$$p(n, \boldsymbol{x}_{(n)} \mid \boldsymbol{\tau}, \theta) = p(\boldsymbol{x}_{(n)} \mid \theta),$$

the latter considered pointwise as functions of θ (*i.e.*, likelihoods). It then follows trivially from Bayes' theorem that, *for any specified* $p(\theta)$, inferences for θ based on assuming n to have been fixed at its observed value will be identical to those based on a likelihood derived from explicit consideration of $\boldsymbol{\tau}$.

Consider now a randomised version of this stopping rule which is defined by $\tau_1(1) = \alpha$, $\tau_1(0) = 0$, $\tau_2(x_1, x_2) = 1$ for all x_1, x_2. In this case, we have

$$\begin{aligned} p(n = 1, x_1 = 1 \mid \boldsymbol{\tau}, \theta) &= p(x_1 = 1 \mid n = 1, \boldsymbol{\tau}, \theta) p(n = 1 \mid \boldsymbol{\tau}, \theta) \\ &= 1 \cdot \alpha \cdot p(x_1 = 1 \mid \theta), \end{aligned}$$

with, for $x = 0, 1$,

$$\begin{aligned} p(n =& 2, x_1 = 0, x_2 = x \mid \boldsymbol{\tau}, \theta) \\ &= p(n = 2 \mid x_1 = 0, \boldsymbol{\tau}, \theta) \\ &\quad \times p(x_1 = 0 \mid \boldsymbol{\tau}, \theta) p(x_2 = x \mid x_1 = 0, n = 2, \boldsymbol{\tau}, \theta) \\ &= 1 \cdot p(x_1 = 0 \mid \theta) p(x_2 = x \mid \theta) \end{aligned}$$

and

$$\begin{aligned} p(n = 2, x_1 = 1, x_2 = x \mid \boldsymbol{\tau}, \theta) = p(n = 2 \mid x_1 = 1, \boldsymbol{\tau}, \theta) p(x_1 = 1 \mid \boldsymbol{\tau}, \theta) \\ \times p(x_2 = x \mid x_1 = 1, n = 2, \boldsymbol{\tau}, \theta) \\ = (1 - \alpha) p(x_1 = 1 \mid \theta) p(x_2 = x \mid \theta). \end{aligned}$$

Thus, for all $(n, \boldsymbol{x}_{(n)})$ having non-zero probability, we again find that

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{\tau}, \theta) \propto p(\boldsymbol{x}_{(n)} \,|\, \theta)$$

as functions of θ, so that the proportionality of the likelihoods once more implies identical inferences from Bayes' theorem, for any given $p(\theta)$.

□

The analysis of the preceding example showed, perhaps contrary to intuition, that, although seemingly biasing the analysis towards beliefs in larger values of θ, the stopping rule does not in fact lead to a different likelihood from that of the a priori fixed sample size. The following, rather trivial, theorem makes clear that this is true for all stopping rules as defined in Definition 2, which we might therefore describe as "likelihood non-informative stopping rules".

Theorem 1. ***Stopping rules are likelihood non-informative***.
For any stopping rule $\boldsymbol{\tau}$, *for (sequential) sampling from a sequence of observables* $x_1, x_2, \ldots$, *having fixed sample size parametric model* $p(\boldsymbol{x}_{(n)} \,|\, n, \boldsymbol{\theta}) = p(\boldsymbol{x}_{(n)} \,|\, \boldsymbol{\theta})$,

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{\tau}, \boldsymbol{\theta}) \propto p(\boldsymbol{x}_{(n)} \,|\, \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta,$$

for all $(n, \boldsymbol{x}_{(n)})$ *such that* $p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{\tau}, \boldsymbol{\theta}) \neq 0$.

Proof. This follows straightforwardly on noting that

$$p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{\tau}, \boldsymbol{\theta}) = \Big[\boldsymbol{\tau}(\boldsymbol{x}_n) \prod_{i=1}^{n-1} (1 - \boldsymbol{\tau}(\boldsymbol{x}_i))\Big] p(\boldsymbol{x}_{(n)} \,|\, \boldsymbol{\theta}),$$

and that the term in square brackets does not depend on $\boldsymbol{\theta}$.

◁

Again, it is a trivial consequence of Bayes' theorem that, *for any specified prior density*, prior to posterior inference for $\boldsymbol{\theta}$ given data $(n, \boldsymbol{x}_{(n)})$ obtained using a likelihood non-informative stopping rule $\boldsymbol{\tau}$ can proceed by acting as if $\boldsymbol{x}_{(n)}$ were obtained using a fixed sample size n. However, a notationally precise rendering of Bayes' theorem,

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, n, \boldsymbol{x}_{(n)}, \boldsymbol{\tau}) &\propto p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{\tau}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \boldsymbol{\tau}) \\ &\propto p(\boldsymbol{x}_{(n)} \,|\, \theta) p(\boldsymbol{\theta} \,|\, \boldsymbol{\tau}), \end{aligned}$$

reveals that *knowledge of* $\boldsymbol{\tau}$ *might well affect the specification of the prior density*! It is for this reason that we use the term "likelihood non-informative" rather than just "non-informative" stopping rules. It cannot be emphasised too often that, although it is often convenient for expository reasons to focus at a given juncture on one or other of the "likelihood" and "prior" components of the model, they are basically inseparable in coherent modelling and analysis of beliefs. This issue is highlighted in the following example.

Example 3. ***"Biased" stopping rule for a normal mean***. Suppose, given θ, that data $x_1, x_2, \ldots$, may be regarded as a sequence of independent normal random quantities with $p(x_i \,|\, \theta) = \mathrm{N}(x_i \,|\, \theta, 1)$, $x_i \in \Re$. Suppose further that an investigator has a particular concern with the parameter value $\theta = 0$ and wants to stop sampling if $\overline{x}_n = \sum_i x_i / n$ ever takes on a value that is "unlikely", assuming $\theta = 0$ to be true.

For any fixed sample size n, if "unlikely" is interpreted as "an event having probability less than or equal to α", for small α, a possible stopping rule, using the fact that $p(\overline{x}_n \,|\, n, \theta) = \mathrm{N}(\overline{x}_n \,|\, \theta, n)$, might be

$$\tau_n(\boldsymbol{x}_{(n)}) = \begin{cases} 1, & \text{if } |\,\overline{x}_n\,| > k(\alpha)/\sqrt{n} \\ 0, & \text{if } |\,\overline{x}_n\,| \le k(\alpha)/\sqrt{n} \end{cases}$$

for suitable $k(\alpha)$ (for example, $k = 1.96$ for $\alpha = 0.05$, $k = 2.57$ for $\alpha = 0.01$, or $k = 3.31$ for $\alpha = 0.001$). It can be shown, using the law of the iterated logarithm, that this is a proper stopping rule, so that termination will certainly occur for some finite n, yielding data $(n, \boldsymbol{x}_{(n)})$. Moreover, defining

$$\begin{aligned} S_n = \Big\{ \boldsymbol{x}_{(n)};\ |\bar{x}_1| \le k(\alpha),\ |\bar{x}_2| \le \frac{k(\alpha)}{\sqrt{2}}, \cdots, \\ |\bar{x}_{n-1}| \le \frac{k(\alpha)}{\sqrt{n-1}},\ |\bar{x}_n| > \frac{k(\alpha)}{\sqrt{n}} \Big\}, \end{aligned}$$

we have

$$\begin{aligned} p(n, \boldsymbol{x}_{(n)} \,|\, \boldsymbol{\tau}, \theta) &= p(\boldsymbol{x}_{(n)} \,|\, n, \boldsymbol{\tau}, \theta) p(n \,|\, \boldsymbol{\tau}, \theta) \\ &= p(\boldsymbol{x}_{(n)} \,|\, S_n, \theta) p(S_n \,|\, \theta) \\ &= p(\boldsymbol{x}_{(n)} \,|\, \theta), \end{aligned}$$

as a function of θ, for all $(n, \boldsymbol{x}_{(n)})$ for which the left-hand side is non-zero. It follows that $\boldsymbol{\tau}$ is a likelihood non-informative stopping rule.

Now consider prior to posterior inference for θ, where, for illustration, we assume the prior specification $p(\theta) = \mathrm{N}(\theta \,|\, \mu, \lambda)$, with precision $\lambda \simeq 0$, to be interpreted as indicating extremely vague prior beliefs about θ, which take no explicit account of the stopping rule $\boldsymbol{\tau}$. Since the latter is likelihood non-informative, we have

$$\begin{aligned} p(\theta \,|\, \boldsymbol{x}_{(n)}, n) &\propto p(\boldsymbol{x}_{(n)} \,|\, n, \theta) p(\theta) \\ &\propto p(\overline{x}_n \,|\, n, \theta) p(\theta) \\ &\propto \mathrm{N}(\overline{x}_n \,|\, \theta, n) \mathrm{N}(\theta \,|\, \mu, \lambda) \end{aligned}$$

by virtue of the sufficiency of $(n, \overline{x}_n)$ for the normal parametric model. The right-hand side is easily seen to be proportional to $\exp\{-\frac{1}{2}Q(\theta)\}$, where

$$Q(\theta) = (n + \tau) \left[\theta - \frac{n\bar{x}_n + \lambda\mu}{n + \lambda} \right]^2,$$

which implies that

$$\begin{aligned} p(\theta \,|\, \boldsymbol{x}_{(n)}, n) &= \mathrm{N}\left(\theta \Big| \frac{n\bar{x}_n + \lambda\mu}{n + \lambda}, (n + \lambda) \right) \\ &\simeq \mathrm{N}(\theta \,|\, \overline{x}_n, n) \end{aligned}$$

for $\lambda \simeq 0$.

One consequence of this vague prior specification is that, having observed $(n, \boldsymbol{x}_{(n)})$, we are led to the posterior probability statement

$$P\Big[\theta \in \Big(\overline{x}_n \pm \frac{k(\alpha)}{\sqrt{n}} \Big) \Big| n, \overline{x}_n \Big] = 1 - \alpha.$$

But the stopping rule $\boldsymbol{\tau}$ ensures that $|\,\overline{x}_n\,| > k(\alpha)/\sqrt{n}$. This means that the value $\theta = 0$ certainly does not lie in the posterior interval to which someone with initially very vague beliefs would attach a high probability. An investigator *knowing* $\theta = 0$ to be the true value can therefore, by using this stopping rule, mislead someone who, unaware of the stopping rule, acts as if initially very vague.

However, let us now consider an analysis which takes into account the stopping rule. The nature of $\boldsymbol{\tau}$ might suggest a prior specification $p(\theta \,|\, \boldsymbol{\tau})$ that recognises $\theta = 0$ as a possibly "special" parameter value, which should be assigned non-zero prior probability (rather than the zero probability resulting from any continuous prior density specification). As an illustration, suppose that we specify

$$p(\theta \,|\, \boldsymbol{\tau}) = \alpha \; 1_{(\theta=0)}(\theta) + (1-\alpha) 1_{(\theta\neq 0)}(\theta) \mathrm{N}(\theta \,|\, 0, \lambda_0),$$

which assigns a "spike" of probability, α, to the special value, $\theta = 0$, and assigns $1-\alpha$ times a $\mathrm{N}(\theta \,|\, 0, \lambda_0)$ density to the range $\theta \neq 0$.

Since $\boldsymbol{\tau}$ is a likelihood non-informative stopping rule and $(n, \overline{x}_n)$ are sufficient statistics for the normal parametric model, we have

$$p(\theta \,|\, n, \boldsymbol{x}_{(n)}, \boldsymbol{\tau}) \propto \mathrm{N}(\overline{x}_n \,|\, \theta, n) p(\theta \,|\, \boldsymbol{\tau}).$$

The complete posterior $p(\theta \,|\, n, \boldsymbol{x}_{(n)}, \boldsymbol{\tau})$ is thus given by

$$\frac{\alpha \; 1_{(\theta=0)}(\theta)\mathrm{N}(\overline{x}_n \,|\, 0, n) + (1-\alpha)1_{(\theta\neq 0)}(\theta)\mathrm{N}(\overline{x}_n \,|\, \theta, n)\mathrm{N}(\theta \,|\, 0, \lambda_0)}{\alpha \; \mathrm{N}(\overline{x}_n \,|\, 0, n) + (1-\alpha)\int_{-\infty}^{\infty} \mathrm{N}(\overline{x}_n \,|\, \theta, n)\mathrm{N}(\theta \,|\, 0, \lambda_0) d\theta}$$

$$= \alpha^* 1_{(\theta=0)}(\theta) + (1-\alpha^*) 1_{(\theta\neq 0)} \mathrm{N}\left(\theta \Big| \frac{n\overline{x}_n}{n+\lambda_0}, n+\lambda_0\right),$$

where, since

$$\int_{-\infty}^{\infty} \mathrm{N}(\overline{x}_n \,|\, \theta, n)\mathrm{N}(\theta \,|\, 0, \lambda_0) d\theta = \mathrm{N}\left(\overline{x}_n \,|\, 0, n\frac{\lambda_0}{n+\lambda_0}\right),$$

it is easily verified that

$$\alpha^* = \left\{1 + \frac{1-\alpha}{\alpha} \cdot \frac{\mathrm{N}(\overline{x}_n \,|\, 0, n\lambda_0(n+\lambda_0)^{-1})}{\mathrm{N}(\overline{x}_n \,|\, 0, n)}\right\}^{-1}$$

$$= \left\{1 + \frac{1-\alpha}{\alpha}\left(1 + \frac{n}{\lambda_0}\right)^{-1/2} \exp\left[\tfrac{1}{2}(\sqrt{n}\overline{x}_n)^2\left(1 + \frac{\lambda_0}{n}\right)^{-1}\right]\right\}^{-1}.$$

The posterior distribution thus assigns a "spike" α^* to $\theta = 0$ and assigns $1-\alpha^*$ times a $\mathrm{N}(\theta \,|\, (n+\lambda_0)^{-1} n\overline{x}_n,\, n+\lambda_0)$ density to the range $\theta \neq 0$.

The behaviour of this posterior density, derived from a prior taking account of $\boldsymbol{\tau}$, is clearly very different from that of the posterior density based on a vague prior taking no account of the stopping rule. For qualitative insight, consider the case where actually $\theta = 0$ and α has been chosen to be very small, so that $k(\alpha)$ is quite large. In such a case, n is likely to be very large and at the stopping point we shall have $\overline{x}_n \simeq k(\alpha)/\sqrt{n}$. This means that

$$\alpha^* \simeq \left[1 + \frac{1-\alpha}{\alpha}\left(1 + \frac{n}{\lambda_0}\right)^{-1/2} \exp\left(\tfrac{1}{2}k^2(\alpha)\right)\right]^{-1} \simeq 1,$$

for large n, so that knowing the stopping rule and then observing that it results in a large sample size leads to an increasing conviction that $\theta = 0$. On the other hand, if θ is appreciably different from 0, the resulting n, and hence α^*, will tend to be small and the posterior will be dominated by the $\mathrm{N}(\theta \,|\, (n+\lambda_0)^{-1} n\overline{x}_n, n+\lambda_0)$ component.

□

1.4. DECISIONS AND INFERENCE SUMMARIES

Our central concern is the representation and revision of beliefs as the basis for decisions. Either beliefs are to be used directly in the choice of an action, or are to be recorded or reported in some selected form, with the possibility or intention of subsequently guiding the choice of a future action. The elements of a decision problem in the inference context are:

(i) $\boldsymbol{a} \in \mathcal{A}$, available "answers" to the inference problem;
(ii) $\boldsymbol{\omega} \in \Omega$, unknown states of the world;
(iii) $u : \mathcal{A} \times \Omega \to \Re$, a function attaching utilities to each consequence $(\boldsymbol{a}, \boldsymbol{\omega})$ of a decision to summarise inference in the form of an "answer", $\boldsymbol{a}$, and an ensuing state of the world, $\boldsymbol{\omega}$;
(iv) $p(\boldsymbol{\omega})$, a specification, in the form of a probability distribution, of current beliefs about the possible states of the world.

The optimal choice of answer to an inference problem is an $\boldsymbol{a} \in \mathcal{A}$ which *maximises the expected utility,*

$$\int_\Omega u(\boldsymbol{a}, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

Alternatively, if instead of working with $u(\boldsymbol{a}, \boldsymbol{\omega})$ we work with a so-called *loss function,*

$$l(\boldsymbol{a}, \boldsymbol{\omega}) = f(\boldsymbol{\omega}) - u(\boldsymbol{a}, \boldsymbol{\omega}),$$

where *f* is an arbitrary, fixed function, the optimal choice of answer is an $\boldsymbol{a} \in \mathcal{A}$ which *minimises the expected loss,*

$$\int_\Omega l(\boldsymbol{a}, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

It is clear from the forms of the expected utilities or losses which have to be calculated in order to choose an optimal answer, that, if beliefs about unknown states of the world are to provide an appropriate basis for future decision making, where, as yet, $\mathcal{A}$ and u (or l) may be unspecified, we need to report the complete belief distribution $p(\boldsymbol{\omega})$.

However, if an immediate application to a particular decision problem, with specified $\mathcal{A}$ and u (or l), is all that is required, the optimal answer—maximising the expected utility or minimising the expected loss—may turn out to involve only limited, specific features of the belief distribution, so that these "summaries" of the full distribution suffice for decision-making purposes.

In the following subsections, we shall illustrate and discuss some of these commonly used forms of summary. Throughout, we shall have in mind the context of parametric and predictive inference, where the unknown states of the world are parameters or future data values (observables), and current beliefs, $p(\boldsymbol{\omega})$, typically reduce to one or other of the familiar forms:

$p(\boldsymbol{\theta})$	initial beliefs about a parameter vector, $\boldsymbol{\theta}$;
$p(\boldsymbol{\theta} \mid \boldsymbol{x})$	beliefs about $\boldsymbol{\theta}$, given data $\boldsymbol{x}$;
$p(\boldsymbol{\psi} \mid \boldsymbol{x})$	beliefs about $\boldsymbol{\psi} = \boldsymbol{g}(\boldsymbol{\theta})$, given data $\boldsymbol{x}$;
$p(\boldsymbol{y} \mid \boldsymbol{x})$	beliefs about future data $\boldsymbol{y}$, given data $\boldsymbol{x}$.

1.4.1. *Point Estimates*

In cases where $\boldsymbol{\omega} \in \Omega$ corresponds to an unknown quantity, so that Ω is $\Re$, or $\Re^k$, or $\Re^+$, or $\Re \times \Re^+$, etc., and the required answer, $\boldsymbol{a} \in \mathcal{A}$, is an estimate of the true value of $\boldsymbol{\omega}$ (so that $\mathcal{A} = \Omega$), the corresponding decision problem is typically referred to as one of *point estimation*.

If $\boldsymbol{\omega} = \boldsymbol{\theta}$ or $\boldsymbol{\omega} = \boldsymbol{\psi}$, we refer to *parametric* point estimation; if $\boldsymbol{\omega} = \boldsymbol{y}$, we refer to *predictive* point estimation. Moreover, since one is almost certain not to get the answer exactly right in an estimation problem, statisticians typically work directly with the loss function concept, rather than with the utility function. A point estimation problem is thus completely defined once $\mathcal{A} = \Omega$ and $l(\boldsymbol{a}, \boldsymbol{\omega})$ are specified. Direct intuition suggests that in the one-dimensional case, distributional summaries such as the mean, median or mode of $p(\boldsymbol{\omega})$ could be reasonable point estimates of a random quantity $\boldsymbol{\omega}$. Clearly, however, these could differ considerably, and more formal guidance may be required as to when and why particular functionals of the belief distribution are justified as point estimates. This is provided by the following definition and result.

Definition 3. ***Bayes estimate***. *A Bayes estimate of $\boldsymbol{\omega}$ with respect to the loss function $l(\boldsymbol{a}, \boldsymbol{\omega})$ and the belief distribution $p(\boldsymbol{\omega})$ is an $\boldsymbol{a} \in \mathcal{A} = \Omega$ which minimises $\int_\Omega l(\boldsymbol{a}, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega}$.*

Theorem 2. ***Forms of Bayes estimates***.

(i) If $\mathcal{A} = \Omega = \Re^k$ and $l(\boldsymbol{a}, \boldsymbol{\omega}) = (\boldsymbol{a} - \boldsymbol{\omega})^t \boldsymbol{H} (\boldsymbol{a} - \boldsymbol{\omega})$, where $\boldsymbol{H}$ is symmetric and definite positive, the Bayes estimate satisfies

$$\boldsymbol{H}\boldsymbol{a} = \boldsymbol{H} E(\boldsymbol{\omega}).$$

If $\boldsymbol{H}^{-1}$ exists, $\boldsymbol{a} = E(\boldsymbol{\omega})$, and so ***the Bayes estimate with respect to quadratic form loss is the mean*** *of $p(\boldsymbol{\omega})$, assuming the mean to exist.*

(ii) If $\mathcal{A} = \Omega = \Re$ and $l(\boldsymbol{a}, \boldsymbol{\omega}) = c_1(a - \omega) 1_{(\omega \le a)}(a) + c_2(\omega - a) 1_{(\omega > a)}(a)$, the ***Bayes estimate with respect to linear loss is the quantile*** *such that*

$$P(\omega \le a) = c_2/(c_1 + c_2).$$

If $c_1 = c_2$, the right-hand side equals $1/2$ and so ***the Bayes estimate with respect to absolute value loss is a median*** *of $p(\boldsymbol{\omega})$.*

(iii) If $\mathcal{A} = \Omega \subseteq \Re^k$ and $l(\boldsymbol{a}, \boldsymbol{\omega}) = 1 - 1_{(B_\varepsilon(\boldsymbol{a}))}(\boldsymbol{\omega})$, where $B_\varepsilon(\boldsymbol{a})$ is a ball of radius ε in Ω centred at $\boldsymbol{a}$, the Bayes estimate maximises

$$\int_{B_\varepsilon(\boldsymbol{a})} p(\boldsymbol{\omega})\, d\boldsymbol{\omega}.$$

As $\varepsilon \to 0$, the function to be maximised tends to $p(\boldsymbol{a})$ and so ***the Bayes estimate with respect to zero-one loss is a mode*** *of $p(\boldsymbol{\omega})$, assuming a mode to exist.*

Proof. Differentiating $\int (\boldsymbol{a} - \boldsymbol{\omega})^t \boldsymbol{H} (\boldsymbol{a} - \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ with respect to $\boldsymbol{a}$ and equating to zero yields

$$2\boldsymbol{H} \int (\boldsymbol{a} - \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = 0.$$

This establishes (i). Since

$$\int l(\boldsymbol{a}, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = c_1 \int_{\{\omega \le a\}} (\boldsymbol{a} - \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega} + c_2 \int_{\{\omega > a\}} (\boldsymbol{\omega} - \boldsymbol{a}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega},$$

differentiating with respect to a and equating to zero yields

$$c_1 \int_{\{\omega \le a\}} p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = c_2 \int_{\{\omega > a\}} p(\boldsymbol{\omega})\, d\boldsymbol{\omega},$$

whence, adding $c_2 \int_{\omega \le a} p(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ to each side, we obtain (ii). Finally, since

$$\int l(\boldsymbol{a}, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = 1 - \int 1_{B_\varepsilon(a)}(\boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega},$$

and this is minimised when $\int_{B_\varepsilon(a)} p(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ is maximised, we have (iii).

$\triangleleft$

Further insight into the nature of case (iii) can be obtained by thinking of a unimodal, continuous $p(\omega)$ in one dimension. It is then immediate by a continuity argument that a should be chosen such that

$$p(a - \varepsilon) = p(a + \varepsilon).$$

In the case of a unimodal, symmetric belief distribution, $p(\omega)$, for a single random quantity ω, the mean, median and mode coincide. In general, for unimodal, positively skewed, densities we have the relation

$$\text{mean} > \text{median} > \text{mode}$$

and the difference can be substantial if $p(\omega)$ is markedly skew. Unless, therefore, there is a very clear need for a point estimate, and a strong rationale for a specific one of the loss functions considered in Theorem 2, the provision of a single number to summarise $p(\omega)$ may be extremely misleading as a summary of the information available about ω. Of course, such a comment acquires even greater force if $p(\omega)$ is multimodal or otherwise “irregular”.

For further discussion of Bayes estimators, see, for example, DeGroot and Rao (1963, 1966), Sacks (1963), Farrell (1964), Brown (1973), Tiao and Box (1974), Berger and Srinivasan (1978), Berger (1979, 1986), Hwang (1985, 1988), de la Horra (1987, 1988, 1992), Ghosh (1992a, 1992b), Irony (1992) and Spall and Maryak (1992).

1.4.2. *Credible regions*

We have emphasised that, from a theoretical perspective, uncertainty about an unknown quantity of interest, $\boldsymbol{\omega}$, needs to be communicated in the form of the full (prior, posterior or predictive) density, $p(\boldsymbol{\omega})$, if formal calculation of expected loss or utility is to be possible for any arbitrary future decision problem. In practice, however, $p(\boldsymbol{\omega})$ may be a somewhat complicated entity and it may be both more convenient, and also sufficient for general orientation regarding the uncertainty about $\boldsymbol{\omega}$, simply to describe regions $C \subseteq \Omega$ of given probability under $p(\boldsymbol{\omega})$. Thus, for example, in the case where $\Omega \subseteq \Re$, the identification of intervals containing 50%, 90%, 95% or 99% of the probability under $p(\omega)$ might suffice to give a good idea of the general quantitative messages implicit in $p(\omega)$. This is the intuitive basis of popular graphical representations of univariate distributions such as *box plots*.

Definition 4. ***Credible Region***. *A region $C \subseteq \Omega$ such that*

$$\int_C p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = 1 - \alpha$$

is said to be a $100(1-\alpha)\%$ *credible region for* $\boldsymbol{\omega}$*, with respect to* $p(\boldsymbol{\omega})$*. If* $\Omega \subseteq \Re$*, connected credible regions will be referred to as* ***credible intervals****. If* $p(\boldsymbol{\omega})$ *is a (prior-posterior-predictive) density, we refer to (prior-posterior-predictive) credible regions.*

Clearly, for any given α there is not a unique credible region—even if we restrict attention to connected regions, as we should normally wish to do for obvious ease of interpretation (at least in cases where $p(\boldsymbol{\omega})$ is unimodal). For given Ω, $p(\boldsymbol{\omega})$ and fixed α, the problem of choosing among the subsets $C \subseteq \Omega$ such that $\int_C p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = 1-\alpha$ could be viewed as a decision problem, provided that we are willing to specify a loss function, $l(C, \boldsymbol{\omega})$, reflecting the possible consequences of quoting the $100(1-\alpha)\%$ credible region C. We now describe the resulting form of credible region when a loss function is used which encapsulates the intuitive idea that, for given α, we would prefer to report a credible region C whose size $||C||$ (volume, area, length) is minimised.

Theorem 3. ***Minimal size credible regions***.
Let $p(\boldsymbol{\omega})$ *be a probability density for* $\boldsymbol{\omega} \in \Omega$ *almost everywhere continuous; given* α*,* $0 < \alpha < 1$*, if* $\mathcal{A} = \{C;\ P(\boldsymbol{\omega} \in C) = 1-\alpha\} \neq \emptyset$ *and*

$$l(C, \boldsymbol{\omega}) = k||C|| - 1_C(\boldsymbol{\omega}), \quad C \in \mathcal{A}, \quad \boldsymbol{\omega} \in \Omega, \quad k > 0,$$

then C *is optimal if and only if it has the property that* $p(\boldsymbol{\omega}_1) \geq p(\boldsymbol{\omega}_2)$ *for all* $\boldsymbol{\omega}_1 \in C$*,* $\boldsymbol{\omega}_2 \notin C$ *(except possibly for a subset of* Ω *of zero probability).*

Proof. It follows straightforwardly that, for any $C \in \mathcal{A}$,

$$\int_\Omega l(C, \boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = k||C|| + 1 - \alpha,$$

so that an optimal C must have minimal size.

If C has the stated property and D is any other region belonging to $\mathcal{A}$, then since $C = (C \cap D) \cup (C \cap D^c)$, $D = (C \cap D) \cup (C^c \cap D)$ and $P(\boldsymbol{\omega} \in C) = P(\boldsymbol{\omega} \in D)$, we have

$$\begin{aligned} \inf_{\boldsymbol{\omega} \in C \cap D^c} p(\boldsymbol{\omega}) ||C \cap D^c|| &\leq \int_{C \cap D^c} p(\boldsymbol{\omega})\, d\boldsymbol{\omega} \\ &= \int_{C^c \cap D} p(\boldsymbol{\omega})\, d\boldsymbol{\omega} \leq \sup_{\boldsymbol{\omega} \in C^c \cap D} p(\boldsymbol{\omega}) ||C^c \cap D|| \end{aligned}$$

with

$$\sup_{\boldsymbol{\omega} \in C^c \cap D} p(\boldsymbol{\omega}) \leq \inf_{\boldsymbol{\omega} \in C \cap D^c} p(\boldsymbol{\omega})$$

so that $||C \cap D^c|| \leq ||C^c \cap D||$, and hence $||C|| \leq ||D||$.

If C does not have the stated property, there exists $A \subseteq C$ such that for all $\boldsymbol{\omega}_1 \in A$, there exists $\boldsymbol{\omega}_2 \notin C$ such that $p(\boldsymbol{\omega}_2) > p(\boldsymbol{\omega}_1)$. Let $B \subseteq C^c$ be such that $P(\boldsymbol{\omega} \in A) = P(\boldsymbol{\omega} \in B)$ and $p(\boldsymbol{\omega}_2) > p(\boldsymbol{\omega}_1)$ for all $\boldsymbol{\omega}_2 \in B$ and $\boldsymbol{\omega}_1 \in A$. Define $D = (C \cap A^c) \cup B$. Then $D \in \mathcal{A}$ and by a similar argument to that given above the result follows by showing that $||D|| < ||C||$. The property of Theorem 3 is worth emphasising in the form of a definition (Box and Tiao, 1965).

Definition 5. ***Highest probability density (HPD) regions***.
A region $C \subseteq \Omega$ *is said to be a* $100(1-\alpha)\%$ *highest probability density region for* $\boldsymbol{\omega}$ *with respect to* $p(\boldsymbol{\omega})$ *if*

(i) $P(\boldsymbol{\omega} \in C) = 1 - \alpha$

(ii) $p(\boldsymbol{\omega}_1) \geq p(\boldsymbol{\omega}_2)$ *for all* $\boldsymbol{\omega}_1 \in C$ *and* $\boldsymbol{\omega}_2 \notin C$, *except possibly for a subset of* Ω *having probability zero.*

If $p(\boldsymbol{\omega})$ *is a (prior-posterior-predictive) density, we refer to highest (prior-posterior-predictive) density regions.*

Clearly, the credible region approach to summarising $p(\boldsymbol{\omega})$ is not particularly useful in the case of discrete Ω, since such regions will only exist for limited choices of α. The above development should therefore be understood as intended for the case of continuous Ω.

For a number of commonly occurring univariate forms of $p(\boldsymbol{\omega})$, there exist tables which facilitate the identification of HPD intervals for a range of values of α (see, for example, Isaacs *et al.*, 1974, Ferrándiz and Sendra,1982, and Lindley and Scott, 1985).

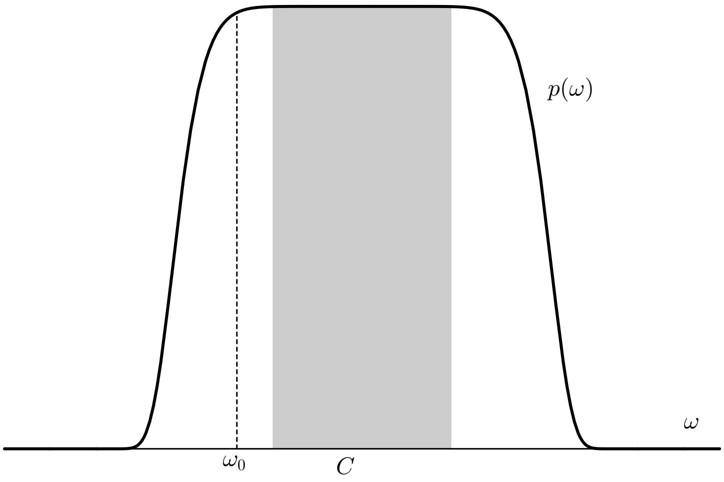

Figure 1a ω_0 *almost as "plausible" as all* $\omega \in C$

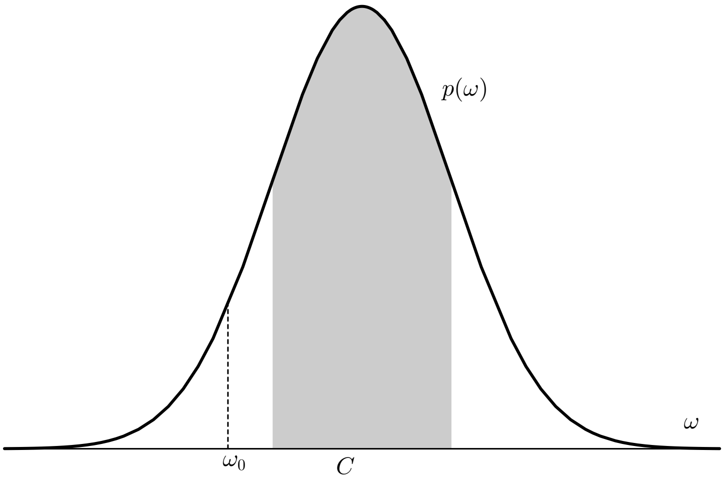

Figure 1b ω_0 *much less "plausible" than most* $\omega \in C$

Although an appropriately chosen selection of HDP regions can serve to give a useful summary of $p(\boldsymbol{\omega})$ when we focus just on the quantity $\boldsymbol{\omega}$, there is a fundamental difficulty which prevents such regions serving, in general, as a proxy for the actual density $p(\boldsymbol{\omega})$. The problem is that of lack of invariance under parameter transformation. Even if $\boldsymbol{v} = g(\boldsymbol{\omega})$ is a one-to-one transformation, it is easy to see that there is no general relation between HPD regions for $\boldsymbol{\omega}$ and $\boldsymbol{v}$. In addition, there is no way of identifying a marginal HPD region for a (possibly transformed) subset of components of $\boldsymbol{\omega}$ from knowledge of the joint HPD region.

In general the derivation of an HPD region requires numerical calculation and, particularly if $p(\boldsymbol{\omega})$ does not exhibit markedly skewed behaviour, it may be satisfactory in practice to quote some more simply calculated credible region. For example, in the univariate case, conventional statistical tables facilitate the identification of intervals which exclude equi-probable tails of $p(\boldsymbol{\omega})$ for many standard distributions. This form has the added advantage of being consistent under one-to-one reparametrisations of $\boldsymbol{\omega}$.

In cases where an HPD credible region C is pragmatically acceptable as a crude summary of the density $p(\boldsymbol{\omega})$, then, particularly for small values of α (for example, 0.05, 0.01), a specific value $\boldsymbol{\omega}_0 \in \Omega$ will tend to be regarded as somewhat "implausible" if $\boldsymbol{\omega}_0 \notin C$. This, of course, provides no justification for actions such as "rejecting the hypothesis that $\boldsymbol{\omega} = \boldsymbol{\omega}_0$". If we wish to consider such actions, we must formulate a proper decision problem, specifying alternative actions and the losses consequent on correct and incorrect actions. Inferences about a specific hypothesised value $\boldsymbol{\omega}_0$ of a random quantity $\boldsymbol{\omega}$ in the absence of alternative hypothesised values are often considered in the general statistical literature under the heading of "significance testing". For the present, it will suffice to note—as illustrated in Figure 1—that even the intuitive notion of "implausibility if $\boldsymbol{\omega}_0 \notin C$" depends much more on the complete characterisation of $p(\boldsymbol{\omega})$ than on an either-or assessment based on an HPD region.

For further discussion of credible regions see, for example, Pratt (1961), Aitchison (1964, 1966), Wright (1986) and DasGupta (1991).

1.4.3. *Hypothesis Testing*

The basic hypothesis testing problem usually considered by statisticians may be described as a decision problem with elements

$$\Omega = \{\omega_0 = [H_0 : \boldsymbol{\theta} \in \Theta_0], \quad \omega_1 = [H_1 : \boldsymbol{\theta} \in \Theta_1]\},$$

together with $p(\boldsymbol{\omega})$, where $\boldsymbol{\theta} \in \Theta = \Theta_0 \cup \Theta_1$, is the parameter labelling a parametric model, $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, $\mathcal{A} = \{a_0, a_1\}$, with $a_1(a_0)$ corresponding to rejecting hypothesis $H_0(H_1)$, and loss function $l(a_i, \omega_j) = l_{ij}$, $i, j \in \{0, 1\}$, with the l_{ij} reflecting the relative seriousness of the four possible consequences and, typically, $l_{00} = l_{11} = 0$.

General discussions of Bayesian hypothesis testing are included in Jeffreys (1939/1961), Good (1950, 1965, 1983), Lindley (1957, 1961b, 1965, 1977), Edwards *et al.* (1963), Pratt (1965), C. A. B. Smith (1965), Farrell (1968), Dickey (1971, 1974, 1977), Lempers (1971), H. Rubin (1971), Zellner (1971), DeGroot (1973), Leamer (1978), Box (1980), Shafer (1982b), Gilio and Scozzafava (1985), A. F. M. Smith, (1986), Berger and Delampady (1987), Berger and Sellke (1987), Hodges (1990, 1992) and Berger and Mortera (1991a, 1994).

1.5. IMPLEMENTATION ISSUES

Given a likelihood $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ and prior density $p(\boldsymbol{\theta})$, the starting point for any form of parametric inference summary or decision about $\boldsymbol{\theta}$ is the joint posterior density

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \,,$$

and the starting point for any predictive inference summary or decision about future observables $\boldsymbol{y}$ is the predictive density

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \int p(\boldsymbol{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\,d\boldsymbol{\theta}.$$

It is clear that to form these posterior and predictive densities there is a technical requirement to perform integrations over the range of $\boldsymbol{\theta}$. Moreover, further summarisation, in order to obtain marginal densities, or marginal moments, or expected utilities or losses in explicitly defined decision problems, will necessitate further integrations with respect to components of $\boldsymbol{\theta}$ or $\boldsymbol{y}$, or transformations thereof.

The key problem in implementing the formal Bayes solution to inference reporting or decision problems is therefore seen to be that of evaluating the required integrals. In cases where the likelihood just involves a single parameter, implementation just involves integration in one dimension and is essentially trivial. However, in problems involving a multiparameter likelihood the task of implementation is anything but trivial, since, if $\boldsymbol{\theta}$ has k components, two k-dimensional integrals are required just to form $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ and $p(\boldsymbol{y}\,|\,\boldsymbol{x})$. Moreover, in the case of $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$, for example, k $(k-1)$-dimensional integrals are required to obtain univariate marginal density values or moments, $\binom{k}{2}$ $(k-2)$-dimensional integrals are required to obtain bivariate marginal densities, and so on. Clearly, if k is at all large, the problem of implementation will, in general, lead to challenging technical problems, requiring simultaneous analytic or numerical approximation of a number of multidimensional integrals.

The above discussion has assumed a given specification of a likelihood and prior density function. However, as although a specific mathematical form for the likelihood in a given context is very often implied or suggested by consideration of symmetry, sufficiency or experience, the mathematical specification of prior densities is typically more problematic. Some of the problems involved—such as the pragmatic strategies to be adopted in translating actual beliefs into mathematical form—relate more to practical methodology than to conceptual and theoretical issues and will be not be discussed in this monograph. However, many of the other problems of specifying prior densities are closely related to the general problems of implementation described above, as exemplified by the following questions:

(i) if the information to be provided by the data is known to be far greater than that implicit in an individual's prior beliefs, is there any necessity for a precise mathematical representation of the latter, or can a Bayesian implementation proceed purely on the basis of this qualitative understanding?

(ii) either in the context of interpersonal analysis, or as a special form of actual individual analysis, is there a formal way of representing the beliefs of an individual whose prior information is to be regarded as minimal, relative to the information provided by the data?

Question (i) will be answered in Chapter 2, where an approximate "large sample" Bayesian theory involving *asymptotic posterior normality* will be presented.

Question (ii) will be answered in in Chapter 3, where the information-based concept of a *reference* prior density will be introduced. An extended historical discussion of this celebrated philosophical problem of how to represent "ignorance" will be given in Chapter 4.

2. Asymptotic Analysis

We know that in representations of belief models for observables involving a parametric model $p(x \mid \boldsymbol{\theta})$ and a prior specification $p(\boldsymbol{\theta})$, the parameter $\boldsymbol{\theta}$ acquired an operational meaning as some form of strong law limit of observables. Given observations $\boldsymbol{x} = (x_1, \ldots, x_n)$, the posterior distribution, $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, then describes beliefs about that strong law limit in the light of the information provided by $x_1, \ldots, x_n$. We now wish to examine various properties of $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ as the number of observations increases; *i.e.*, as $n \to \infty$. Intuitively, we would hope that beliefs about $\boldsymbol{\theta}$ would become more and more concentrated around the "true" parameter value; *i.e.*, the corresponding strong law limit. Under appropriate conditions, we shall see that this is, indeed, the case.

2.1. DISCRETE ASYMPTOTICS

We begin by considering the situation where $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \}$ consists of a countable (possibly finite) set of values, such that the parametric model corresponding to the true parameter, $\boldsymbol{\theta}_t$, is "distinguishable" from the others, in the sense that the logarithmic divergences, $\int p(x \mid \boldsymbol{\theta}_t) \log[p(x \mid \boldsymbol{\theta}_t)/p(x \mid \boldsymbol{\theta}_i)]\, dx$ are strictly larger than zero, for all $i \neq t$.

Theorem 4. ***Discrete asymptotics***.
Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be observations for which the parametric model $p(x \mid \boldsymbol{\theta})$ is defined, where $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots\}$, and the prior $p(\boldsymbol{\theta}) = \{p_1, p_2, \ldots\}$, $p_i > 0$, $\sum_i p_i = 1$. Suppose that $\boldsymbol{\theta}_t \in \Theta$ is the true value of $\boldsymbol{\theta}$ and that, for all $i \neq t$,

$$\int p(x \mid \boldsymbol{\theta}_t) \log \left[\frac{p(x \mid \boldsymbol{\theta}_t)}{p(x \mid \boldsymbol{\theta}_i)} \right] dx > 0;$$

then

$$\lim_{n \to \infty} p(\boldsymbol{\theta}_t \mid \boldsymbol{x}) = 1, \qquad \lim_{n \to \infty} p(\boldsymbol{\theta}_i \mid \boldsymbol{x}) = 0,\ i \neq t.$$

Proof. By Bayes' theorem, and assuming that $p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^n p(x_i|\boldsymbol{\theta})$,

$$\begin{aligned} p(\boldsymbol{\theta}_i \mid \boldsymbol{x}) &= p_i \, \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)}{p(\boldsymbol{x})} \\ &= \frac{p_i \, \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)/p(\boldsymbol{x} \mid \boldsymbol{\theta}_t)\}}{\sum_i p_i \, \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)/p(\boldsymbol{x} \mid \boldsymbol{\theta}_t)\}} \\ &= \frac{\exp\{\log p_i + S_i\}}{\sum_i \exp\{\log p_i + S_i\}}, \end{aligned}$$

where

$$S_i = \sum_{j=1}^n \log \frac{p(x_j \mid \boldsymbol{\theta}_i)}{p(x_j \mid \boldsymbol{\theta}_t)} \; .$$

Conditional on $\boldsymbol{\theta}_t$, the latter is the sum of n independent identically distributed random quantities and hence, by the strong law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n} S_i = \int p(x \mid \boldsymbol{\theta}_t) \log \left[\frac{p(x \mid \boldsymbol{\theta}_i)}{p(x \mid \boldsymbol{\theta}_t)} \right] dx.$$

The right-hand side is negative for all $i \neq t$, and equals zero for $i = t$, so that, as $n \to \infty$, $S_t \to 0$ and $S_i \to -\infty$ for $i \neq t$, which establishes the result.

$\triangleleft$

An alternative way of expressing the result of Theorem 3, established for countable Θ, is to say that the posterior distribution function for $\boldsymbol{\theta}$ ultimately degenerates to a step function with a single (unit) step at $\boldsymbol{\theta} = \boldsymbol{\theta}_t$. In fact, this result can be shown to hold, under suitable regularity conditions, for much more general forms of Θ. However, the proofs require considerable measure-theoretic machinery and the reader is referred to Berk (1966, 1970) for details.

A particularly interesting result is that if the true $\boldsymbol{\theta}$ is *not* in Θ, the posterior degenerates onto the value in Θ which gives the parametric model closest in logarithmic divergence to the true model.

2.2. CONTINUOUS ASYMPTOTICS

Let us now consider what can be said in the case of general Θ about the forms of probability statements implied by $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ for large n. Proceeding heuristically for the moment, without concern for precise regularity conditions, we note that, in the case of a parametric representation for an exchangeable sequence of observables,

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}) &\propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta}) \\ &\propto \exp\left\{ \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) \right\}. \end{aligned}$$

If we now expand the two logarithmic terms about their respective maxima, $\boldsymbol{m}_0$ and $\hat{\boldsymbol{\theta}}_n$, assumed to be determined by setting $\nabla \log p(\boldsymbol{\theta}) = 0$, $\nabla \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) = 0$, respectively, we obtain

$$\begin{aligned} \log p(\boldsymbol{\theta}) &= \log p(\boldsymbol{m}_0) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m}_0)^t \boldsymbol{H}_0 (\boldsymbol{\theta} - \boldsymbol{m}_0) + R_0 \\ \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) &= \log p(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}_n) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^t \boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) + R_n, \end{aligned}$$

where R_0, R_n denote remainder terms and

$$\boldsymbol{H}_0 = \left(-\frac{\partial^2 \log p(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{m}_0} \quad \boldsymbol{H}(\hat{\boldsymbol{\theta}}_n) = \left(-\frac{\partial^2 \log p(\boldsymbol{x} \mid \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \right) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}.$$

Assuming regularity conditions which ensure that R_0, R_n are small for large n, and ignoring constants of proportionality, we see that

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}) &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m}_0)^t \boldsymbol{H}_0 (\boldsymbol{\theta} - \boldsymbol{m}_0) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^t H(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) \right\} \\ &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m}_n)^t \boldsymbol{H}_n (\boldsymbol{\theta} - \boldsymbol{m}_n) \right\}, \end{aligned}$$

with

$$\begin{aligned}\boldsymbol{H}_n &= \boldsymbol{H}_0 + \boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)\\ \boldsymbol{m}_n &= \boldsymbol{H}_n^{-1}\left(\boldsymbol{H}_0\boldsymbol{m}_0 + \boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)\hat{\boldsymbol{\theta}}_n\right),\end{aligned}$$

where $\boldsymbol{m}_0$ (the *prior mode*) maximises $p(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_n$ (the *maximum likelihood estimate*) maximises $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$. The Hessian matrix, $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)$, measures the local curvature of the log-likelihood function at its maximum, $\hat{\boldsymbol{\theta}}_n$, and is often called the *observed information matrix*.

This heuristic development thus suggests that $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ will, for large n, tend to resemble a multivariate normal distribution, $\mathrm{N}_k(\boldsymbol{\theta}\,|\,\boldsymbol{m}_n, \boldsymbol{H}_n)$ whose mean is a matrix weighted average of a prior (modal) estimate and an observation-based (maximum likelihood) estimate, and whose precision matrix is the sum of the prior precision matrix and the observed information matrix.

Other approximations suggest themselves: for example, for large n the prior precision will tend to be small compared with the precision provided by the data and could be ignored. Also, since, by the strong law of large numbers, for all i, j,

$$\begin{aligned}\lim_{n\to\infty}\left\{\frac{1}{n}\left(-\frac{\partial^2 \log p(\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right\} &= \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{l=1}^{n}\left(-\frac{\partial^2 \log p(x_l\,|\,\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right\}\\ &= \int p(x\,|\,\boldsymbol{\theta})\left(-\frac{\partial^2 \log p(x\,|\,\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)dx\end{aligned}$$

we see that $H(\hat{\boldsymbol{\theta}}_n) \to n\boldsymbol{I}(\hat{\boldsymbol{\theta}}_n)$, where $\boldsymbol{I}(\boldsymbol{\theta})$, defined by

$$(\boldsymbol{I}(\boldsymbol{\theta}))_{ij} = \int p(x\,|\,\boldsymbol{\theta})\left(-\frac{\partial^2 \log p(x\,|\,\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)dx,$$

is the so-called *Fisher (or expected) information matrix*. We might approximate $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$, therefore, by either $\mathrm{N}_k(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}}_n, H(\hat{\boldsymbol{\theta}}_n))$ or $\mathrm{N}_k(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}}_n, nI(\hat{\boldsymbol{\theta}}_n))$, where k is the dimension of $\boldsymbol{\theta}$.

In the case of $\theta \in \Theta \subseteq \Re$,

$$H(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log p(\boldsymbol{x}\,|\,\theta),$$

so that the approximate posterior variance is the negative reciprocal of the rate of change of the first derivative of $\log p(\boldsymbol{x}\,|\,\theta)$ in the neighbourhood of its maximum. Sharply peaked log-likelihoods imply small posterior uncertainty and vice-versa.

There is a large literature on the regularity conditions required to justify mathematically the heuristics presented above. Those who have contributed to the field include: Laplace (1812), Jeffreys (1939/1961, Chapter 4), LeCam (1953, 1956, 1958, 1966, 1970, 1986), Lindley (1961b), Freedman (1963b, 1965), Walker (1969), Chao (1970), Dawid (1970), DeGroot (1970, Chapter 10), Ibragimov and Hasminski (1973), Heyde and Johnstone (1979), Hartigan (1983, Chapter 4), Bermúdez (1985), Chen (1985), Sweeting and Adekola (1987), Fu and Kass (1988), Fraser and McDunnough (1989), Sweeting (1992) and J. K. Ghosh *et al.* (1994). Related work on higher-order expansion approximations in which the normal appears as a leading term includes that of Hartigan (1965), R. A. Johnson (1967, 1970), Johnson and Ladalla (1979) and Crowder (1988). The account given below is based on Chen (1985).

In what follows, we assume that $\boldsymbol{\theta} \in \Theta \subseteq \Re^k$ and that $\{p_n(\boldsymbol{\theta}), n = 1, 2, \ldots\}$ is a sequence of posterior densities for $\boldsymbol{\theta}$, typically of the form $p_n(\boldsymbol{\theta}) = p(\boldsymbol{\theta}\,|\,x_1, \ldots, x_n)$, derived from an exchangeable sequence with parametric model $p(x\,|\,\boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$, although the mathematical development to be given does not require this. We define $L_n(\boldsymbol{\theta}) = \log p_n(\boldsymbol{\theta})$, and assume throughout that, for every n, there is a strict local maximum, $\boldsymbol{m}_n$, of p_n (or, equivalently, L_n) satisfying:

$$\boldsymbol{L}'_n(\boldsymbol{m}_n) = \nabla L_n(\boldsymbol{\theta}) \,|\, _{\boldsymbol{\theta}=\boldsymbol{m}_n} = 0$$

and implying the existence and positive-definiteness of

$$\Sigma_n = \left(-\boldsymbol{L}''_n(\boldsymbol{m}_n)\right)^{-1},$$

where $\left[\boldsymbol{L}''_n(\boldsymbol{m}_n)\right]_{ij} = \left(\partial^2 L_n(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j\right) \,|\, _{\boldsymbol{\theta}=\boldsymbol{m}_n}$.

Defining $|\,\boldsymbol{\theta}\,| = (\boldsymbol{\theta}^t\boldsymbol{\theta})^{1/2}$ and $B_\delta(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} \in \Theta;\, |\,\boldsymbol{\theta} - \boldsymbol{\theta}^*\,| < \delta\}$, we shall show that the following three basic conditions are sufficient to ensure a valid normal approximation for $p_n(\boldsymbol{\theta})$ in a small neighbourhood of $\boldsymbol{m}_n$ as n becomes large.

(c1) "*Steepness*". $\overline{\sigma}_n^2 \to 0$ as $n \to \infty$, where $\overline{\sigma}_n^2$ is the largest eigenvalue of Σ_n.

(c2) "*Smoothness*". For any $\varepsilon > 0$, there exists N and $\delta > 0$ such that, for any $n > N$ and $\boldsymbol{\theta} \in B_\delta(\boldsymbol{m}_n)$, $\boldsymbol{L}''_n(\boldsymbol{\theta})$ exists and satisfies

$$\boldsymbol{I} - \boldsymbol{A}(\varepsilon) \le \boldsymbol{L}''_n(\boldsymbol{\theta})\{\boldsymbol{L}''(\boldsymbol{m}_n)\}^{-1} \le \boldsymbol{I} + \boldsymbol{A}(\varepsilon),$$

where $\boldsymbol{I}$ is the $k \times k$ identity matrix and $\boldsymbol{A}(\varepsilon)$ is a $k \times k$ symmetric positive-semidefinite matrix whose largest eigenvalue tends to zero as $\varepsilon \to 0$.

(c3) "*Concentration*". For any $\delta > 0$, $\int_{B_\delta(\boldsymbol{m}_n)} p_n(\boldsymbol{\theta})d\boldsymbol{\theta} \to 1$ as $n \to \infty$.

Essentially, we shall see that (c1), (c2) together ensure that, for large n, inside a small neighbourhood of $\boldsymbol{m}_n$ the function p_n becomes highly peaked and behaves like the multivariate normal density kernel $\exp\{-\frac{1}{2}\,(\boldsymbol{\theta} - \boldsymbol{m}_n)^t\,\Sigma_n^{-1}\,(\boldsymbol{\theta} - \boldsymbol{m}_n)\}$. The final condition (c3) ensures that the probability outside any neighbourhood of $\boldsymbol{m}_n$ becomes negligible. We do not require any assumption that the $\boldsymbol{m}_n$ themselves converge, nor do we need to insist that $\boldsymbol{m}_n$ be a global maximum of p_n. We implicitly assume, however, that the limit of $p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{1/2}$ exists as $n \to \infty$, and we shall now establish a bound for that limit.

Theorem 5. ***Bounded concentration***.
The conditions (c1), (c2) imply that

$$\lim_{n\to\infty} p_n(\boldsymbol{m}_n)\,|\Sigma_n|^{1/2} \le (2\alpha)^{-k/2},$$

with equality if and only if (c3) holds.

Proof. Given $\varepsilon > 0$, consider $n > N$ and $\delta > 0$ as given in (c2). Then, for any $\boldsymbol{\theta} \in B_\delta(\boldsymbol{m}_n)$, a simple Taylor expansion establishes that

$$\begin{aligned} p_n(\boldsymbol{\theta}) &= p_n(\boldsymbol{m}_n)\exp\{L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{m}_n)\} \\ &= p_n(\boldsymbol{m}_n)\exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{m}_n)^t(\boldsymbol{I} + \boldsymbol{R}_n)\Sigma_n^{-1}(\boldsymbol{\theta} - \boldsymbol{m}_n)\right\}, \end{aligned}$$

where

$$\boldsymbol{R}_n = \boldsymbol{L}''_n(\boldsymbol{\theta}^+)\{\boldsymbol{L}''_n(\boldsymbol{m}_n)\}^{-1}(\boldsymbol{m}_n) - \boldsymbol{I},$$

for some $\boldsymbol{\theta}^+$ lying between $\boldsymbol{\theta}$ and $\boldsymbol{m}_n$. It follows that

$$P_n(\delta) = \int_{B_\delta(\boldsymbol{m}_n)} p_n(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

is bounded above by

$$P_n^+(\delta) = p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{\,1/2}\,|\,\boldsymbol{I} - \boldsymbol{A}(\varepsilon)\,|^{\,-1/2} \int_{|\,\boldsymbol{z}\,|<s_n} \exp\left\{-\tfrac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right\} d\boldsymbol{z}$$

and below by

$$P_n^-(\delta) = p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{\,1/2}\,|\,\boldsymbol{I} + \boldsymbol{A}(\varepsilon)\,|^{\,-1/2} \int_{|\,\boldsymbol{z}\,|<t_n} \exp\left\{-\tfrac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right\} d\boldsymbol{z},$$

where $s_n = \delta(1-\underline{\alpha}(\varepsilon))^{1/2}/\underline{\sigma}_n$ and $t_n = \delta(1+\underline{\alpha}(\varepsilon))^{1/2}/\overline{\sigma}_n$, with $\overline{\sigma}_n^2(\underline{\sigma}_n^2)$ and $\overline{\alpha}(\varepsilon)(\underline{\alpha}(\varepsilon))$ the largest (smallest) eigenvalues of Σ_n and $\boldsymbol{A}(\varepsilon)$, respectively, since, for any $k \times k$ matrix $\boldsymbol{V}$,

$$B_{\delta/\overline{V}}(0) \subseteq \left\{\boldsymbol{z}; (\boldsymbol{z}^t\boldsymbol{V}\boldsymbol{z})^{1/2} < \delta\right\} \subseteq B_{\delta/\underline{V}}(0),$$

where $\overline{V}^2(\underline{V}^2)$ are the largest (smallest) eigenvalues of $\boldsymbol{V}$.

Since (c1) implies that both s_n and t_n tend to infinity as $n \to \infty$, we have

$$\begin{aligned}|\boldsymbol{I} - \boldsymbol{A}(\varepsilon)|^{1/2} \lim_{n\to\infty} P_n(\delta) &\le \lim_{n\to\infty} p_n(\boldsymbol{m}_n)|\Sigma_n|^{1/2}(2\pi)^{k/2} \\ &\le |\boldsymbol{I} + \boldsymbol{A}(\varepsilon)|^{1/2} \lim_{n\to\infty} P_n(\delta),\end{aligned}$$

and the required inequality follows from the fact that $|\boldsymbol{I} \pm \boldsymbol{A}(\varepsilon)| \to 1$ as $\varepsilon \to 0$ and $P_n(\delta) \le 1$ for all n. Clearly, we have equality if and only if $\lim_{n\to\infty} P_n(\delta) = 1$, which is condition (c3). ◁

We can now establish the main result, which may colloquially be stated as "$\boldsymbol{\theta}$ has an asymptotic posterior $\mathrm{N}_k(\boldsymbol{\theta}|\boldsymbol{m}_n, \Sigma_n^{-1})$ distribution, where $\boldsymbol{L}_n'(\boldsymbol{m}_n) = 0$ and $\Sigma_n^{-1} = -\boldsymbol{L}_n''(\boldsymbol{m}_n)$."

Theorem 6. ***Asymptotic posterior normality*.**
For each n, consider $p_n(\cdot)$ as the density function of a random quantity $\boldsymbol{\theta}_n$, and define, $\boldsymbol{\phi}_n = \Sigma_n^{-1/2}(\boldsymbol{\theta}_n - \boldsymbol{m}_n)$. Then, given (c1) and (c2), (c3) is a necessary and sufficient condition for $\boldsymbol{\phi}_n$ to converge in distribution to $\boldsymbol{\phi}$, where $p(\boldsymbol{\phi}) = (2\pi)^{-k/2}\exp\left\{-\frac{1}{2}\boldsymbol{\phi}^t\boldsymbol{\phi}\right\}$.

Proof. Given (c1) and (c2), and writing $\boldsymbol{b} \ge \boldsymbol{a}$, for $\boldsymbol{a}, \boldsymbol{b} \in \Re^k$, to denote that all components of $\boldsymbol{b} - \boldsymbol{a}$ are non-negative, it suffices to show that, as $n \to \infty$, $P_n(\boldsymbol{a} \le \boldsymbol{\phi}_n \le \boldsymbol{b}) \to P(\boldsymbol{a} \le \boldsymbol{\phi} \le \boldsymbol{b})$ if and only if (c3) holds.

We first note that

$$P_n(\boldsymbol{a} \le \boldsymbol{\phi}_n \le \boldsymbol{b}) = \int_{\Theta_n} p_n(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where, by (c1), for any $\delta > 0$ and sufficiently large n,

$$\Theta_n = \left\{\boldsymbol{\theta}; \Sigma_n^{1/2}\boldsymbol{a} \le (\boldsymbol{\theta} - \boldsymbol{m}_n) \le \Sigma_n^{1/2}\boldsymbol{b}\right\} \subset B_\delta(\boldsymbol{m}_n).$$

It then follows, by a similar argument to that used in Theorem 5, that, for any $\varepsilon > 0$, $P_n(\boldsymbol{a} \leq \boldsymbol{\phi}_n \leq \boldsymbol{b})$ is bounded above by

$$P_n(\boldsymbol{m}_n)\,|\boldsymbol{I} - \boldsymbol{A}(\varepsilon)|^{-1/2}\,|\Sigma_n|^{1/2} \int_{Z(\varepsilon)} \exp\left\{-\tfrac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right\} d\boldsymbol{z},$$

where

$$Z(\varepsilon) = \left\{\boldsymbol{z}; [\boldsymbol{I} - \boldsymbol{A}(\varepsilon)]^{1/2}\,\boldsymbol{a} \leq \boldsymbol{z} \leq [\boldsymbol{I} - \boldsymbol{A}(\varepsilon)]^{1/2}\,\boldsymbol{b}\right\},$$

and is bounded below by a similar quantity with $+\boldsymbol{A}(\varepsilon)$ in place of $-\boldsymbol{A}(\varepsilon)$.

Given (c1), (c2), as $\varepsilon \to 0$ we have

$$\lim_{n\to\infty} P_n(\boldsymbol{a} \leq \boldsymbol{\phi}_n \leq \boldsymbol{b}) = \lim_{n\to\infty} p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{1/2} \int_{Z(0)} \exp\left\{-\tfrac{1}{2}\boldsymbol{z}^t\boldsymbol{z}\right\} d\boldsymbol{z},$$

where $Z(0) = \{\boldsymbol{z}; \boldsymbol{a} \leq \boldsymbol{z} \leq \boldsymbol{b}\}$. The result follows from Theorem 5.

◁

Conditions (c1) and (c2) are often relatively easy to check in specific applications, but (c3) may not be so directly accessible. It is useful therefore to have available alternative conditions which, given (c1), (c2), imply (c3). Two such are provided by the following:

(c4) For any $\delta > 0$, there exists an integer N and c, $d \in \Re^+$ such that, for any $n > N$ and $\boldsymbol{\theta} \notin B_\delta(\boldsymbol{m}_n)$,

$$L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{m}_n) < -c\left\{(\boldsymbol{\theta} - \boldsymbol{m}_n)^t\Sigma_n^{-1}(\boldsymbol{\theta} - \boldsymbol{m}_n)\right\}^d.$$

(c5) As (c4), but, with $G(\boldsymbol{\theta}) = \log g(\boldsymbol{\theta})$ for some density (or normalisable positive function) $g(\boldsymbol{\theta})$ over Θ,

$$L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{m}_n) < -c\,|\Sigma_n|^{-d} + G(\boldsymbol{\theta}).$$

Theorem 7. ***Alternative conditions***.
Given (c1), (c2), either (c4) or (c5) implies (c3).

Proof. It is straightforward to verify that

$$\int_{\Theta - B_\delta(\boldsymbol{m}_n)} p_n(\boldsymbol{\theta})d\boldsymbol{\theta} \leq p_n(\boldsymbol{m}_n)\,|\Sigma_n|^{1/2} \int_{|\,z\,| > \delta/\overline{\sigma}_n} \exp\left\{-c(\boldsymbol{z}^t\boldsymbol{z})^d\right\} d\boldsymbol{z},$$

given (c4), and similarly, that

$$\int_{\Theta - B_\delta(\boldsymbol{m}_n)} p_n(\boldsymbol{\theta})d\boldsymbol{\theta} \leq p_n(\boldsymbol{m}_n)\,|\Sigma_n|^{1/2}\,|\Sigma_n|^{-1/2} \exp\left\{-c\,|\Sigma_n|^{-d}\right\},$$

given (c4). Since $p_n(\boldsymbol{m}_n)\,|\,\Sigma_n\,|^{1/2}$ is bounded and the remaining terms or the right-hand side clearly tend to zero, it follows that the left-hand side tends to zero as $n \to \infty$.

◁

To understand better the relative ease of checking (c4) or (c5) in applications, we note that, if $p_n(\boldsymbol{\theta})$ is based on data $\boldsymbol{x}$,

$$L_n(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) - \log p(\boldsymbol{x}),$$

so that $L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{m}_n)$ does not involve the, often intractable, normalising constant $p(\boldsymbol{x})$. Moreover, (c4) does not even require the use of a proper prior for the vector $\boldsymbol{\theta}$.

We shall illustrate the use of (c4) for the general case of canonical conjugate analysis for exponential families.

Theorem 8. ***Asymptotic normality under conjugate analysis***.
Suppose that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are data resulting from a random sample of size n from the canonical exponential family form

$$p(\boldsymbol{y} \,|\, \boldsymbol{\psi}) = a(\boldsymbol{y}) \exp\left\{\boldsymbol{y}^t \boldsymbol{\psi} - b(\boldsymbol{\psi})\right\}$$

with a prior density of the form

$$p(\boldsymbol{\psi} \,|\, n_0, \boldsymbol{y}_0) = c(n_0, \boldsymbol{y}_0) \exp\left\{n_0 \boldsymbol{y}_0^t \boldsymbol{\psi} - n_0 b(\boldsymbol{\psi})\right\},$$

i.e., its canonical conjugate. For each n, consider the corresponding posterior density

$$p_n(\boldsymbol{\psi}) = p(\boldsymbol{\psi} \,|\, n_0 + n, n_0 \boldsymbol{y}_0 + n \overline{\boldsymbol{y}}_n),$$

with $\overline{\boldsymbol{y}}_n = \sum_{i=1}^n \boldsymbol{y}_i / n$, to be the density function for a random quantity $\boldsymbol{\psi}_n$, and define $\boldsymbol{\phi}_n = \Sigma_n^{-1/2}(\boldsymbol{\psi}_n - \boldsymbol{b}'(\boldsymbol{m}_n))$, where

$$\boldsymbol{b}'(\boldsymbol{m}_n) = \nabla b(\boldsymbol{\psi})\Big|_{\boldsymbol{\psi}=\boldsymbol{m}_n} = \frac{n_0 \boldsymbol{y}_0 + n \overline{\boldsymbol{y}}_n}{n_0 + n}$$

$$\left(\boldsymbol{b}''(\boldsymbol{m}_n)\right)_{ij} = \left(\frac{\partial^2 b(\boldsymbol{\psi})}{\partial \psi_i \partial \psi_j}\right)\Big|_{\boldsymbol{\psi}=\boldsymbol{m}_n} = (n_0 + n)\left(\Sigma_n^{-1}\right)_{ij}.$$

Then $\boldsymbol{\phi}_n$ converges in distribution to $\boldsymbol{\phi}$, where

$$p(\boldsymbol{\phi}) = (2\pi)^{-k/2} \exp\left\{-\tfrac{1}{2}\boldsymbol{\phi}^t \boldsymbol{\phi}\right\}.$$

Proof. Colloquially, we have to prove that $\boldsymbol{\psi}$ has an asymptotic posterior $\mathrm{N}_k(\boldsymbol{\psi} \,|\, \boldsymbol{b}'(\boldsymbol{m}_n), \Sigma_n^{-1})$ distribution, where $\boldsymbol{b}'(\boldsymbol{m}_n) = (n_0 + n)^{-1}(n_0 \boldsymbol{y}_0 + n \overline{\boldsymbol{y}}_n)$ and $\Sigma_n^{-1} = (n_0 + n)^{-1} \boldsymbol{b}''(\boldsymbol{m}_n)$. From a mathematical perspective,

$$p_n(\boldsymbol{\psi}) \propto \exp\left\{(n_0 + n) h(\boldsymbol{\psi})\right\},$$

where $h(\boldsymbol{\psi}) = [\boldsymbol{b}'(\boldsymbol{m}_n)]^t \boldsymbol{\psi} - b(\boldsymbol{\psi})$, with $b(\boldsymbol{\psi})$ a continuously differentiable and strictly convex function. It follows that, for each n, $p_n(\boldsymbol{\psi})$ is unimodal with a maximum at $\boldsymbol{\psi} = \boldsymbol{m}_n$ satisfying $\nabla h(\boldsymbol{m}_n) = 0$. By the strict concavity of $h(\cdot)$, for any $\delta > 0$ and $\theta \notin B_\delta(\boldsymbol{m}_n)$, we have, for some $\boldsymbol{\psi}^+$ between $\boldsymbol{\psi}$ and $\boldsymbol{m}_n$, with angle θ between $\boldsymbol{\psi} - \boldsymbol{m}_n$ and $\nabla h(\boldsymbol{\psi}^+)$,

$$\begin{aligned} h(\boldsymbol{\psi}) - h(\boldsymbol{m}_n) &= (\boldsymbol{\psi} - \boldsymbol{m}_n)^t \nabla h(\boldsymbol{\psi}^+) \\ &= |\, \boldsymbol{\psi} - \boldsymbol{m}_n \,| \;\; |\, \nabla h(\boldsymbol{\psi}^+) \,| \cos\theta \\ &< -c \,|\, \boldsymbol{\psi} - \boldsymbol{m}_n \,|, \end{aligned}$$

for $c = \inf \left\{ \,|\, \nabla h(\boldsymbol{\psi}^+) \,|\, ; \boldsymbol{\psi} \notin B_\delta(\boldsymbol{m}_n) \right\} > 0$. It follows that

$$\begin{aligned} L_n(\boldsymbol{\psi}) - L_n(\boldsymbol{m}_n) &< -(n_0 + n) \,|\, \boldsymbol{\psi} - \boldsymbol{m}_n \,| \\ &< -c_1 \left\{ (\boldsymbol{\psi} - \boldsymbol{m}_n)^t \Sigma_n^{-1} (\boldsymbol{\psi} - \boldsymbol{m}_n) \right\}^{1/2}, \end{aligned}$$

where $c_1 = c\lambda^{-1}$, with λ^2 the largest eigenvalue of $\boldsymbol{b}''(\boldsymbol{m}_n)$, and hence that (c4) is satisfied. Conditions (c1), (c2) follows straightforwardly from the fact that

$$\begin{aligned} (n_0 + n)\Sigma_n^{-1} &= \boldsymbol{b}''(m_n), \\ \boldsymbol{L}_n''(\boldsymbol{\psi})\{\boldsymbol{L}_n''(\boldsymbol{m}_n)\}^{-1} &= \boldsymbol{b}''(\boldsymbol{\psi})\{\boldsymbol{b}''(\boldsymbol{m}_n)\}^{-1}, \end{aligned}$$

the latter not depending on $n_0 + n$, and so the result follows by Theorems 6 and 7.

◁

Example 4. ***(Continued)***. Suppose that Be$(\theta \,|\, \alpha_n, \beta_n)$, where $\alpha_n = \alpha + r_n$, and $\beta_n = \beta + n - r_n$, is the posterior derived from n Bernoulli trials with r_n successes and a Be$(\theta \,|\, \alpha, \beta)$ prior. Proceeding directly,

$$\begin{aligned} L_n(\theta) = \log p_n(\theta) &= \log p(\boldsymbol{x} \,|\, \theta) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{x}) \\ &= (\alpha_n - 1)\log\theta + (\beta_n - 1)\log(1-\theta) - \log p(\boldsymbol{x}) \end{aligned}$$

so that

$$L_n'(\theta) = \frac{(\alpha_n - 1)}{\theta} - \frac{(\beta_n - 1)}{1 - \theta}$$

and

$$L_n''(\theta) = -\frac{(\alpha_n - 1)}{\theta^2} - \frac{(\beta_n - 1)}{(1 - \theta)^2}\,.$$

It follows that

$$m_n = \frac{\alpha_n - 1}{(\alpha_n + \beta_n - 2)}\,, \quad \left(-L_n''(m_n)\right)^{-1} = \frac{(\alpha_n - 1)(\beta_n - 1)}{(\alpha_n + \beta_n - 2)^3}\,.$$

Condition (c1) is clearly satisfied since $(-L_n''(m_n))^{-1} \to 0$ as $n \to \infty$; condition (c2) follows from the fact that $L_n''(\theta)$ is a continuous function of θ. Finally, (c4) may be verified with an argument similar to the one used in the proof of Theorem 6.

Taking $\alpha = \beta = 1$ for illustration (*i.e.*, a uniform prior density), we see that

$$m_n = \frac{r_n}{n}\,, \quad (-L_n''(m_n))^{-1} = \frac{1}{n} \cdot \frac{r_n}{n}\left(1 - \frac{r_n}{n}\right)\,,$$

and hence that the asymptotic posterior for θ is

$$\mathrm{N}\left(\theta \,\middle|\, \frac{r_n}{n}, \left\{\frac{1}{n} \cdot \frac{r_n}{n}\left(1 - \frac{r_n}{n}\right)\right\}^{-1}\right).$$

(As an aside, we note the interesting "duality" between this asymptotic form for θ given n, r_n, and the asymptotic distribution for r_n/n given θ, which, by the central limit theorem, has the form

$$\mathrm{N}\left(\frac{r_n}{n} \,\middle|\, \theta, \left\{\frac{1}{n}\theta(1-\theta)\right\}^{-1}\right).$$

□

2.3. ASYMPTOTICS UNDER TRANSFORMATIONS

The result of Theorem 8 is given in terms of the canonical parametrisation of the exponential family underlying the conjugate analysis. This prompts the obvious question as to whether the asymptotic posterior normality "carries over", with appropriate transformations of the mean and covariance, to an arbitrary (one-to-one) reparametrisation of the model. More generally, we could ask the same question in relation to Theorem 6. A partial answer is provided by the following.

Theorem 9. ***Asymptotic normality under transformation****.*
With the notation and background of Theorem 6, suppose that $\boldsymbol{\theta}$ *has an asymptotic* $\mathrm{N}_k(\boldsymbol{\theta}|\boldsymbol{m}_n, \Sigma_n^{-1})$ *distribution, with the additional assumptions that, with respect to a parametric model* $p(\boldsymbol{x}|\boldsymbol{\theta}_0)$, $\bar{\sigma}_n^2 \to 0$ *and* $\boldsymbol{m}_n \to \boldsymbol{\theta}_0$ *in probability, and* $\bar{\sigma}_n^2 = O_p(\underline{\sigma}_n^2)$, *where* $\bar{\sigma}_n^2$ $(\underline{\sigma}_n^2)$ *is the largest (smallest) eigenvalue of* Σ_n^2. *Then, if* $\boldsymbol{\nu} = \boldsymbol{g}(\boldsymbol{\theta})$ *is a transformation such that, at* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

is non-singular with continuous entries, $\boldsymbol{\nu}$ *has an asymptotic distribution*

$$\mathrm{N}_k\left(\boldsymbol{\nu}\;\middle|\; \boldsymbol{g}(\boldsymbol{m}_n), [\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{m}_n)\Sigma_n \boldsymbol{J}^t_{\boldsymbol{g}}(\boldsymbol{m}_n)]^{-1}\right).$$

Proof. This is a generalization and Bayesian reformulation of classical results presented in Serfling (1980, Section 3.3). For details, see Mendoza (1994).

◁

For any finite n, the adequacy of the normal approximation provided by Theorem 9 may be highly dependent on the particular transformation used. Anscombe (1964a, 1964b) analyses the choice of transformations which improve asymptotic normality. A related issue is that of selecting appropriate parametrisations for various numerical approximation methods (Hills and Smith, 1992, 1993).

The expression for the asymptotic posterior precision matrix (inverse covariance matrix) given in Theorem 9 is often rather cumbersome to work with. A simpler, alternative form is given by the following.

Corollary 1. ***Asymptotic precision after transformation****.*
In Theorem 9, if $\boldsymbol{H}_n = \Sigma_n^{-1}$ *denotes the asymptotic precision matrix for* $\boldsymbol{\theta}$, *then the asymptotic precision matrix for* $\boldsymbol{\nu} = \boldsymbol{g}(\boldsymbol{\theta})$ *has the form*

$$\boldsymbol{J}^t_{\boldsymbol{g}^{-1}}(\boldsymbol{g}(\boldsymbol{m}_n))\boldsymbol{H}_n \boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{g}(\boldsymbol{m}_n)),$$

where

$$\boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\nu}) = \frac{\partial \boldsymbol{g}^{-1}(\boldsymbol{\nu})}{\partial \boldsymbol{\nu}}$$

is the Jacobian of the inverse transformation.

Proof. This follows immediately by reversing of the roles of $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$.

$\triangleleft$

In many applications, we simply wish to consider one-to-one transformations of a single parameter. The next result provides a convenient summary of the required transformation result.

Corollary 2. ***Asymptotic normality after scalar transformation***.
Suppose that given the conditions of Theorems 6, and 9 with scalar θ, the sequence m_n tends in probability to θ_0 under $p(\boldsymbol{x}|\theta_0)$, and that $L''_n(m_n) \to 0$ in probability as $n \to \infty$. Then, if $\nu = g(\theta)$ is such that $g'(\theta) = dg(\theta)/\, d\theta$ is continuous and non-zero at $\theta = \theta_0$, the asymptotic posterior distribution for ν is

$$\mathrm{N}(\nu|g(m_n), -L''_n(m_n)[g'(m_n)]^{-2}).$$

Proof. The conditions ensure, by Theorem 6, that θ has an asymptotic posterior distribution of the form $\mathrm{N}(\theta|m_n, -L''_n(m_n))$, so that the result follows from Theorem 9.

$\triangleleft$

Example 4. ***(Continued)***. Suppose, again, that $\mathrm{Be}(\theta \mid \alpha_n, \beta_n)$, where $\alpha_n = \alpha + r_n$, and $\beta_n = \beta + n - r_n$, is the posterior distribution of the parameter of a Bernoulli distribution after n trials, and suppose now that we are interested in the asymptotic posterior distribution of the variance stabilising transformation

$$\nu = g(\theta) = 2 \sin^{-1} \sqrt{\theta}\,.$$

Straightforward application of Corollary 2 to Theorem 9, leads to the asymptotic distribution

$$\mathrm{N}(\nu|2 \sin^{-1}(\sqrt{r_n/n}), n).$$

□

It is clear from the presence of the term $[g'(m_n)]^{-2}$ in the form of the asymptotic precision given in Corollary 2 to Theorem 9 that things will go wrong if $g'(m_n) \to 0$ as $n \to \infty$. This is dealt with in the result presented by the requirement that $g'(\theta_0) \neq 0$, where $m_n \to \theta_0$ in probability. A concrete illustration of the problems that arise when such a condition is not met is given by the following.

Example 5. ***Non-normal asymptotic posterior***. Suppose that the asymptotic posterior for a parameter $\theta \in \Re$ is given by $\mathrm{N}(\theta|\bar{x}_n, n)$, $n\bar{x}_n = x_1 + \cdots + x_n$, perhaps derived from $\mathrm{N}(x_i|\theta, 1)$, $i = 1, \ldots, n$, with $\mathrm{N}(\theta|0, \lambda)$, having $\lambda \approx 0$. Now consider the transformation $\nu = g(\theta) = \theta^2$, and suppose that the actual value of θ generating the x_i through $\mathrm{N}(x_i|\theta, 1)$ is $\theta = 0$.

Intuitively, it is clear that ν cannot have an asymptotic normal distribution since the sequence $\bar{x}_n^2$ is converging in probability to 0 through *strictly positive* values. Technically, $g'(0) = 0$ and the condition of the corollary is not satisfied. In fact, it can be shown that the asymptotic posterior distribution of $n\nu$ is χ^2 in this case.

□

One attraction of the availability of the results given in Theorem 9 and its corollary is that verification of the conditions for asymptotic posterior normality (as in, for example, Theorem 6) may be much more straightforward under one choice of parametrisation of the likelihood than under another. The result given enables us to identify the posterior normal form for any convenient choice of parameters, subsequently deriving the form for the parameters of interest by straightforward transformation. An indication of the usefulness of this result is given in the following example (and further applications can be found in Chapter 3).

Example 6. ***Asymptotic posterior normality for a ratio***. Suppose that we have a random sample $x_1, \ldots, x_n$ from the model,

$$p(\boldsymbol{x} \,|\, \theta_1) = \prod_{i=1}^{n} \mathrm{N}(x_i|\theta_1, 1),$$

with prior $p(\theta_1) = \mathrm{N}(\theta_1|0, \lambda_1)$, and, independently, another random sample $y_1, \cdots, y_n$ from the model

$$p(\boldsymbol{y} \,|\, \theta_2) = \prod_{j=1}^{n} \mathrm{N}(y_j|\theta_1, 1)$$

with prior $p(\theta_2) = \mathrm{N}(\theta_2|0, \lambda_2)$, and let us further suppose that $\lambda_1 \approx 0$, $\lambda_2 \approx 0$ and $\theta_2 \neq 0$. We are interested in the posterior distribution of $\phi_1 = \theta_1/\theta_2$ as $n \to \infty$.

First, we note that, for large n, it is very easily verified that the joint posterior distribution for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is given by

$$\mathrm{N}_2\left\{\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \middle| \begin{pmatrix}\bar{x}_n\\ \bar{y}_n\end{pmatrix}, \begin{pmatrix} n & 0\\ 0 & n\end{pmatrix}\right\},$$

where $n\bar{x}_n = x_1 + \cdots + x_n$, $n\bar{y}_n = y_1 + \cdots + y_n$. Secondly, we note that the marginal asymptotic posterior for ϕ_1 can be obtained by defining an appropriate ϕ_2 such that $(\theta_1, \theta_2) \to (\phi_1, \phi_2)$ is a one-to-one transformation, obtaining the distribution of $\boldsymbol{\phi} = (\phi_1, \phi_2)$ using Theorem 9, and subsequently marginalising to ϕ_1.

An obvious choice for ϕ_2 is $\phi_2 = \theta_2$, so that, in the notation of Theorem 9, $\boldsymbol{g}(\theta_1, \theta_2) = (\phi_1, \phi_2)$ and

$$\boldsymbol{J_g}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial\phi_1}{\partial\theta_1} & \frac{\partial\phi_1/}{\partial\theta_2}\\ \frac{\partial\phi_2}{\partial\theta_1} & \frac{\partial\phi_2/}{\partial\theta_2}\end{pmatrix} = \begin{pmatrix} \frac{1}{\theta_2} & -\frac{\theta_1}{\theta_2^2}\\ 0 & 1\end{pmatrix}.$$

The determinant of this, θ_2^{-1}, is non-zero for $\theta_2 \neq 0$, and the conditions of Theorem 9 are clearly satisfied. It follows that the asymptotic posterior of $\boldsymbol{\phi}$ is

$$\mathrm{N}_2\left(\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix} \middle| \begin{pmatrix}\bar{x}_n/\bar{y}_n\\ \bar{y}_n\end{pmatrix},\quad n\bar{y}_n^2 \begin{pmatrix} 1 + (\bar{x}_n/\bar{y}_n)^2 & -\bar{x}_n\\ -\bar{x}_n & \bar{y}_n^2\end{pmatrix}^{-1}\right),$$

so that the required asymptotic posterior for $\phi_1 = \theta_1/\theta_2$ is

$$\mathrm{N}\left(\phi_1 \middle| \frac{\bar{x}_n}{\bar{y}_n},\quad n\bar{y}_n^2\left(\frac{\bar{y}_n^2}{\bar{x}_n^2 + \bar{y}_n^2}\right)\right).$$

Any reader remaining unappreciative of the simplicity of the above analysis may care to examine the form of the likelihood function, etc., corresponding to an initial parametrisation directly in terms of ϕ_1, ϕ_2, and to contemplate verifying directly the conditions of Theorem 6 using the ϕ_1, ϕ_2 parametrisation.

□

3. Reference Analysis

In Chapter 2, we have examined situations where data corresponding to large sample sizes come to dominate prior information, leading to inferences which are negligibly dependent on the initial state of information. We now turn to consider specifying prior distributions in situations where it is felt that, *even for moderate sample sizes*, the data should be expected to dominate prior information because of the "vague" nature of the latter. However, the problem of characterising a "*non-informative*" or "*objective*" prior distribution, representing "*prior ignorance*", "*vague prior knowledge*" and "*letting the data speak for themselves*" is far more complex than the apparent intuitive immediacy of these words and phrases would suggest.

In Chapter 4, we shall provide a brief review of the fascinating history of the quest for this "baseline", limiting prior form. It is important however to begin by making clear that "mere words" are an inadequate basis for clarifying such a slippery concept. Put bluntly: data cannot ever speak entirely for themselves; every prior specification has *some* informative posterior or predictive implications; and "vague" is itself much too vague an idea to be useful. There is no "objective" prior that represents ignorance. On the other hand, we all recognise that there *is* a pragmatically important need for a form of prior to posterior analysis capturing, *in some well-defined sense*, the notion of the prior having a minimal effect, relative to the data, on the final inference. Such a *reference analysis* might be required as an approximation to actual individual beliefs; more typically, it might be required as a limiting "what if?" baseline in considering a range of prior to posterior analyses, or as a *default* option when there are insufficient resources for detailed elicitation of actual prior knowledge.

The setting for our development of such a reference analysis will be the general decision-theoretic framework, together with the specific information-theoretic tools that have emerged as key measures of the discrepancies (or "distances") between probability distributions. From the approach we adopt, it will be clear that the *reference prior* component of the analysis is simply a mathematical tool. It has considerable pragmatic importance in implementing a *reference analysis*, whose role and character will be precisely defined, but it is not a privileged, "uniquely non-informative" or "objective" prior. Its main use will be to provide a "conventional" prior, to be used when a default specification having a claim to being *non-influential* in the sense described above is required. We seek to move away, therefore, from the rather philosophically muddled debates about "prior ignorance" that have all too often confused these issues, and towards well-defined decision-theoretic and information-theoretic procedures.

3.1. REFERENCE DECISIONS

Consider a specific form of decision problem with possible decisions $d \in \mathcal{D}$ providing possible answers, $a \in \mathcal{A}$, to an inference problem, with unknown state of the world $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$, utilities for consequences $(a, \boldsymbol{\omega})$ given by $u(d(\boldsymbol{\omega}_1)) = u(a, \boldsymbol{\omega}_1)$ and the availability of an experiment e which consists of obtaining an observation $\boldsymbol{x}$ having parametric model $p(\boldsymbol{x} \mid \boldsymbol{\omega}_2)$ and a prior probability density $p(\boldsymbol{\omega}) = p(\boldsymbol{\omega}_1 \mid \boldsymbol{\omega}_2)p(\boldsymbol{\omega}_2)$ for the unknown state of the world, $\boldsymbol{\omega}$. This general structure describes a situation where practical consequences depend directly on the $\boldsymbol{\omega}_1$ component of $\boldsymbol{\omega}$, whereas inference from data $\boldsymbol{x} \in X$ provided by experiment e takes place indirectly, through the $\boldsymbol{\omega}_2$ component of $\boldsymbol{\omega}$ as described by $p(\boldsymbol{\omega}_1 \,|\, \boldsymbol{\omega}_2)$. If $\boldsymbol{\omega}_1$ is a function of $\boldsymbol{\omega}_2$, the prior density is, of course, simply $p(\boldsymbol{\omega}_2)$.

To avoid subscript proliferation, let us now, without any risk of confusion, indulge in a harmless abuse of notation by writing $\boldsymbol{\omega}_1 = \boldsymbol{\omega}, \boldsymbol{\omega}_2 = \boldsymbol{\theta}$. This both simplifies the exposition and has the mnemonic value of suggesting that $\boldsymbol{\omega}$ is the state of the world of ultimate interest (since it occurs in the utility function), whereas $\boldsymbol{\theta}$ is a parameter in the usual sense (since it occurs in the probability model). Often $\boldsymbol{\omega}$ is just some function $\boldsymbol{\omega} = \boldsymbol{\phi}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$; if $\boldsymbol{\omega}$ is not a function of $\boldsymbol{\theta}$, the relationship between $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$ is that described in their joint distribution $p(\boldsymbol{\omega}, \boldsymbol{\theta}) = p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})$.

Now, for given conditional prior $p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})$ and utility function $u(a, \boldsymbol{\omega})$, let us examine, *in utility terms*, the influence of the prior $p(\boldsymbol{\theta})$, relative to the observational information provided by e. We note that if a_0^* denotes the optimal answer under $p(\boldsymbol{\omega})$ and a_x^* denotes the optimal answer under $p(\boldsymbol{\omega} \,|\, \boldsymbol{x})$, then the expected (utility) value of the experiment e, given the prior $p(\boldsymbol{\theta})$, is

$$v_u\{e, p(\boldsymbol{\theta})\} = \int p(\boldsymbol{x}) \int p(\boldsymbol{\omega} \,|\, \boldsymbol{x})\, u(a_x^*, \boldsymbol{\omega})\, d\boldsymbol{\omega}\, d\boldsymbol{x} - \int p(\boldsymbol{\omega})\, u(a_0^*, \boldsymbol{\omega})\, d\boldsymbol{\omega},$$

since

$$\int p(\boldsymbol{x}) \int u(a_0^*, \boldsymbol{\omega})\, p(\boldsymbol{\omega} \,|\, \boldsymbol{x}) d\boldsymbol{\omega} d\boldsymbol{x} = \int p(\boldsymbol{\omega})\, u(a_0^*, \boldsymbol{\omega})\, d\boldsymbol{\omega},$$

where, assuming $\boldsymbol{\omega}$ is independent of $\boldsymbol{x}$, given $\boldsymbol{\theta}$,

$$p(\boldsymbol{\omega}) = \int p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \qquad p(\boldsymbol{\omega} \,|\, \boldsymbol{x}) = \int \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})}{p(\boldsymbol{x})} p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

and

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

If $e(k)$ denotes the experiment consisting of k independent replications of e, that is yielding observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ with joint parametric model $\prod_{i=1}^k p(\boldsymbol{x}_i \,|\, \boldsymbol{\theta})$, then $v_u\{e(k), p(\boldsymbol{\theta})\}$, the expected utility value of the experiment $e(k)$, has the same mathematical form as $v_u\{e, p(\boldsymbol{\theta})\}$, but with $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ and $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^k p(\boldsymbol{x}_i \,|\, \boldsymbol{\theta})$. Intuitively, at least in suitably regular cases, as $k \to \infty$ we obtain, from $e(\infty)$, perfect (*i.e.*, complete) information about $\boldsymbol{\theta}$, so that, assuming the limit to exist,

$$v_u\{e(\infty), p(\boldsymbol{\theta})\} = \lim_{k \to \infty} v_u\{e(k), p(\boldsymbol{\theta})\}$$

is the expected (utility) *value of perfect information*, about $\boldsymbol{\theta}$, given $p(\boldsymbol{\theta})$.

Clearly, the more valuable the information contained in $p(\boldsymbol{\theta})$, the less will be the expected value of perfect information about $\boldsymbol{\theta}$; conversely, the less valuable the information contained in the prior, the more we would expect to gain from exhaustive experimentation. This, then, suggests a well-defined "thought experiment" procedure for characterising a "minimally valuable prior": choose, from the class of priors which has been identified as compatible with other assumptions about $(\boldsymbol{\omega}, \boldsymbol{\theta})$, that prior, $\pi(\boldsymbol{\theta})$, say, which *maximises the expected value of perfect information about* $\boldsymbol{\theta}$. Such a prior will be called a u-*reference prior*; the posterior distributions,

$$\pi(\boldsymbol{\omega} \,|\, \boldsymbol{x}) = \int p(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) d\boldsymbol{\theta}$$

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta})$$

derived from combining $\pi(\boldsymbol{\theta})$ with actual data $\boldsymbol{x}$, will be called *u-reference posteriors*; and the optimal decision derived from $\pi(\boldsymbol{\omega}\,|\,\boldsymbol{x})$ and $u(a, \boldsymbol{\omega})$ will be called a *u-reference decision*.

It is important to note that the limit above is *not* taken in order to obtain some form of asymptotic "approximation" to reference distributions; the "exact" reference prior is *defined* as that which maximises the value of *perfect* information about $\boldsymbol{\theta}$, *not* as that which maximises the expected value of the experiment.

Example 7. *Prediction with quadratic loss*. Suppose that a sequence of n observables, $\boldsymbol{x} = (x_1, \ldots, x_n)$, is assumed to be a random sample of size n from an $N(x\,|\,\mu, \lambda)$ parametric model, with known precision λ, and that a prior for μ is selected from the class

$$\{N(\mu\,|\,\mu_0, \lambda_0), \quad \mu_0 \in \Re, \quad \lambda_0 \geq 0\}.$$

Assuming a quadratic loss function, the decision problem is to provide a point estimate for x_{n+1}, given $x_1, \ldots, x_n$. We shall derive a reference analysis of this problem, for which $\mathcal{A} = \Re$, $\omega = x_{n+1}$, and $\theta = \mu$. Moreover,

$$u(a, \omega) = -(a - x_{n+1})^2, \quad p(\boldsymbol{x}\,|\,\theta) = \prod_{i=1}^{n} N(x_i\,|\,\mu, \lambda)$$

and, for given μ_0, λ_0, we have

$$p(\omega, \theta) = p(x_{n+1}, \mu) = p(x_{n+1}\,|\,\mu)p(\mu) = N(x_{n+1}\,|\,\mu, \lambda)N(\mu\,|\,\mu_0, \lambda_0).$$

For the purposes of the "thought experiment", let $\boldsymbol{z}_k = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ denote the (imagined) outcomes of k replications of the experiment yielding the observables $(x_1, \ldots, x_{kn})$, say, and let us denote the future observation to be predicted (x_{kn+1}) simply by x. Then

$$\begin{aligned} v_u\{e(k), N(\mu\,|\,\mu_0, \lambda_0)\} = &- \int p(\boldsymbol{z}_k) \inf_a \int p(x\,|\,\boldsymbol{z}_k)(a - x)^2 dx d\boldsymbol{z}_k \\ &+ \inf_a \int p(x)(a - x)^2 dx. \end{aligned}$$

However, we know from Theorem 2 that optimal estimates with respect to quadratic loss functions are given by the appropriate means, so that

$$\begin{aligned} v_u\{e(k), N(\mu\,|\,\mu_0, \lambda_0)\} &= - \int p(\boldsymbol{z}_k) V[x\,|\,\boldsymbol{z}_k] d\boldsymbol{z}_k + V[x] \\ &= -V[x\,|\,\boldsymbol{z}_k] + V[x], \end{aligned}$$

since, by virtue of the normal distributional assumptions, the predictive variance of x given $\boldsymbol{z}_k$ does not depend explicitly on $\boldsymbol{z}_k$. In fact, straightforward manipulations reveal that

$$\begin{aligned} v_u\{e(\infty), N(\mu\,|\,\mu_0, \lambda_0)\} &= \lim_{k \to \infty} v_u\{e(k), N(\mu\,|\,\mu_0, \lambda_0)\} \\ &= \lim_{k \to \infty} \left\{- \left[\lambda^{-1} + (\lambda_0 + kn\lambda)^{-1}\right] + (\lambda^{-1} + \lambda_0^{-1})\right\} = \lambda_0^{-1}, \end{aligned}$$

so that the *u-reference prior* corresponds to the choice $\lambda_0 = 0$, with μ_0 arbitrary.

Thus, given actual data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, the u-reference decision, *i.e.*, the reference prediction of the next observation under squared error loss, is simply the sample mean $\overline{x}$.

□

Example 8. ***Variance estimation***. Suppose that $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is assumed to be a random sample from $N(x \,|\, 0, \lambda)$, and that the prior for λ is selected form the class of gamma distributions centred on λ_0, so that $p(\lambda) = \text{Ga}(\lambda \,|\, \alpha, \alpha\lambda_0^{-1})$, $\alpha > 0$. The decision problem is to provide a point estimate for $\sigma^2 = \lambda^{-1}$, assuming a *standardised* quadratic loss function, so that

$$u(a, \sigma^2) = -\left[\frac{(a - \sigma^2)}{\sigma^2}\right]^2 = -(a\lambda - 1)^2.$$

Thus, we have $\mathcal{A} = \Re^+$, $\theta = \lambda$, $w = \sigma^2$, and

$$p(\boldsymbol{x}, \lambda) = \prod_{i=1}^{n} N(x_i \,|\, 0, \lambda)\, \text{Ga}(\lambda \,|\, \alpha, \alpha\lambda_0^{-1}).$$

Let $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ denote the outcome of k replications of the experiment. Then

$$\begin{aligned} v_u\{e(k), p(\lambda)\} = &-\int p(\boldsymbol{z}_k) \inf_a \int p(\lambda \,|\, \boldsymbol{z}_k)\, (a\lambda - 1)^2 \, d\lambda \, dz_k \\ &+ \inf_a \int p(\lambda)\, (a\lambda - 1)^2 \, d\lambda, \end{aligned}$$

where

$$p(\lambda) = \text{Ga}(\lambda \,|\, \alpha, \alpha\lambda_0^{-1}), \quad p(\lambda \,|\, z_k) = \text{Ga}\left(\lambda \,|\, \alpha + \frac{kn}{2}, \, \alpha\lambda_0^{-1} + \frac{kns^2}{2}\right),$$

and $kns^2 = \sum_i \sum_j x_{ij}^2$. Since

$$\inf_a \int \text{Ga}(\lambda \,|\, \alpha, \beta)\, (a\lambda - 1)^2 \, d\lambda = \frac{1}{\alpha + 1}\,,$$

and this is attained when $a = \beta/(\alpha + 1)$, one has

$$\begin{aligned} v_u\{e(\infty), p(\lambda)\} &= \lim_{k \to \infty} v_u\{e(k), p(\lambda)\} \\ &= \lim_{k \to \infty} \left\{ -\frac{1}{1 + \alpha + (kn)/2} + \frac{1}{1 + \alpha} \right\} = \frac{1}{1 + \alpha}\,. \end{aligned}$$

This is maximised when $\alpha = 0$ and, hence, the u-*reference prior* corresponds to the choice $\alpha = 0$, with λ_0 arbitrary. Given *actual* data, $\boldsymbol{x} = (x_1, \ldots, x_n)$, the u-*reference posterior* for λ is $\text{Ga}(\lambda \,|\, n/2, ns^2/2)$, where $ns^2 = \sum_i x_i^2$ and, thus, the u-*reference decision* is to give the estimate

$$\hat{\sigma^2} = \frac{ns^2/2}{(n/2) + 1} = \frac{\Sigma x_i^2}{n + 2}\,.$$

Hence, the reference estimator of σ^2 with respect to *standardised* quadratic loss is *not* the usual s^2, but a slightly smaller multiple of s^2.

It is of interest to note that, from a frequentist perspective, $\hat{\sigma}^2$ is the best invariant estimator of σ^2 and is admissible. Indeed, $\hat{\sigma}^2$ dominates s^2 or any smaller multiple of s^2 in terms of frequentist risk (cf. Example 45 in Berger, 1985a, Chapter 4). Thus, the u-reference approach has led to the “correct” multiple of s^2 as seen from a frequentist perspective.

□

Explicit reference decision analysis is possible when the parameter space is fine, so that $\Theta = \{\theta_1, \ldots, \theta_M\}$. In this case, the expected value of perfect information may be written as

$$v_u\{e(\infty), p(\theta)\} = \sum_{i=1}^{M} p(\theta_i) \sup_{\mathcal{D}} u(d(\theta_i)) - \sup_{\mathcal{D}} \sum_{i=1}^{M} p(\theta_i)\, u(d(\theta_i)),$$

and the u-reference prior, which is that $\pi(\theta)$ which maximises $v_u\{e(\infty), p(\theta)\}$, may be explicitly obtained by standard algebraic manipulations. For further information, see Bernardo (1981a) and Rabena (1998).

3.2. ONE-DIMENSIONAL REFERENCE DISTRIBUTIONS

It is known (Bernardo, 1979a) that reporting beliefs is itself a decision problem, where the "inference answer" space consists of the class of possible belief distributions that could be reported about the quantity of interest, and the utility function is a proper scoring rule which—in pure inference problems—may be identified with the logarithmic scoring rule.

Our development of reference analysis from now on will concentrate on this case, for which we simply denote $v_u\{\cdot\}$ by $v\{\cdot\}$, and replace the term "u-reference" by "reference".

In discussing reference decisions, we have considered a rather general utility structure where practical interest centred on a quantity $\boldsymbol{\omega}$ related to the $\boldsymbol{\theta}$ of an experiment by a conditional probability specification, $p(\boldsymbol{\omega} \mid \boldsymbol{\theta})$. Here, we shall consider the case where the quantity of interest is θ itself, with $\theta \in \Theta \subset \Re$. More general cases will be considered later.

If an experiment e consists of an observation $\boldsymbol{x} \in X$ having parametric model $p(\boldsymbol{x} \mid \theta)$, with $\boldsymbol{\omega} = \theta, \mathcal{A} = \{q(\cdot); q(\theta) > 0, \int_\Theta q(\theta) d\theta = 1\}$ and the utility function is the logarithmic scoring rule

$$u\{q(\cdot), \theta\} = A \log q(\theta) + B(\theta),$$

the expected utility value of the experiment e, given the prior density $p(\theta)$, is

$$v\{e, p(\theta)\} = \int p(\boldsymbol{x}) \int u\{q_x(\cdot), \theta\} p(\theta \mid \boldsymbol{x})\, d\theta d\boldsymbol{x} - \int u\{q_0(\cdot), \theta\}\, p(\theta)\, d\theta,$$

where $q_0(\cdot), q_x(\cdot)$ denote the optimal choices of $q(\cdot)$ with respect to $p(\theta)$ and $p(\theta \mid \boldsymbol{x})$, respectively. Noting that the logarithmic scoring rule $u\{q(\cdot), \theta\} = A \log q(\theta) + B(\theta)$, is a proper scoring rule, so that, for any $p(\theta)$,

$$\sup_q \int u\{q(\cdot), \theta\} p(\theta)\, d\theta = \int u\{p(\cdot), \theta\} p(\theta)\, d\theta,$$

it is easily seen that

$$v\{e, p(\theta)\} \propto \int p(\boldsymbol{x}) \int p(\theta \mid \boldsymbol{x}) \log \frac{p(\theta \mid \boldsymbol{x})}{p(\theta)}\, d\theta\, d\boldsymbol{x} = I\{e, p(\theta)\}$$

so that, with this utility function, the *value* to be expected from the experiment e becomes proportional to the *amount information* about θ which e may be expected to provide.

The corresponding expected information from the (hypothetical) experiment $e(k)$ yielding the (imagined) observation $\boldsymbol{z}_k = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ with parametric model

$$p(\boldsymbol{z}_k \mid \theta) = \prod_{i=1}^{k} p(\boldsymbol{x}_i \mid \theta)$$

is given by

$$I\{e(k), p(\theta)\} = \int p(\boldsymbol{z}_k) \int p(\theta \,|\, \boldsymbol{z}_k) \log \frac{p(\theta \,|\, \boldsymbol{z}_k)}{p(\theta)} \, d\theta d\boldsymbol{z}_k,$$

and so the expected (utility) value of perfect information about θ is

$$I\{e(\infty), p(\theta)\} = \lim_{k\to\infty} I\{e(k), p(\theta)\},$$

provided that this limit exists. This quantity measures the *missing information* about θ as a function of the prior $p(\theta)$.

The *reference prior* for θ, denoted by $\pi(\theta)$, is thus defined to be that prior which maximises the missing information functional. Given actual data $\boldsymbol{x}$, the *reference posterior* $\pi(\theta \,|\, \boldsymbol{x})$ to be reported is simply derived from Bayes' theorem, as $\pi(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\pi(\theta)$.

Unfortunately, $\lim_{k\to\infty} I\{e(k), p(\theta)\}$ is typically infinite (unless θ can only take a finite range of values) and a direct approach to deriving $\pi(\theta)$ along these lines cannot be implemented. However, a natural way of overcoming this technical difficulty is available: we derive the sequence of priors $\pi_k(\theta)$ which maximise $I\{e(k), p(\theta)\}, k = 1, 2, \ldots$, and subsequently take $\pi(\theta)$ to be a suitable limit. This approach will now be developed in detail.

Let e be the experiment which consists of one observation $\boldsymbol{x}$ from $p(\boldsymbol{x} \,|\, \theta)$, $\theta \in \Theta \subseteq \Re$. Suppose that we are interested in reporting inferences about θ and that no restrictions are imposed on the form of the prior distribution $p(\theta)$. It is easily verified that the amount of information about θ which k independent replications of e may be expected to provide may be rewritten as

$$I^\theta\{e(k), p(\theta)\} = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} \, d\theta,$$

where

$$f_k(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p(\theta \,|\, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\}$$

and $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is a possible outcome from $e(k)$, so that

$$p(\theta \,|\, \boldsymbol{z}_k) \propto \prod_{i=1}^{k} p(\boldsymbol{x}_i \,|\, \theta) p(\theta)$$

is the posterior distribution for θ after $\boldsymbol{z}_k$ has been observed. Moreover, for any prior $p(\theta)$ one must have the constraint $\int p(\theta) \, d\theta = 1$ and, therefore, the prior $\pi_k(\theta)$ which maximises $I^\theta\{e(k), p(\theta)\}$ must be an extremal of the functional

$$F\{p(\cdot)\} = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} d\theta + \lambda \left\{ \int p(\theta) \, d\theta - 1 \right\}.$$

Since this is of the form $F\{p(\cdot)\} = \int g\{p(\cdot)\} \, d\theta$, where, as a functional of $p(\cdot)$, g is twice continuously differentiable, any function $p(\cdot)$ which maximises F must satisfy the condition

$$\frac{\partial}{\partial \varepsilon} F\{p(\cdot) + \varepsilon\tau(\cdot)\} \Big|_{\varepsilon=0} = 0, \quad \text{for all } \tau.$$

It follows that, for any function τ,

$$\int \left\{ \tau(\theta) \log f_k(\theta) + \frac{p(\theta)}{f_k(\theta)} f'_k(\theta) - \tau(\theta) \left(1 + \log p(\theta)\right) + \tau(\theta)\lambda \right\} d\theta = 0,$$

where, after some algebra,

$$f'_k(\theta) = \frac{\partial}{\partial\varepsilon}\left\{\exp\left[\int p(\boldsymbol{z}_k\,|\,\theta)\log\frac{p(\boldsymbol{z}\,|\,\theta)\{p(\theta)+\varepsilon\tau(\theta)\}}{\int p(\boldsymbol{z}_k\,|\,\theta)\{p(\theta)+\varepsilon\tau(\theta)\}\,d\theta}d\boldsymbol{z}_k\right]\right\}\Big|_{\varepsilon=0}$$
$$= f_k(\theta)\frac{\tau(\theta)}{p(\theta)}\,.$$

Thus, the required condition becomes

$$\int \tau(\theta)\,\{\log f_k(\theta) - \log p(\theta) + \lambda\}\; d\theta = 0, \quad \text{for all } \tau(\theta),$$

which implies that the desired extremal should satisfy, for all $\theta \in \Theta$,

$$\log f_k(\theta) - \log p(\theta) + \lambda = 0$$

and hence that $p(\theta) \propto f_k(\theta)$.

Note that, for each k, this only provides an *implicit* solution for the prior which maximises $I^\theta\{e(k), p(\theta)\}$, since $f_k(\theta)$ depends on the prior through the posterior distribution $p(\theta\,|\,\boldsymbol{z}_k) = p(\theta\,|\,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k)$. However, for large values of k, an asymptotic approximation, $p^*(\theta\,|\,\boldsymbol{z}_k)$, say, may be found to the posterior distribution of θ, which *is* independent of the prior $p(\theta)$. It follows that, under suitable regularity conditions, the sequence of positive functions

$$p^*_k(\theta) = \exp\left\{\int p(\boldsymbol{z}_k\,|\,\theta)\log p^*(\theta\,|\,\boldsymbol{z}_k)d\boldsymbol{z}_k\right\}$$

will induce, by formal use of Bayes' theorem, a sequence of posterior distributions

$$\pi_k(\theta\,|\,\boldsymbol{x}) \propto p(\boldsymbol{x}\,|\,\theta)p^*_k(\theta)$$

with the same limiting distributions that would have been obtained from the sequence of posteriors derived from the sequence of priors $\pi_k(\theta)$ which maximise $I^\theta\{e(k), p(\theta)\}$. This completes our motivation for Definition 6. For further information see Bernardo (1979b) and ensuing discussion. For a concise introduction to these ideas, see Bernardo (1997a) and Bernardo and Ramón (1998).

Definition 6. ***One-dimensional reference distributions***.
Let $\boldsymbol{x}$ be the result of an experiment e which consists of one observation from

$$p(\boldsymbol{x}\,|\,\theta), \boldsymbol{x} \in X, \quad \theta \in \Theta \subseteq \Re,$$

let $\boldsymbol{z}_k = \{\boldsymbol{x}_1 \ldots, \boldsymbol{x}_k\}$ be the result of k independent replications of e, and define

$$f^*_k(\theta) = \exp\left\{\int p(\boldsymbol{z}_k\,|\,\theta)\log p^*(\theta\,|\,\boldsymbol{z}_k)d\boldsymbol{z}_k\right\},$$

where

$$p^*(\theta\,|\,\boldsymbol{z}_k) = \frac{\prod_{i=1}^k p(\boldsymbol{x}_i\,|\,\theta)}{\int \prod_{i=1}^k p(\boldsymbol{x}_i\,|\,\theta)\,d\theta}\,.$$

The ***reference posterior*** *density of θ after $\boldsymbol{x}$ has been observed, $\pi(\theta\,|\,\boldsymbol{x})$, is defined to be the limit in the convergence of information sense of*

$$\pi_k(\theta\,|\,\boldsymbol{x}) = c_k(\boldsymbol{x})p(\boldsymbol{x}\,|\,\theta)f^*_k(\theta),$$

where $c_k(\boldsymbol{x})$'s are the required normalising constants, assuming the limit to exist, i.e., such that, for all $\boldsymbol{x} \in X$,

$$\lim_{k\to\infty} \int \pi_k(\theta \,|\, \boldsymbol{x}) \log \frac{\pi_k(\theta \,|\, \boldsymbol{x})}{\pi(\theta \,|\, \boldsymbol{x})}\, d\theta = 0.$$

Any positive function $\pi(\theta)$ such that, for some $c(\boldsymbol{x}) > 0$ and all $\theta \in \Theta$,

$$\pi(\theta \,|\, \boldsymbol{x}) = c(\boldsymbol{x})\, p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta)$$

*will be called a **reference prior** for θ relative to the experiment e.*

It should be clear from the argument which motivates the definition that any asymptotic approximation to the posterior distribution may be used in place of the asymptotic approximation $p^*(\theta \,|\, \boldsymbol{z}_k)$ defined above. The use of convergence in the information sense, the natural convergence in this context, rather than just pointwise convergence, is necessary to avoid possibly pathological behaviour; for details, see Berger and Bernardo (1992c).

Although most of the following discussion refers to reference priors, it must be stressed that *only reference posterior* distributions are directly interpretable in probabilistic terms. The positive functions $\pi(\theta)$ are merely pragmatically convenient *tools* for the derivation of reference posterior distributions via Bayes' theorem. An explicit form for the reference prior is immediately available from Definition 7, and it will be clear from later illustrative examples that the forms which arise may have no direct probabilistic interpretation.

We should stress that the definitions and "Theorems" in this section are by and large heuristic in the sense that they are lacking statements of the technical conditions which would make the theory rigorous. Making the statements and proofs precise, however, would require a different level of mathematics from that used in this monograph, and is still an active area of research. The reader interested in the technicalities involved is referred to Berger and Bernardo (1989, 1992a, 1992b, 1992c) and Berger *et al.* (1989); see, also, Bernardo (1997a) and Bernardo and Ramón (1998).

Theorem 10. ***Explicit form of the reference prior***.
A reference prior for θ relative to the experiment which consists of one observation from $p(\boldsymbol{x} \,|\, \theta)$, $\boldsymbol{x} \in X$, $\theta \in \Theta \subseteq \Re$, is given, provided the limit exists, and convergence in the information sense is verified, by

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f_k^*(\theta)}{f_k^*(\theta_0)}\,, \qquad \theta \in \Theta$$

where $c > 0$, $\theta_0 \in \Theta$,

$$f_k^*(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

with $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ a random sample from $p(\boldsymbol{x} \,|\, \theta)$, and $p^*(\theta \,|\, \boldsymbol{z}_k)$ is an asymptotic approximation to the posterior distribution of θ.

Proof. Using $\pi(\theta)$ as a formal prior,

$$\pi(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\pi(\theta) \propto p(\boldsymbol{x} \,|\, \theta) \lim_{k\to\infty} \frac{f_k^*(\theta)}{f_k^*(\theta_0)} \propto \lim_{k\to\infty} \frac{p(\boldsymbol{x} \,|\, \theta) f_k^*(\theta)}{\int p(\boldsymbol{x} \,|\, \theta) f_k^*(\theta)\, d\theta}\,,$$

and hence

$$\pi(\theta \,|\, \boldsymbol{x}) = \lim_{k\to\infty} \pi_k(\theta \,|\, \boldsymbol{x}), \quad \pi_k(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta) f_k^*(\theta)$$

as required. Note that, under suitable regularity conditions, the limits above will not depend on the particular asymptotic approximation, as $k \to \infty$ to the posterior distribution used to derive $f_k^*(\theta)$.

◁

If the parameter space is finite, it turns out that the reference prior is uniform, independently of the experiment performed.

Theorem 11. ***Reference prior in the finite case***. *Let $\boldsymbol{x}$ be the result of one observation from $p(\boldsymbol{x} \,|\, \theta)$, where $\theta \in \Theta = \{\theta_1, \ldots, \theta_M\}$. Then, any function of the form $\pi(\theta_i) = a$, $a > 0$, $i = 1, \ldots, M$, is a reference prior and the reference posterior is*

$$\pi(\theta_i \,|\, \boldsymbol{x}) = c(\boldsymbol{x}) p(\boldsymbol{x} \,|\, \theta_i), \quad i = 1, \ldots, M$$

where $c(\boldsymbol{x})$ is the required normalising constant.

Proof. We have already established (Theorem 4) that if Θ is finite then, for any strictly positive prior, $p(\theta_i \,|\, x_1, \ldots, x_k)$ will converge to 1 if θ_i is the true value of θ. It follows that the integral in the exponent of

$$f_k(\theta_i) = \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta_i) \log p(\theta_i \,|\, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\}, \quad i = 1, \ldots, M,$$

will converge to zero as $k \to \infty$. Hence, a reference prior is given by

$$\pi(\theta_i) = \lim_{k\to\infty} \frac{f_k(\theta_i)}{f_k(\theta_j)} = 1.$$

The general form of reference prior follows immediately.

◁

The preceding result for the case of a finite parameter space is easily derived from first principles. Indeed, in this case the expected missing information is finite and equals the entropy

$$H\{p(\theta)\} = -\sum_{i=1}^{M} p(\theta_i) \log p(\theta_i)$$

of the prior. This is maximised if and only if the prior is uniform.

The technique encapsulated in Definition 6 for identifying the reference prior depends on the asymptotic behaviour of the posterior for the parameter of interest under (imagined) replications of the experiment to be actually analysed. Thus far, our derivations have proceeded on the basis of an assumed single observation from a parametric model, $p(\boldsymbol{x} \,|\, \theta)$. The next Theorem establishes that for experiments involving a sequence of $n \geq 1$ observations, which are to be modelled as if they are a random sample, conditional on a parametric model, the reference prior does not depend on the size of the experiment and can thus be derived on the basis of a single observation experiment. Note, however, that for experiments involving more structured designs (for example, in linear models) the situation is much more complicated.

Theorem 12. ***Independence of sample size.***
Let $e_n, n \geq 1$, be the experiment which consists of the observation of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in X$, $\theta \in \Theta$, and let $\mathcal{P}_n$ denote the class of reference priors for θ with respect to e_n, derived in accordance with Definition 6, by considering the sample to be a single observation from $\prod_{i=1}^n p(\boldsymbol{x}_i \,|\, \theta)$. Then $\mathcal{P}_1 = \mathcal{P}_n$, for all n.

Proof. If $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is the result of a k-fold independent replicate of e_1, then, by Theorem 10, $\mathcal{P}_1$ consists of $\pi(\theta)$ of the form

$$\pi(\theta) = c \lim_{k \to \infty} \frac{f_k^*(\theta)}{f_k^*(\theta_0)} ,$$

with $c > 0$, θ, $\theta_0 \in \Theta$ and

$$f_k^*(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{z}_k)\, d\boldsymbol{z}_k \right\},$$

where $p^*(\theta \,|\, \boldsymbol{z}_k)$ is an asymptotic approximation (as $k \to \infty$) to the posterior distribution of θ given $\boldsymbol{z}_k$.

Now consider $\boldsymbol{z}_{nk} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{2n}, \ldots, \boldsymbol{x}_{kn}\}$; this can be seen as the result of a k-fold independent replicate of e_n, so that $\mathcal{P}_n$ consists of $\pi(\theta)$ of the form

$$\pi(\theta) = c \lim_{k \to \infty} \frac{f_{nk}^*(\theta)}{f_{nk}^*(\theta_0)} .$$

But $\boldsymbol{z}_{nk}$ can equally be considered as a nk-fold independent replicate of e_1 and so the limiting ratios are clearly identical.

◁

In considering experiments involving random samples from distributions admitting a sufficient statistic of fixed dimension, it is natural to wonder whether the reference priors derived from the distribution of the sufficient statistic are identical to those derived from the joint distribution for the sample. The next theorem guarantees us that this is indeed the case.

Theorem 13. ***Compatibility with sufficient statistics.***
Let $e_n, n \geq 1$, be the experiment which consists of the observation of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in X, \theta \in \Theta$, where, for all n, the latter admits a sufficient statistic $\boldsymbol{t}_n = \boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Then, for any n, the classes of reference priors derived by considering replications of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{t}_n$ respectively, coincide, and are identical to the class obtained by considering replications of e_1.

Proof. If $\boldsymbol{z}_k$ denotes a k-fold replicate of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\boldsymbol{y}_k$ denotes the corresponding k-fold replicate of $\boldsymbol{t}_n$, then, by the definition of a sufficient statistic, $p(\theta \,|\, \boldsymbol{z}_k) = p(\theta \,|\, \boldsymbol{y}_k)$, for any prior $p(\theta)$. It follows that the corresponding asymptotic distributions are identical, so that $p^*(\theta \,|\, \boldsymbol{z}_k) = p^*(\theta \,|\, \boldsymbol{y}_k)$. We thus have

$$\begin{aligned} f_k^*(\theta) &= \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\} \\ &= \exp\left\{ \int p(\boldsymbol{z}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{y}_k) d\boldsymbol{z}_k \right\} \\ &= \exp\left\{ \int p(\boldsymbol{y}_k \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{y}_k) d\boldsymbol{y}_k \right\} \end{aligned}$$

so that, by Definition 6, the reference priors are identical. Identity with those derived from e_1 follows from Theorem 12.

◁

Given a parametric model, $p(\boldsymbol{x}\,|\,\theta)$, $x\in X$, $\theta\in\Theta$, we could, of course, reparametrise and work instead with $p(\boldsymbol{x}\,|\,\phi)$, $x\in X$, $\phi=\phi(\theta)$, for any monotone one-to-one mapping $g:\Theta\to\Phi$. The question now arises as to whether reference priors for θ and ϕ, derived from the parametric models $p(\boldsymbol{x}\,|\,\theta)$ and $p(\boldsymbol{x}\,|\,\phi)$, respectively, are consistent, in the sense that their ratio is the required Jacobian element. The next Theorem establishes this form of consistency and can clearly be extended to mappings which are piecewise monotone.

Theorem 14. ***Invariance under one-to-one transformations***.
Suppose that $\pi_\theta(\theta)$, $\pi_\phi(\phi)$ are reference priors derived by considering replications of experiments consisting of a single observation from $p(\boldsymbol{x}\,|\,\theta)$, with $\boldsymbol{x}\in X,\theta\in\Theta$ and from $p(\boldsymbol{x}\,|\,\phi)$, with $x\in X,\phi\in\Phi$, respectively, where $\phi=g(\theta)$ and $g:\Theta\to\Phi$ is a one-to-one monotone mapping. Then, for some $c>0$ and for all $\phi\in\Phi$:

(i) $\pi_\phi(\phi)=c\,\pi_\theta\left(g^{-1}(\phi)\right)$, *if Θ is discrete;*

(ii) $\pi_\phi(\phi)=c\,\pi_\theta\left(g^{-1}(\phi)\right)\,|\,J_\phi\,|$, *if* $J_\phi=\dfrac{\partial g^{-1}(\phi)}{\partial\phi}$ *exists.*

Proof. If Θ is discrete, so is Φ and the result follows from Theorem 11. Otherwise, if $\boldsymbol{z}_k$ denotes a k-fold replicate of a single observation from $p(\boldsymbol{x}\,|\,\theta)$, then, for any proper prior $p(\theta)$, the corresponding prior for ϕ is given by $p_\phi(\phi)=p_\theta\left(g^{-1}(\phi)\right)|J_\phi|$ and hence, for all $\phi\in\Phi$,

$$p_\phi(\phi\,|\,\boldsymbol{z}_k)=p_\theta\left(g^{-1}(\phi)\,|\,\boldsymbol{z}_k\right)|J_\phi|.$$

It follows that, as $k\to\infty$, the asymptotic posterior approximations are related by the same Jacobian element and hence

$$\begin{aligned} f_k^*(\theta)&=\exp\left\{\int p(\boldsymbol{z}_k\,|\,\theta)\log p^*(\theta\,|\,\boldsymbol{z}_k)d\boldsymbol{z}_k\right\}\\ &=|J_\phi|^{-1}\exp\left\{\int p(\boldsymbol{z}_k\,|\,\phi)\log p^*(\phi\,|\,\boldsymbol{z}_k)d\boldsymbol{z}_k\right\}\\ &=|J_\phi|^{-1}\,f_k^*(\phi). \end{aligned}$$

The second result now follows from Theorem 10.

◁

Corollary 1. ***Invariance under piecewise invertible functions***.
Let $p(\boldsymbol{x}\,|\,\theta)$, $\theta\in\Theta\subset\Re$, be a regular one-parameter model. If the quantity of interest $\phi=\phi(\theta)$ is piecewise invertible, then the corresponding reference prior $\pi_\phi(\theta)$ is the same as if θ were the parameter of interest.

Proof. Let $\phi=\phi(\theta)$, with $\phi(\theta)=\phi_i(\theta)$, $\theta\in\Theta_i$, where each of the $\phi_i(\theta)$'s is one-to-one in Θ_i; thus, $\theta=\{\phi,\omega\}$, where $\omega=i$ iff $\theta\in\Theta_i$. The reference prior $\pi_\phi(\theta)$ only depends on the asymptotic posterior of θ which, for sufficiently large samples, will concentrate on that subset Θ_i of the parameter space to which the true θ belongs. Since $\phi(\theta)$ is one-to-one within Θ_i and, by Theorem 14, reference priors are consistent under one-to-one reparametrizations, the stated result follows.

◁

The assumed existence of the asymptotic posterior distributions that would result from an imagined k-fold replicate of the experiment under consideration clearly plays a key role in the derivation of the reference prior. However, it is important to note that no assumption has thus far been required concerning the form of this asymptotic posterior distribution. As we shall see later, we shall typically consider the case of asymptotic posterior normality, but the following example shows that the technique is by no means restricted to this case.

Example 9. ***Uniform model***. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, be a random sample from a uniform distribution on $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, $\theta \in \Re$, and let $p(\theta)$ be some prior density for θ. If

$$\boldsymbol{t}_n = \left[x^{(n)}_{\min}, x^{(n)}_{\max}\right], \quad x^{(n)}_{\min} = \min\{x_1, \ldots, x_n\}, \quad x^{(n)}_{\max} = \max\{x_1, \ldots, x_n\},$$

then $\boldsymbol{t}_n$ is a sufficient statistic for θ, and

$$p(\theta \,|\, \boldsymbol{x}) = p(\theta \,|\, \boldsymbol{t}_n) \propto p(\theta), \qquad x^{(n)}_{\max} - \tfrac{1}{2} \le \theta \le x^{(n)}_{\min} + \tfrac{1}{2} \,.$$

It follows that, as $k \to \infty$, a k-fold replicate of e with a uniform prior will result in the posterior uniform distribution

$$p^*(\theta \,|\, \boldsymbol{t}_{kn}) \propto c, \qquad x^{(kn)}_{\max} - \tfrac{1}{2} \le \theta \le x^{(kn)}_{\min} + \tfrac{1}{2} \,.$$

It is easily verified that

$$\int p(\boldsymbol{t}_{kn} \,|\, \theta) \log p^*(\theta \,|\, \boldsymbol{t}_{kn}) d\boldsymbol{t}_{kn} = E\left[-\log\left\{1 - (x^{(kn)}_{\max} - x^{(kn)}_{\min})\right\} \,\middle|\, \theta\right] \,,$$

the expectation being with respect to the distribution of $\boldsymbol{t}_{kn}$. For large k, the right-hand side is well-approximated by

$$-\log\left\{1 - \left(E\left[x^{(kn)}_{\max}\right] - E\left[x^{(kn)}_{\min}\right]\right)\right\} \,,$$

and, noting that the distributions of

$$u = x^{(kn)}_{\max} - \theta - \tfrac{1}{2} \,, \quad v = x^{(kn)}_{\min} - \theta + \tfrac{1}{2}$$

are $\text{Be}(u \,|\, kn, 1)$ and $\text{Be}(v \,|\, 1, kn)$, respectively, we see that the above reduces to

$$-\log\left[1 - \frac{kn}{kn+1} + \frac{1}{kn+1}\right] = \log\left(\frac{kn+1}{2}\right) \,.$$

It follows that $f^*_{kn}(\theta) = (kn+1)/2$, and hence that

$$\pi(\theta) = c \lim_{k \to \infty} \frac{(kn+1)/2}{(kn+1)/2} = c \,.$$

Any reference prior for this problem is therefore a constant and, therefore, given a set of actual data $\boldsymbol{x} = (x_1, \ldots, x_n)$, the reference posterior distribution is

$$\pi(\theta \,|\, \boldsymbol{x}) \propto c \,, \qquad x^{(n)}_{\max} - \tfrac{1}{2} \le \theta \le x^{(n)}_{\min} + \tfrac{1}{2} \,,$$

a uniform density over the set of θ values which remain possible after $\boldsymbol{x}$ has been observed. □

Typically, under suitable regularity conditions, the asymptotic posterior distribution $p^*(\theta \mid \boldsymbol{z}_{kn})$, corresponding to an imagined k-fold replication of an experiment e_n involving a random sample of n from $p(\boldsymbol{x} \mid \theta)$, will only depend on $\boldsymbol{z}_{kn}$ through an *asymptotically sufficient, consistent estimate of* θ, a concept which is made precise in the next theorem. In such cases, the reference prior can easily be identified from the form of the asymptotic posterior distribution.

Theorem 15. ***Explicit form of the reference prior when there is an asymptotically sufficient, consistent estimator***.
Let e_n be the experiment which consists of the observation of a random sample $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from $p(\boldsymbol{x} \mid \theta), \boldsymbol{x} \in X, \theta \in \Theta \subseteq \Re$, and let $\boldsymbol{z}_{kn}$ be the result of a k-fold replicate of e_n. If there exists $\hat\theta_{kn} = \hat\theta_{kn}(\boldsymbol{z}_{kn})$ such that, with probability one

$$\lim_{k\to\infty} \hat\theta_{kn} = \theta,$$

and such that

$$\lim_{k\to\infty} \int p^*(\theta \mid \boldsymbol{z}_{kn}) \log \frac{p^*(\theta \mid \boldsymbol{z}_{kn})}{p^*(\theta \mid \hat\theta_{kn})} d\boldsymbol{z}_{kn} = 0,$$

then, for any $c > 0, \theta_0 \in \Theta$, reference priors are defined by

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f^*_{kn}(\theta)}{f^*_{kn}(\theta_0)} \,,$$

where

$$f^*_{kn}(\theta) = p^*(\theta \mid \hat\theta_{kn})\Big|_{\hat\theta_{kn}=\theta} \,.$$

Proof. As k increases, it follows from the assumptions that

$$\begin{aligned} f^*_{kn}(\theta) &= \exp\left\{\int p(\boldsymbol{z}_{kn} \mid \theta) \log p^*(\theta \mid \boldsymbol{z}_{kn}) d\boldsymbol{z}_{kn}\right\} \\ &= \exp\left\{\int p(\boldsymbol{z}_{kn} \mid \theta) \log p^*(\theta \mid \hat\theta_{kn}) d\boldsymbol{z}_{kn}\right\} \{1 + o(k)\} \\ &= \exp\left\{\int p(\hat\theta_{kn} \mid \theta) \log p^*(\theta \mid \hat\theta_{kn}) d\hat\theta_{kn}\right\} \{1 + o(k)\} \\ &= \exp\left\{\log p^*(\theta \mid \hat\theta_{kn})\Big|_{\hat\theta_{kn}=\theta}\right\} \{1 + o(k)\} \\ &= \; p^*(\theta \mid \hat\theta_{kn})\Big|_{\hat\theta_{kn}=\theta} \{1 + o(k)\}. \end{aligned}$$

The result now follows from Theorem 10.

◁

Example 10. ***Deviation from uniformity model***. Let e_n be the experiment which consists of obtaining a random sample from $p(x \mid \theta), 0 \le x \le 1, \theta > 0$, where

$$p(x \mid \theta) = \begin{cases} \theta\{2x\}^{\theta-1} & \text{for } \; 0 \le x \le \frac{1}{2} \\ \theta\{2(1-x)\}^{\theta-1} & \text{for } \; \frac{1}{2} \le x \le 1 \end{cases}$$

defines a one-parameter probability model on $[0,1]$, which finds application (see Bernardo and Bayarri, 1985) in exploring deviations from the standard uniform model on $[0,1]$ (which is given by $\theta = 1$).

It is easily verified that if $\boldsymbol{z}_{kn} = \{x_1, \ldots, x_{kn}\}$ results from a k-fold replicate of e_n, then a sufficient statistic t_{kn} is given by

$$t_{kn} = -\frac{1}{nk} \sum_{i=1}^{kn} \left\{ \log\{2x_i\} 1_{[0,1/2]}(x_i) + \log\{2(1-x_i)\}\, 1_{]1/2,1]}(x_i) \right\}$$

and, for any prior $p(\theta)$,

$$p(\theta \,|\, \boldsymbol{z}_{kn}) = p(\theta \,|\, t_{kn}) \propto p(\theta)\, \theta^{kn} \exp\{-kn(\theta-1)t_{kn}\}.$$

It is also easily shown that $p(t_{kn} \,|\, \theta) = \text{Ga}(t_{kn} \,|\, kn, kn\theta)$, so that

$$E[t_{kn} \,|\, \theta] = \frac{1}{\theta}\,, \quad V[t_{kn} \,|\, \theta] = \frac{1}{kn\theta^2}\,,$$

from which we can establish that $\hat{\theta}_{kn} = t_{kn}^{-1}$ is a sufficient, consistent estimate of θ. It follows that

$$p^*(\theta \,|\, \hat{\theta}_{kn}) \propto \theta^{kn} \exp\left\{ -\frac{kn(\theta-1)}{\hat{\theta}_{kn}} \right\}$$

provides, for large k, an asymptotic posterior approximation which satisfies the conditions required in Theorem 15. From the form of the right-hand side, we see that

$$\begin{aligned} p^*(\theta \,|\, \hat{\theta}_{kn}) &= \text{Ga}(\theta \,|\, kn+1, kn/\hat{\theta}_{kn}) \\ &= \frac{(kn/\hat{\theta}_{kn})^{kn+1}}{\Gamma(kn+1)}\, \theta^{kn} \exp\left\{ \frac{-kn\theta}{\hat{\theta}_{kn}} \right\}\,, \end{aligned}$$

so that

$$f_{kn}^*(\theta) = p^*(\theta \,|\, \hat{\theta}_{kn})\Big|_{\hat{\theta}_{kn}=\theta} = \frac{(kn)^{kn+1} e^{-nk}}{\Gamma(kn+1)\theta}\,,$$

and, from Theorem 10, for some $c > 0$, $\theta_0 > 0$,

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f_{kn}^*(\theta)}{f_{kn}^*(\theta_0)} = \frac{c\theta_0}{\theta} \propto \frac{1}{\theta}\,.$$

The reference posterior for θ having observed actual data $\boldsymbol{x} = (x_1, \ldots, x_n)$, producing the sufficient statistic $t = t(\boldsymbol{x})$, is therefore

$$\begin{aligned} \pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, t) &\propto p(\boldsymbol{x} \,|\, \theta)\, \frac{1}{\theta} \\ &\propto \theta^{n-1} \exp\{-n(\theta-1)t\}, \end{aligned}$$

which is a $\text{Ga}(\theta \,|\, n, nt)$ distribution.

□

Under regularity conditions, the asymptotic posterior distribution of θ tends to normality. In such cases, we can obtain a characterisation of the reference prior directly in terms of the parametric model in which θ appears.

Theorem 16. ***Reference priors under asymptotic normality***.
Let e_n be the experiment which consists of the observation of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $p(\boldsymbol{x} \mid \theta)$, $\boldsymbol{x} \in X$, $\theta \in \Theta \subset \Re$. Then, if the asymptotic posterior distribution of θ, given a k-fold replicate of e_n, is normal with precision $knh(\hat{\theta}_{kn})$, where $\hat{\theta}_{kn}$ is a consistent estimate of θ, reference priors have the form

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

Proof. Under regularity conditions, it follows that an asymptotic approximation to the posterior distribution of θ, given a k-fold replicate of e_n, is

$$p^*(\theta \mid \hat{\theta}_{kn}) = N\left(\theta \mid \hat{\theta}_{kn}, knh(\hat{\theta}_{kn})\right),$$

where $\hat{\theta}_{kn}$ is some consistent estimator of θ. Thus, by Theorem 15,

$$\begin{aligned} f^*_{kn}(\theta) &= p^*(\theta \mid \hat{\theta}_{kn})\Big|_{\hat{\theta}_{kn}=\theta} \\ &= (2\pi)^{-1/2}(kn)^{1/2}\{h(\theta)\}^{1/2}, \end{aligned}$$

and therefore, for some $c > 0$, $\theta_0 \in \Theta$,

$$\pi(\theta) = c \lim_{k\to\infty} \frac{f^*_{kn}(\theta)}{f^*_{kn}(\theta_0)} = \frac{\{h(\theta)\}^{1/2}}{\{h(\theta_0)\}^{1/2}} \propto \{h(\theta)\}^{1/2},$$

as required.

$\triangleleft$

The result of Theorem 16 is closely related to the "rules" independently proposed by Jeffreys (1946, 1939/1961) and by Perks (1947) to derive "non-informative" priors. Typically, under the conditions where asymptotic posterior normality obtains we find that

$$h(\theta) = \int p(\boldsymbol{x} \mid \theta)\left(-\frac{\partial^2}{\partial\theta^2}\log p(\boldsymbol{x} \mid \theta)\right) d\boldsymbol{x},$$

i.e., Fisher's information function (Fisher, 1925), and hence the reference prior,

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

becomes Jeffreys' (or Perks') prior. See Polson (1992a) for a related derivation.

It should be noted however that, even under conditions which guarantee asymptotic normality, Jeffreys' formula is not necessarily the easiest way of deriving a reference prior. As illustrated in Examples 9 and 10 above, it may be simpler to apply Theorem 10 using an directly derived asymptotic approximation to the posterior distribution.

It is important to stress that reference distributions are, by definition, a function of the *entire* probability model $p(\boldsymbol{x} \mid \theta)$, $\boldsymbol{x} \in X$, $\boldsymbol{\theta} \in \Theta$, not only of the observed likelihood.

Technically, this is a consequence of the fact that the amount of information which an experiment may be *expected* to provide is the value of an integral over the entire sample space X, which, therefore, has to be specified.

Example 11. ***Binomial and negative binomial models***. Consider an experiment which consists of the observation of n Bernoulli trials, with n fixed in advance, so that

$$p(x \,|\, \theta) = \theta^x (1-\theta)^{1-x}, \quad x \in \{0,1\}, \quad 0 \le \theta \le 1,$$

$$h(\theta) = -\sum_{x=0}^{1} p(x \,|\, \theta) \frac{\partial^2}{\partial \theta^2} \log p(x \,|\, \theta) = \theta^{-1}(1-\theta)^{-1},$$

and $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Hence, by Theorem 6, the reference prior is

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

If $r = \sum_{i=1}^n x_i$, the reference posterior,

$$\pi(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\pi(\theta) \propto \theta^{r-1/2}(1-\theta)^{n-r-1/2},$$

is the beta distribution $\text{Be}(\theta \,|\, r+\frac{1}{2}, n-r+\frac{1}{2})$. Note that $\pi(\theta \,|\, \boldsymbol{x})$ is proper, whatever the number of successes r. In particular, if $r = 0$, $\pi(\theta \,|\, \boldsymbol{x}) = \text{Be}(\theta \,|\, \frac{1}{2}, n+\frac{1}{2})$, from which sensible inference summaries can be made, *even though there are no observed successes*. (Compare this with the Haldane (1948) prior, $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, which produces an improper posterior until at least one success is observed.)

Consider now, however, an experiment which consists of counting the number x of Bernoulli trials which it is necessary to perform in order to observe a prespecified number of successes, $r \ge 1$. The probability model for this situation is the negative binomial

$$p(x \,|\, \theta) = \binom{x-1}{r-1} \theta^r (1-\theta)^{x-r}, \quad x = r, r+1, \ldots$$

from which we obtain

$$h(\theta) = -\sum_{x=r}^{\infty} p(x \,|\, \theta) \frac{\partial^2}{\partial \theta^2} \log p(x \,|\, \theta) = r\theta^{-2}(1-\theta)^{-1}$$

and hence, by Theorem 6, the reference prior is $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$. The reference posterior is given by

$$\pi(\theta \,|\, x) \propto p(x \,|\, \theta)\pi(\theta) \propto \theta^{r-1}(1-\theta)^{x-r-1/2}, \quad x = r, r+1, \ldots,$$

which is the beta distribution $\text{Be}(\theta \,|\, r, x-r+\frac{1}{2})$. Again, we note that this distribution is proper, whatever the number of observations x required to obtain r successes. Note that $r = 0$ is *not* possible under this model: the use of an inverse binomial sampling design implicitly assumes that r successes *will* eventually occur *for sure*, which is not true in direct binomial sampling. This difference in the underlying assumption about θ is duly reflected in the slight difference which occurs between the respective reference prior distributions.

Geisser (1984) and ensuing discussion provides further analysis and discussion of this canonical example. See also Bernard (1996).

□

In reporting results, scientists are typically required to specify not only the data but *also* the conditions under which the data were obtained (the *design* of the experiment), so that the data analyst has available the *full* specification of the probability model $p(\boldsymbol{x} \mid \theta)$, $\boldsymbol{x} \in X, \theta \in \Theta$. In order to carry out the reference analysis described in this chapter, such a full specification is clearly required.

We want to stress, however, that the preceding argument is totally compatible with a full personalistic view of probability. A reference prior is nothing but a (limiting) form of rather *specific* beliefs; namely, those which maximise the missing information which a *particular* experiment could possibly be expected to provide. Consequently, different experiments generally define different types of limiting beliefs. To report the corresponding reference posteriors (possibly for a range of possible alternative models) is only part of the general prior-to-posterior mapping which interpersonal or sensitivity considerations would suggest should always be carried out. Reference analysis provides an answer to an important "what if?" question: namely, what can be said about the parameter of interest *if* prior information were minimal *relative* to the information which infinite replications of a well-defined, specific experiment may be expected to provide.

3.3. RESTRICTED REFERENCE DISTRIBUTIONS

When analysing the inferential implications of the result of an experiment for a quantity of interest, θ, where, for simplicity, we continue to assume that $\theta \in \Theta \subseteq \Re$, it is often interesting, either *per se*, or on a "what if?" basis, to *condition* on some assumed features of the prior distribution $p(\theta)$, thus defining a restricted class, Q, say, of priors which consists of those distributions compatible with such conditioning. The concept of a reference posterior may easily be extended to this situation by maximising the missing information which the experiment may be expected to provide *within* this restricted class of priors.

Repeating the argument which motivated the definition of (unrestricted) reference distributions, we are led to seek the limit of the sequence of posterior distributions, $\pi_k(\theta \mid \boldsymbol{x})$, which correspond to the sequence of priors, $\pi_k(\theta)$, which are obtained by maximising, *within* Q, the amount of information

$$I\left\{e(k), p(\theta)\right\} = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)}\, d\theta,$$

where

$$f_k(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \theta) \log p(\theta \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

which could be expected from k independent replications $z = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ of the single observation experiment.

Definition 7. ***Restricted reference distributions*.**
Let $\boldsymbol{x}$ be the result of an experiment e which consists of one observation from $p(\boldsymbol{x} \mid \theta)$, $x \in X$, with $\theta \in \Theta \subseteq \Re$, let Q be a subclass of the class of all prior distributions for θ, let $\boldsymbol{z}_k = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ be the result of k independent replications of e and define

$$f_k^*(\theta) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \theta) \log p^*(\theta \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

where

$$p^*(\theta \mid \boldsymbol{z}_k) = \frac{\prod_{i=1}^k p(\boldsymbol{x}_i \mid \theta)}{\int \prod_{i=1}^k p(\boldsymbol{x}_i \mid \theta)\, d\theta}$$

The Q-reference posterior density of θ after $\boldsymbol{x}$ has been observed, $\pi^Q(\theta \,|\, \boldsymbol{x})$, is defined to be the limit, in the information convergence sense, of

$$\pi_k^Q(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, \pi_k^Q(\theta),$$

so that, assuming the limit to exist,

$$\lim_{k\to\infty} \int \pi_k^Q(\theta \,|\, \boldsymbol{x}) \log \frac{\pi_k^Q(\theta \,|\, \boldsymbol{x})}{\pi^Q(\theta \,|\, \boldsymbol{x})} = 0,$$

where $\pi_k^Q(\theta)$ is the prior which minimises, within Q the logarithmic divergence

$$\int p(\theta) \log \frac{p(\theta)}{f_k^*(\theta)}\, d\theta.$$

A positive function $\pi^Q(\theta)$ in Q such that

$$\pi^Q(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, \pi^Q(\theta), \quad \textit{for all } \theta \in \Theta,$$

is then called a Q-reference prior for θ relative to the experiment e.

The intuitive content of Definition 7 is illuminated by the following result, which essentially establishes that the Q-reference prior is the closest prior in Q to the unrestricted reference prior $\pi(\theta)$, in the sense of minimising its logarithmic divergence from $\pi(\theta)$.

Theorem 17. ***The restricted reference prior as an approximation*****.**
Suppose that an unrestricted reference prior $\pi(\theta)$ relative to a given experiment is proper; then, if it exists, a Q-reference prior $\pi_Q(\theta)$ satisfies

$$\int \pi^Q(\theta) \log \frac{\pi^Q(\theta)}{\pi(\theta)}\, d\theta = \inf_{p\in\mathcal{Q}} \int p(\theta) \log \frac{p(\theta)}{\pi(\theta)}\, d\theta.$$

Proof. It follows from Theorem 10 that $\pi(\theta)$ is proper if and only if

$$\int f_k^*(\theta)\; d\theta = c_k < \infty,$$

in which case,

$$\pi(\theta) = \lim_{k\to\infty} \pi_k(\theta) = \lim_{k\to\infty} c_k^{-1} f_k^*(\theta).$$

Moreover,

$$\begin{aligned}\int p(\theta) \log \frac{f_k^*(\theta)}{p(\theta)}\, d\theta &= -\int p(\theta) \log \frac{c_k^{-1} p(\theta)}{c_k^{-1} f_k^*(\theta)} d\theta \\ &= \log c_k - \int p(\theta) \log \frac{p(\theta)}{\pi_k(\theta)} d\theta,\end{aligned}$$

which is maximised if the integral is minimised. Let $\pi_k^Q(\theta)$ be the prior which minimises the integral within Q. Then, by Definition 7,

$$\pi^Q(\theta \,|\, x) \propto p(x \,|\, \theta) \lim_{k\to\infty} \pi_k^Q(\theta) = p(x \,|\, \theta)\pi^Q(\theta),$$

where, by the continuity of the divergence functional, $\pi^Q(\theta)$ is the prior which minimises, within Q,

$$\int p(\theta) \log \left\{ \frac{p(\theta)}{\lim\limits_{k\to\infty} \pi_k(\theta)} \right\} d\theta = \int p(\theta) \log \left\{ \frac{p(\theta)}{\pi(\theta)} \right\} d\theta.$$

$\triangleleft$

If $\pi(\theta)$ is not proper, it is necessary to apply Definition 7 directly in order to characterise $\pi^Q(\theta)$. The following result provides an explicit solution for the rather large class of problems where the conditions which define Q may be expressed as a collection of expected value restrictions.

Theorem 18. ***Explicit form of restricted reference priors***.
Let e be an experiment which provides information about θ, and, for given $\{(g_i(\cdot), \beta_i),$ $i = 1, \ldots, m\}$, let Q be the class of prior distributions $p(\theta)$ of θ which satisfy

$$\int g_i(\theta)p(\theta)d\theta = \beta_i, \quad i = 1, \ldots, m.$$

Let $\pi(\theta)$ be an unrestricted reference prior for θ relative to e; then, a Q-reference prior of θ relative to e, if it exists, is of the form

$$\pi^Q(\theta) \propto \pi(\theta) \exp\left\{\sum_{i=1}^m \lambda_i g_i(\theta)\right\},$$

where the λ_i's are constants determined by the conditions which define Q.

Proof. The calculus of variations argument which underlay the derivation of reference priors may be extended to include the additional restrictions imposed by the definition of Q, thus leading us to seek an extremal of the functional

$$\int p(\theta) \log \frac{f_k^*(\theta)}{p(\theta)}\, d\theta + \lambda \left\{\int p(\theta)\, d\theta - 1\right\} + \sum_{i=1}^m \lambda_i \left\{\int g_i(\theta)\, p(\theta)\, d\theta - \beta_i\right\},$$

corresponding to the assumption of a k-fold replicate of e. A standard argument now shows that the solution must satisfy

$$\log f_k^*(\theta) - \log p(\theta) + \lambda + \sum_{i=1}^m \lambda_i g_i(\theta) \equiv 0$$

and hence that

$$p(\theta) \propto f_k^*(\theta) \exp\left\{\sum_{i=1}^m \lambda_i g_i(\theta)\right\}.$$

Taking $k \longrightarrow \infty$, the result follows from Theorem 10.

◁

Example 12. ***Location models***. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a location model $p(x \mid \theta) = h(x - \theta)$, $x \in \Re$, $\theta \in \Re$, and suppose that the prior mean and variance of θ are restricted to be $E[\theta] = \mu_0$, $V[\theta] = \sigma_0^2$. Under suitable regularity conditions, the asymptotic posterior distribution of θ will be of the form $p^*(\theta \mid x_1, \ldots, x_n) \propto f(\hat{\theta}_n - \theta)$, where $\hat{\theta}_n$ is an asymptotically sufficient, consistent estimator of θ. Thus, by Theorem 15,

$$\pi(\theta) \propto p^*(\theta \mid \hat{\theta}_n)\Big|_{\hat{\theta}_n = \theta} \propto f(0),$$

which is constant, so that the unrestricted reference prior will be *uniform*. It now follows from Theorem 26 that the restricted reference prior will be

$$\pi^Q(\theta) \propto \exp\left\{\lambda_1\theta + \lambda_2(\theta - \mu_0)^2\right\},$$

with $\int \theta\pi^Q(\theta)\, d\theta = \mu_0$ and $\int (\theta - \mu_0)^2 \pi^Q(\theta)\, d\theta = \sigma_0^2$. Thus, the restricted reference prior, $\pi^Q(\theta)$, is the *normal* distribution $\mathrm{N}(\theta \mid \mu_0, \sigma_0^{-2})$, with the specified mean and variance.

□

3.4. NUISANCE PARAMETERS

The development given thus far has assumed that θ was one-dimensional and that interest was centred on θ or on a one-to-one transformation of θ. We shall next consider the case where $\boldsymbol{\theta}$ is two-dimensional and interest centres on reporting inferences for a one-dimensional function, $\phi = \phi(\boldsymbol{\theta})$. Without loss of generality, we may rewrite the vector parameter in the form $\boldsymbol{\theta} = (\phi, \lambda)$, $\phi \in \Phi$, $\lambda \in \Lambda$, where ϕ is the parameter of interest and λ is a nuisance parameter. The problem is to *identify a reference prior for $\boldsymbol{\theta}$, when the decision problem is that of reporting marginal inferences for ϕ,* assuming a logarithmic score (utility) function.

To motivate our approach to this problem, consider $\boldsymbol{z}_k$ to be the result of a k-fold replicate of the experiment which consists in obtaining a single observation, $\boldsymbol{x}$, from $p(\boldsymbol{x} \mid \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \phi, \lambda)$. Recalling that $p(\boldsymbol{\theta})$ can be thought of in terms of the decomposition

$$p(\boldsymbol{\theta}) = p(\phi, \lambda) = p(\phi)p(\lambda \mid \phi),$$

suppose, for the moment, that a *suitable reference form,* $\pi(\lambda \mid \phi)$, for $p(\lambda \mid \phi)$ has been specified and that only $\pi(\phi)$ remains to be identified. Theorem 10 then implies that the "marginal reference prior" for ϕ is given by

$$\pi(\phi) \propto \lim_{k \to \infty} \left[f_k^*(\phi)/f_k^*(\phi_0) \right], \quad \phi, \phi_0 \in \Phi,$$

where

$$f_k^*(\phi) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \phi) \log p^*(\phi \mid \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

$p^*(\phi \mid \boldsymbol{z}_k)$ is an asymptotic approximation to the marginal posterior for ϕ, and

$$\begin{aligned} p(\boldsymbol{z}_k \mid \phi) &= \int p(\boldsymbol{z}_k \mid \phi, \lambda)\pi(\lambda \mid \phi)\, d\lambda \\ &= \int \prod_{i=1}^{k} p(\boldsymbol{x}_i \mid \phi, \lambda)\pi(\lambda \mid \phi)\, d\lambda. \end{aligned}$$

By conditioning throughout on ϕ, we see from Theorem 10 that the "conditional reference prior" for λ given ϕ has the form

$$\pi(\lambda \mid \phi) \propto \lim_{k \to \infty} \left[\frac{f_k^*(\lambda \mid \phi)}{f_k^*(\lambda_0 \mid \phi)} \right], \quad \lambda, \lambda_0 \in \Lambda, \phi \in \Phi,$$

where

$$f_k^*(\lambda \mid \phi) = \exp\left\{ \int p(\boldsymbol{z}_k \mid \phi, \lambda) \log p^*(\lambda \mid \phi, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

$p^*(\lambda \mid \phi, \boldsymbol{z}_k)$ is an asymptotic approximation to the conditional posterior for λ given ϕ, and

$$p(\boldsymbol{z}_k \mid \phi, \lambda) = \prod_{i=1}^{k} p(\boldsymbol{x}_i \mid \phi, \lambda).$$

Given actual data $\boldsymbol{x}$, the marginal reference posterior for ϕ, corresponding to the reference prior

$$\pi(\boldsymbol{\theta}) = \pi(\phi, \lambda) = \pi(\phi)\pi(\lambda \mid \phi)$$

derived from the above procedure, would then be

$$\begin{aligned}\pi(\phi\,|\,\boldsymbol{x}) &\propto \int \pi(\phi,\lambda\,|\,\boldsymbol{x})\,d\lambda\\ &\propto \pi(\phi)\int p(\boldsymbol{x}\,|\,\phi,\lambda)\pi(\lambda\,|\,\phi)d\lambda.\end{aligned}$$

This would appear, then, to provide a straightforward approach to deriving reference analysis procedures in the presence of nuisance parameters. *However, there is a major difficulty*.

In general, as we have already seen, reference priors are typically *not* proper probability densities. This means that the integrated form derived from $\pi(\lambda\,|\,\phi)$,

$$p(\boldsymbol{z}_k\,|\,\phi) = \int p(\boldsymbol{z}_k\,|\,\phi,\lambda)\pi(\lambda\,|\,\phi)\,d\lambda,$$

which plays a key role in the above derivation of $\pi(\phi)$, will typically not be a proper probability model. The above approach cannot directly be applied in such cases, and a more subtle strategy is required to overcome this technical problem. However, before turning to the required details, we present an example, involving *finite* parameter ranges, where the approach outlined above does produce an interesting solution.

Example 13. ***Induction***. Consider a large, finite dichotomised population, all of whose elements individually may or may not have a specified property. A random sample is taken without replacement from the population, the sample being large in absolute size, but still relatively small compared with the population size. *All* the elements sampled turn out to have the specified property. Many commentators have argued that, in view of the large absolute size of the sample, one should be led to believe quite strongly that all elements of the *population* have the property, irrespective of the fact that the population size is greater still, an argument related to Laplace's rule of succession. (See, for example, Wrinch and Jeffreys, 1921, Jeffreys, 1939/1961, pp. 128–132 and Geisser, 1980a.)

Let us denote the population size by N, the sample size by n, the observed number of elements having the property by x, and the actual number of elements in the population having the property by θ. The probability model for the sampling mechanism is then the hypergeometric, which, for possible values of x, has the form

$$p(x\,|\,\theta) = \frac{\dbinom{\theta}{x}\dbinom{N-\theta}{n-x}}{\dbinom{N}{n}}\,.$$

If $p(\theta = r)$, $r = 0,\ldots,N$ defines a prior distribution for θ, the posterior probability that $\theta = N$, having observed $x = n$, is given by

$$p(\theta = N\,|\,x = n) = \frac{p(x = n\,|\,\theta = N)p(\theta = N)}{\sum_{r=n}^{N} p(x = n\,|\,\theta = r)p(\theta = r)}\,.$$

Suppose we considered θ to be the parameter of interest, and wished to provide a reference analysis. Then, since the set of possible values for θ is finite, Theorem 11 implies that

$$p(\theta = r) = \frac{1}{N+1},\quad r = 0,1,\ldots,N,$$

is the corresponding reference prior. Straightforward calculation then establishes that

$$p(\theta = N \,|\, x = n) = \frac{n+1}{N+1}\,,$$

which is *not* close to unity when n is large but n/N is small.

However, careful consideration of the problem suggests that it is *not* θ which is the *quantity of interest*: rather it is the parameter

$$\phi = \begin{cases} 1 & \text{if } \theta = N \\ 0 & \text{if } \theta \neq N. \end{cases}$$

To obtain a representation of θ in the form (ϕ, λ), let us define

$$\lambda = \begin{cases} \lambda_0 & \text{if } \theta = N \\ \theta & \text{if } \theta \neq N, \end{cases}$$

for some arbitrary λ_0. By Theorem 11, the reference priors $\pi(\phi)$ and $\pi(\lambda \,|\, \phi)$ are both uniform over the appropriate ranges, and are given by

$$\pi(\phi = 0) = \pi(\phi = 1) = \tfrac{1}{2}\,,$$

$$\pi(\lambda = \lambda_0 \,|\, \phi = 1) = 1, \quad \pi(\lambda = r \,|\, \phi = 0) = \frac{1}{N}\,, \quad r = 0, 1, \ldots, N-1.$$

These imply a reference prior for θ of the form

$$p(\theta) = \begin{cases} \dfrac{1}{2} & \text{if } \theta = N \\ \dfrac{1}{2N} & \text{if } \theta \neq N \end{cases}$$

and straightforward calculation establishes that

$$p(\theta = N \,|\, x = n) = \left[1 + \frac{1}{(n+1)}\left(1 - \frac{n}{N}\right)\right]^{-1} \approx \frac{n+1}{n+2}\,,$$

which clearly displays the irrelevance of the sampling fraction and the approach to unity for large n (see Bernardo, 1985b, for further discussion).

□

We return now to the general problem of defining a reference prior for $\boldsymbol{\theta} = (\phi, \lambda)$, $\phi \in \Phi$, $\lambda \in \Lambda$, where ϕ is the parameter vector of interest and λ is a nuisance parameter. We shall refer to the pair (ϕ, λ) as an *ordered parametrisation* of the model. We recall that the problem arises because in order to obtain the marginal reference prior $\pi(\phi)$ for the first parameter we need to work with the integrated model

$$p(\boldsymbol{z}_k \,|\, \phi) = \int p(\boldsymbol{z}_k \,|\, \phi, \lambda)\pi(\lambda \,|\, \phi)\, d\lambda.$$

However, this will only be a proper model if the conditional prior $\pi(\lambda \,|\, \phi)$ for the second parameter is a proper probability density and, typically, this will not be the case.

This suggests the following strategy: identify an increasing sequence $\{\Lambda_i\}$ of subsets of Λ, which may depend on ϕ, such that $\bigcup_i \Lambda_i = \Lambda$ and such that, on each Λ_i, the conditional reference prior, $\pi(\lambda \,|\, \phi)$ restricted to Λ_i can be normalised to give *proper* conditional reference prior $\pi_i(\lambda \,|\, \phi)$. For each i, a proper integrated model can then be obtained and a marginal reference prior $\pi_i(\phi)$ identified. The required reference prior $\pi(\phi, \lambda)$ is then obtained by taking the limit, as $i \to \infty$, of $\{\pi_i(\phi, \lambda) = \pi_i(\lambda \,|\, \phi)\, \pi_i(\phi)\}$ This strategy clearly requires a choice of the Λ_i's to be made, but in any specific problem a "natural" sequence usually suggests itself. We formalise this procedure in the next definition.

Definition 8. ***Reference distributions given a nuisance parameter***.
Let $\boldsymbol{x}$ be the result of an experiment e which consists of one observation from the probability model $p(\boldsymbol{x}\,|\,\phi,\lambda)$, $\boldsymbol{x}\in X$, $(\phi,\lambda)\in\Phi\times\Lambda\subset\Re\times\Re$. The reference posterior, $\pi(\phi\,|\,\boldsymbol{x})$, for the parameter of interest ϕ, relative to the experiment e and to the increasing sequences of subsets of Λ, $\{\Lambda_i(\phi)\}$, $\phi\in\Phi$, $\bigcup_i\Lambda_i(\phi)=\Lambda$, is defined to be the result of the following procedure:

(i) applying Definition 7 to the model $p(\boldsymbol{x}\,|\,\phi,\lambda)$, for fixed ϕ, obtain the conditional reference prior, $\pi(\lambda\,|\,\phi)$, for Λ;

(ii) for each ϕ, normalise $\pi(\lambda\,|\,\phi)$ within each $\Lambda_i(\phi)$ to obtain a sequence of proper priors, $\pi_i(\lambda\,|\,\phi)$;

(iii) use these to obtain a sequence of integrated models

$$p_i(\boldsymbol{x}_k\,|\,\phi)=\int_{\Lambda_i(\phi)}p(\boldsymbol{x}_k\,|\,\phi,\lambda)\pi_i(\lambda\,|\,\phi)\,d\lambda;$$

(iv) use those to derive the sequence of reference priors

$$\pi_i(\phi)=c\lim_{k\to\infty}\frac{f_k^*(\phi)}{f_k^*(\phi_0)},$$
$$f_k^*(\phi)=\exp\left\{\int p_i(\boldsymbol{z}_k\,|\,\phi)\log p^*(\phi\,|\,\boldsymbol{z}_k)d\boldsymbol{z}_k\right\},$$

and, for data $\boldsymbol{x}$, obtain the corresponding reference posteriors

$$\pi_i(\phi\,|\,\boldsymbol{x})\propto\pi_i(\phi)\int_{\Lambda_i(\phi)}p(\boldsymbol{x}\,|\,\phi,\lambda)\pi_i(\lambda\,|\,\phi)\,d\lambda;$$

(v) define $\pi(\phi\,|\,\boldsymbol{x})$ as the limit, in the convergence of information sense, of $\{\pi_i(\phi\,|\,\boldsymbol{x})\}$ i.e., such that

$$\lim_{i\to\infty}\int\pi_i(\phi\,|\,\boldsymbol{x})\log\frac{\pi_i(\phi\,|\,\boldsymbol{x})}{\pi(\phi\,|\,\boldsymbol{x})}\,d\phi=0.$$

The reference prior, relative to the ordered parametrisation (ϕ,λ), is any positive function $\pi(\phi,\lambda)$, such that

$$\pi(\phi\,|\,\boldsymbol{x})\propto\int p(\boldsymbol{x}\,|\,\phi,\lambda)\,\pi(\phi,\lambda)\,d\lambda.$$

This will typically be simply obtained as

$$\pi(\phi,\lambda)=\lim_{i\to\infty}\frac{\pi_i(\phi)\pi_i(\lambda\,|\,\phi)}{\pi_i(\phi_0)\pi_i(\lambda_0\,|\,\phi_0)}\,.$$

Ghosh and Mukerjee (1992a) showed that, in effect, the reference prior thus defined maximises the missing information about the parameter of interest, ϕ, subject to the condition that, given ϕ, the missing information about the nuisance parameter, λ, is maximised.

In a model involving a parameter of interest and a nuisance parameter, the form chosen for the latter is, of course, arbitrary. Thus, $p(\boldsymbol{x}\,|\,\phi,\lambda)$ can be written alternatively as $p(\boldsymbol{x}\,|\,\phi,\psi)$, for any $\psi=\psi(\phi,\lambda)$ for which the transformation $(\phi,\lambda)\to(\phi,\psi)$ is one-to-one. Intuitively, we would hope that the reference posterior for ϕ derived according to Definition 9 would not depend on the particular form chosen for the nuisance parameters. The following theorem establishes that this is the case.

Theorem 19. ***Invariance with respect to the choice of the nuisance parameter***. *Let e be an experiment which consists in obtaining one observation from model $p(\boldsymbol{x}\,|\,\phi, \lambda)$, $(\phi, \lambda) \in \Phi \times \Lambda \subset \Re \times \Re$, and let e' be an experiment which consists in obtaining one observation from $p(\boldsymbol{x}\,|\,\phi, \psi)$, $(\phi, \psi) \in \Phi \times \Psi \subseteq \Re \times \Re$, where $(\phi, \lambda) \to (\phi, \psi)$ is one-to-one transformation, with $\psi = g_\phi(\lambda)$. Then, the reference posteriors for ϕ, relative to $[e, \{\Lambda_i(\phi)\}]$ and $[e', \{\Psi_i(\phi)\}]$, where $\Psi_i(\phi) = g_\phi\{\Lambda_i(\phi)\}$, are identical.*

Proof. By Theorem 14, for given ϕ,

$$\pi_\psi(\psi\,|\,\phi) = \pi_\lambda(g_\phi^{-1}(\psi)\,|\,\phi)\,|\,J_{g_\phi^{-1}}(\psi)\,|\,,$$

where

$$\psi = g_\phi(\lambda), \qquad J_\psi(\phi) = \frac{\partial g_\phi^{-1}(\psi)}{\partial \psi}\,.$$

Hence, if we define

$$\Psi_i(\phi) = \{\psi;\ \psi = g_\phi(\lambda),\ \lambda \in \Lambda_i(\phi)\}$$

and normalise $\pi_\psi(\psi\,|\,\phi)$ over $\Psi_i(\phi)$ and $\pi_\lambda(g_\phi^{-1}(\psi)\,|\,\phi)$ over $\Lambda_i(\phi)$, we see that the normalised forms are consistently related by the appropriate Jacobian element. If we denote these normalised forms, for simplicity, by $\pi_i(\lambda\,|\,\phi)$, $\pi_i(\psi\,|\,\phi)$, we see that, for the integrated models used in steps (iii) and (iv) of Definition 8,

$$\begin{aligned} p_i(\boldsymbol{x}\,|\,\phi) &= \int_{\Lambda_i(\phi)} p(\boldsymbol{x}\,|\,\phi, \lambda)\pi_i(\lambda\,|\,\phi)\, d\lambda \\ &= \int_{\Psi_i(\phi)} p(\boldsymbol{x}\,|\,\phi, \psi)\pi_i(\psi\,|\,\phi)\, d\psi, \end{aligned}$$

and hence that the procedure will lead to identical forms of $\pi(\phi\,|\,\boldsymbol{x})$.

◁

Alternatively, we may wish to consider retaining the same form of nuisance parameter, λ, but redefining the parameter of interest to be a one-to-one function of ϕ. Thus, $p(\boldsymbol{x}\,|\,\phi, \lambda)$ might be written as $p(\boldsymbol{x}\,|\,\gamma, \lambda)$, where $\gamma = g(\phi)$ is now the parameter vector of interest. Intuitively, we would hope that the reference posterior for γ would be consistently related to that of ϕ by means of the appropriate Jacobian element. The next theorem establishes that this is indeed the case.

Theorem 20. ***Invariance under one-to-one transformations***.
Let e be an experiment which consists in obtaining one observation from $p(\boldsymbol{x}\,|\,\phi, \lambda)$, $\phi \in \Phi$, $\lambda \in \Lambda$, and let e' be an experiment which consists in obtaining one observation from $p(\boldsymbol{x}\,|\,\gamma, \lambda)$, $\gamma \in \Gamma, \lambda \in \Lambda$, where $\gamma = g(\phi)$. Then, given data $\boldsymbol{x}$, the reference posteriors for ϕ and γ, relative to $[e, \{\Lambda_i(\phi)\}]$ and $[e', \{\Phi_i(\gamma)\}]$, $\Phi_i(\gamma) = \Lambda_i\{g(\phi)\}$ are related by:

(i) $\pi_\gamma(\gamma\,|\,\boldsymbol{x}) = \pi_\phi(g^{-1}(\gamma)\,|\,\boldsymbol{x})$, *if Φ is discrete;*

(ii) $\pi_\gamma(\gamma\,|\,\boldsymbol{x}) = \pi_\phi(g^{-1}(\gamma)\,|\,\boldsymbol{x})\,|\,J_{g^{-1}}(\gamma)\,|\,$, *if* $J_{g^{-1}}(\gamma) = \dfrac{\partial g^{-1}(\gamma)}{\partial \gamma}$ *exists.*

Proof. In all cases, step (i) of Definition 8 clearly results in a conditional reference prior $\pi(\lambda\,|\,\phi) = \pi(\lambda\,|\,g^{-1}(\gamma))$. For discrete Φ, λ, $\pi_i(\phi)$ and $\pi_i(\gamma)$ defined by steps (ii)–(iv) of Definition 8 are both uniform distributions, by Theorem 10, and the result follows straightforwardly. If $J_{g^{-1}}(\gamma)$ exists, $\pi_i(\phi)$ and $\pi_i(\gamma)$ defined by steps (ii)–(iv) of Definition 8 are related by the claimed Jacobian element, $|\,J_{g^{-1}}(\gamma)\,|$, by Theorem 14, and the result follows immediately.

◁

In Theorem 15, we saw that the identification of explicit forms of reference prior can be greatly simplified if the approximate asymptotic posterior distribution is of the form

$$p^*(\theta \,|\, \boldsymbol{z}_k) = p^*(\theta \,|\, \hat{\theta}_k),$$

where $\hat{\theta}_k$ is an asymptotically sufficient, consistent estimate of θ. Theorem 16 establishes that even greater simplification results when the asymptotic distribution is normal. We shall now extend this to the nuisance parameter case.

Theorem 21. ***Bivariate reference priors under asymptotic normality.*** *Let e_n be the experiment which consists of the observation of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $p(\boldsymbol{x} \,|\, \phi, \lambda)$, $(\phi, \lambda) \in \Phi \times \Lambda \subseteq \Re \times \Re$, and let $\{\Lambda_i(\phi)\}$ be suitably defined sequences of subsets of λ, as required by Definition 8. Suppose that the joint asymptotic posterior distribution of (ϕ, λ), given a k-fold replicate of e_n, is multivariate normal with precision matrix $kn\boldsymbol{H}(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$, where $(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$ is a consistent estimate of (ϕ, λ) and suppose that $\hat{h}_{ij} = h_{ij}(\hat{\phi}_{kn}, \hat{\lambda}_{kn})$, $i = 1, 2$, $j = 1, 2$, is the partition of $\boldsymbol{H}$ corresponding to ϕ, λ. Then*

$$\pi(\lambda \,|\, \phi) \propto \{h_{22}(\phi, \lambda)\}^{1/2};$$
$$\pi(\phi, \lambda) = \pi(\lambda \,|\, \phi) \lim_{i\to\infty} \left\{ \frac{\pi_i(\phi) c_i(\phi)}{\pi_i(\phi_0) c_i(\phi_0)} \right\}, \quad \phi_0 \in \Phi,$$

define a reference prior relative to the ordered parametrisation (ϕ, λ), where

$$\pi_i(\phi) \propto \exp\left\{ \int_{\Lambda_i(\phi)} \pi_i(\lambda \,|\, \phi) \log\left(\{h_\phi(\phi, \lambda)\}^{1/2} \right) d\lambda \right\},$$

with

$$\pi_i(\lambda \,|\, \phi) = c_i(\phi) \pi(\lambda \,|\, \phi) = \frac{\pi(\lambda \,|\, \phi)}{\int_{\Lambda_i(\phi)} \pi(\lambda \,|\, \phi)\, d\lambda},$$

and

$$h_\phi = (h_{11} - h_{12} h_{22}^{-1} h_{21}).$$

Proof. The conditional distribution of λ given ϕ is asymptotically normal with precision $knh_{22}(\phi_{kn}, \hat{\lambda}_{kn})$. The first part of Theorem 21 then follows from Theorem 16.

The asymptotic marginal distribution of ϕ is univariate normal with precision $kn\hat{h}_\phi$, where $h_\phi = (h_{11} - h_{12} h_{22}^{-1} h_{21})$. To derive the form of $\pi_i(\phi)$, we note that if $\boldsymbol{z}_k \in Z$ denotes the result of a k-fold replication of e_n,

$$f^*_{kn}(\phi) = \exp\left\{ \int_Z \pi_i(\boldsymbol{z}_k \,|\, \phi) \log p^*(\phi \,|\, \boldsymbol{z}_k) d\boldsymbol{z}_k \right\},$$

where, with $\pi_i(\lambda \,|\, \phi)$ denoting the normalised version of $\pi(\lambda \,|\, \phi)$ over $\Lambda_i(\phi)$, the integrand has the form

$$\begin{aligned}
&\int_Z \left[\int_{\Lambda_i(\phi)} p(\boldsymbol{z}_k \,|\, \phi, \lambda) \pi_i(\lambda \,|\, \phi)\, d\lambda \right] \log N(\phi \,|\, \hat{\phi}_{kn}, kn\hat{h}_\phi) d\boldsymbol{z}_k \\
&\qquad = \int_{\Lambda_i(\phi)} \pi_i(\lambda \,|\, \phi) \left[\int_Z p(\boldsymbol{z}_k \,|\, \phi, \lambda) \log N(\phi \,|\, \hat{\phi}_{kn}, kn\hat{h}_\phi) d\boldsymbol{z}_k \right] d\lambda \\
&\qquad \approx \int_{\Lambda_i(\phi)} \pi_i(\lambda \,|\, \phi) \log \left[\frac{\{h_\phi(\phi, \lambda)\}}{2\pi} \right]^{1/2} d\lambda,
\end{aligned}$$

for large k, so that

$$\pi_i(\phi) = \lim_{k\to\infty} \frac{f^*_{kn}(\phi)}{f^*_{kn}(\phi_0)}$$

has the stated form. Since, for data $\boldsymbol{x}$, the reference prior $\pi(\phi, \lambda)$ is defined by

$$\begin{aligned}\pi(\phi \,|\, \boldsymbol{x}) &= \lim_{i\to\infty} \pi_i(\phi \,|\, \boldsymbol{x}) \propto \lim_{i\to\infty} p_i(\boldsymbol{x} \,|\, \phi)\pi_i(\phi)\\ &\propto \lim_{i\to\infty} \pi_i(\phi) \int_{\Lambda_i} p(\boldsymbol{x} \,|\, \phi, \lambda) c_i(\phi)\pi(\lambda \,|\, \phi) d\lambda\\ &\propto \int p(\boldsymbol{x} \,|\, \phi, \lambda)\pi(\phi, \lambda) d\lambda,\end{aligned}$$

the result follows.

◁

In many cases, the forms of $\{h_{22}(\phi, \lambda)\}$ and $\{h_\phi(\phi, \lambda)\}$ factorise into products of separate functions of ϕ and λ, and the subsets $\{\Lambda_i\}$ do not depend on ϕ. In such cases, the reference prior takes on a very simple form.

Corollary 1. ***Factorisation***.
Suppose that, under the conditions of Theorem 21, we choose a suitable increasing sequence of subsets $\{\Lambda_i\}$ *of* Λ*, which do not depend on* ϕ*, and suppose also that*

$$\{h_\phi(\phi, \lambda)\}^{1/2} = f_1(\phi) g_1(\lambda), \quad \{h_{22}(\phi, \lambda)\}^{1/2} = f_2(\phi) g_2(\lambda).$$

Then a reference prior relative to the ordered parametrisation (ϕ, λ) *is*

$$\pi(\phi, \lambda) \propto f_1(\phi) g_2(\lambda)$$

Proof. By Theorem 21, $\pi(\lambda \,|\, \phi) \propto f_2(\phi) g_2(\lambda)$, and hence

$$\pi_i(\lambda \,|\, \phi) = a_i \, g_2(\lambda),$$

where $a_i^{-1} = \int_{\Lambda_i} g_2(\lambda) \; d\lambda$. It then follows that

$$\begin{aligned}\pi_i(\phi) &\propto \exp\left\{ \int_{\Lambda_i} a_i g_2(\lambda) \log[f_1(\phi) g_1(\lambda)] \; d\lambda \right\}\\ &\propto b_i \, f_1(\phi),\end{aligned}$$

where $b_i = \int_{\Lambda_i} a_i g_2(\lambda) \log g_1(\lambda) \; d\lambda$, and the result easily follows.

◁

Example 14. ***Mean and standard deviation of a normal model***. Let e_n be the experiment which consists in the observation of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a normal distribution, with both mean, μ, and standard deviation, σ, unknown. We shall first obtain a reference analysis for μ, taking σ to be the nuisance parameter.

Since the distribution belongs to the exponential family, asymptotic normality obtains and the results of Theorem 21 can be applied. We therefore first obtain the Fisher (expected) information matrix, whose elements we recall are given by

$$h_{ij}(\mu, \sigma) = \int \mathrm{N}(x \,|\, \mu, \sigma^{-2}) \left\{ -\frac{\partial^2 \log N(x \,|\, \mu, \sigma^{-2})}{\partial\theta_i \partial\theta_j} \right\} dx,$$

from which it is easily verified that the asymptotic precision matrix as a function of $\boldsymbol{\theta} = (\mu, \sigma)$ is given by

$$\boldsymbol{H}_{\boldsymbol{\theta}}(\mu, \sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix},$$

$$\{h_\mu(\mu, \sigma)\}^{1/2} = \sigma^{-1},$$

$$\{h_{22}(\mu, \sigma)\}^{1/2} = \sqrt{2}\,\sigma^{-1}.$$

This implies that

$$\pi(\sigma \,|\, \mu) \propto \{h_{22}(\mu, \sigma)\}^{1/2} \propto \sigma^{-1},$$

so that, for example, $\Lambda_i = \{\sigma; e^{-i} \leq \sigma \leq e^i\}, i = 1, 2, \ldots$, provides a suitable sequence of subsets of $\Lambda = \Re^+$ not depending on μ, over which $\pi(\sigma \,|\, \mu)$ can be normalised and the corollary to Theorem 21 can be applied. It follows that

$$\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma \,|\, \mu) \propto 1 \times \sigma^{-1}$$

provides a reference prior relative to the ordered parametrisation (μ, σ). The corresponding reference posterior for μ, given $\boldsymbol{x}$, is

$$\begin{aligned} \pi(\mu \,|\, \boldsymbol{x}) &\propto \int p(\boldsymbol{x} \,|\, \mu, \sigma)\pi(\mu, \sigma)\, d\sigma \\ &\propto \pi(\mu) \int \prod_{i=1}^{n} \mathrm{N}(x_i \,|\, \mu, \sigma)\pi(\sigma \,|\, \mu)\, d\sigma \\ &\propto \int \sigma^{-n} \exp\left\{-\frac{n}{2\sigma^2}\left[(\overline{x} - \mu)^2 + s^2\right]\right\} \sigma^{-1}\, d\sigma \\ &\propto \int \lambda^{n/2-1} \exp\left\{-\frac{n\lambda}{2}\left[(\overline{x} - \mu)^2 + s^2\right]\right\} d\lambda \\ &\propto \left[s^2 + (\mu - \overline{x})^2\right]^{-n/2} \\ &= \mathrm{St}(\mu \,|\, \overline{x}, (n-1)s^{-2}, n-1), \end{aligned}$$

where $ns^2 = \Sigma(x_i - \overline{x})^2$, so that

$$t = \sqrt{n-1}\left(\frac{\overline{x} - \mu}{s}\right) = \sqrt{n}\,(\overline{x} - \mu)\Big/\sqrt{\frac{\sum(x_j - \overline{x})^2}{n-1}}$$

has a standard t distribution with $n - 1$ degrees of freedom.

If we now reverse the roles of μ and σ, so that the latter is now the parameter of interest and μ is the nuisance parameter, we obtain, writing $\boldsymbol{\phi} = (\sigma, \mu)$

$$\boldsymbol{H}_{\boldsymbol{\phi}}(\sigma, \mu) = \begin{pmatrix} 2\sigma^{-2} & 0 \\ 0 & \sigma^{-2} \end{pmatrix},$$

so that $\{h_\sigma(\sigma, \mu)\}^{1/2} = \sqrt{2}\sigma^{-1}$, $h_{22}(\sigma, \mu)\}^{1/2} = \sigma^{-1}$ and, by a similar analysis to the above,

$$\pi(\mu \,|\, \sigma) \propto \sigma^{-1}$$

so that, for example, $\Lambda_i = \{\mu; -i \leq \mu \leq i\}, i = 1, 2, \ldots$ provides a suitable sequence of subsets of $\Lambda = \Re$ not depending on σ, over which $\pi(\mu \,|\, \sigma)$ can be normalised and the corollary to Theorem 21 can be applied. It follows that

$$\pi(\mu, \sigma) = \pi(\sigma)\pi(\mu \,|\, \sigma) \propto 1 \times \sigma^{-1}$$

provides a reference prior relative to the ordered parametrisation (σ, μ). The corresponding reference posterior for σ, given $\boldsymbol{x}$, is

$$\begin{aligned}\pi(\sigma \,|\, \boldsymbol{x}) &\propto \int p(\boldsymbol{x} \,|\, \mu, \sigma)\ \pi(\mu, \sigma)\ d\mu \\ &\propto \pi(\sigma) \int \prod_{i=1}^{n} \mathrm{N}(x_i \,|\, \mu, \sigma)\ \pi(\mu \,|\, \sigma)\ d\mu,\end{aligned}$$

the right-hand side of which can be written in the form

$$\sigma^{-n} \exp\left\{-\frac{ns^2}{2\sigma^2}\right\} \int \sigma^{-1} \exp\left\{-\frac{n}{2\sigma^2}(\mu - \overline{x})^2\right\}\ d\mu.$$

Noting, by comparison with a $N(\mu \,|\, \overline{x}, n\lambda)$ density, that the integral is a constant, and changing the variable to $\lambda = \sigma^{-1}$, implies that

$$\begin{aligned}\pi(\lambda \,|\, \boldsymbol{x}) &\propto \lambda^{(n-1)/2-1} \exp\left\{\tfrac{1}{2} ns^2 \lambda\right\} \\ &= \mathrm{Ga}\left(\lambda \mid \tfrac{1}{2}(n-1),\ \tfrac{1}{2}ns^2\right),\end{aligned}$$

or, alternatively,

$$\begin{aligned}\pi(\lambda ns^2 \,|\, \boldsymbol{x}) &= \mathrm{Ga}\left(\lambda ns^2 \mid \tfrac{1}{2}(n-1),\ \tfrac{1}{2}\right) \\ &= \chi^2(\lambda ns^2 \,|\, n-1),\end{aligned}$$

so that ns^2/σ^2 has a chi-squared distribution with $n-1$ degrees of freedom.

□

One feature of the above example is that the reference prior did not, in fact, depend on which of the parameters was taken to be the parameter of interest. In the following example the form does change when the parameter of interest changes.

Example 15. ***Standardised normal mean***. We consider the same situation as that of Example 14, but we now take $\phi = \mu/\sigma$ to be the parameter of interest. If σ is taken as the nuisance parameter (by Theorem 19 the choice is irrelevant), $\boldsymbol{\psi} = (\phi, \sigma) = \boldsymbol{g}(\mu, \sigma)$ is clearly a one-to-one transformation, with

$$\boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) = \begin{pmatrix} \dfrac{\partial \mu}{\partial \phi} & \dfrac{\partial \mu}{\partial \sigma} \\ \dfrac{\partial \sigma}{\partial \phi} & \dfrac{\partial \sigma}{\partial \sigma} \end{pmatrix} = \begin{pmatrix} \sigma & \phi \\ 0 & 1 \end{pmatrix}$$

and using Corollary 1 to Theorem 9.

$$\boldsymbol{H}_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \boldsymbol{J}^t_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{g}^{-1}(\boldsymbol{\psi})) \boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) = \begin{pmatrix} 1 & \phi\sigma^{-1} \\ \phi\sigma^{-1} & \sigma^{-2}(2 + \phi^2) \end{pmatrix}.$$

Again, the sequence $\Lambda_i = \{\sigma; e^{-i} \leq \sigma \leq e^i\}, i = 1, 2, \ldots$, provides a reasonable basis for applying the corollary to Theorem 21. It is easily seen that

$$\begin{aligned}&\{h_\phi(\phi, \sigma)\}^{1/2} = \{h_{11(\phi,\sigma)} - h_{12(\phi,\sigma)} h_{22}^{-1}(\phi, \sigma) h_{21}(\phi, \sigma)\}^{1/2} = (1 + \tfrac{1}{2}\phi^2)^{-1/2}, \\ &\{h_{22}(\phi, \sigma)\}^{1/2} = (2 + \phi^2)^{1/2} \sigma^{-1},\end{aligned}$$

so that the reference prior relative to the ordered parametrisation (ϕ, σ) is given by

$$\pi(\phi, \sigma) \propto (1 + \tfrac{1}{2}\phi^2)^{-1/2}\sigma^{-1}.$$

In the (μ, σ) parametrisation this corresponds to

$$\pi(\mu, \sigma) \propto \left(1 + \frac{1}{2}\frac{\mu^2}{\sigma^2}\right)^{-1/2} \sigma^{-2},$$

which is clearly different from the form obtained in Example 14. Further discussion of this example will be provided in Example 22, in Chapter 4.

□

We conclude this subsection by considering a rather more involved example, where a natural choice of the required $\Lambda_i(\phi)$ subsequence *does* depend on ϕ. In this case, we use Theorem 21, since its corollary does not apply.

Example 16. *Product of normal means*. Consider the case where independent random samples $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ are to be taken, respectively, from $N(x \,|\, \alpha, 1)$ and $N(y \,|\, \beta, 1)$, $\alpha > 0$, $\beta > 0$, so that the complete parametric model is

$$p(\boldsymbol{x}, \boldsymbol{y} \,|\, \alpha, \beta) = \prod_{i=1}^{n} N(x_i \,|\, \alpha, 1) \prod_{j=1}^{m} N(y_j \,|\, \beta, 1),$$

for which, writing $\boldsymbol{\theta} = (\alpha, \beta)$ the Fisher information matrix is easily seen to be

$$\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \boldsymbol{H}(\alpha, \beta) = \begin{pmatrix} n & 0 \\ 0 & m \end{pmatrix}.$$

Suppose now that we make the one-to-one transformation $\boldsymbol{\psi} = (\phi, \lambda) = (\alpha\beta, \alpha/\beta) = \boldsymbol{g}(\alpha, \beta) = \boldsymbol{g}(\boldsymbol{\theta})$, so that $\phi = \alpha\beta$ is taken to be the parameter of interest and $\lambda = \alpha/\beta$ is taken to be the nuisance parameter. Such a parameter of interest arises, for example, when inference about the area of a rectangle is required from data consisting of measurements of its sides.

The Jacobian of the inverse transformation is given by

$$\boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) = \begin{pmatrix} \dfrac{\partial\alpha}{\partial\phi} & \dfrac{\partial\alpha}{\partial\lambda} \\ \dfrac{\partial\beta}{\partial\phi} & \dfrac{\partial\beta}{\partial\lambda} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \left(\dfrac{\lambda}{\phi}\right)^{1/2} & \left(\dfrac{\phi}{\lambda}\right)^{1/2} \\ \left(\dfrac{1}{\phi\lambda}\right)^{1/2} & -\dfrac{1}{\lambda}\left(\dfrac{\phi}{\lambda}\right)^{1/2} \end{pmatrix}$$

and hence, using Corollary 1 to Theorem 9

$$\boldsymbol{H}_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \boldsymbol{J}^t_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) \boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{g}^{-1}(\boldsymbol{\psi})) \boldsymbol{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{\psi}) = \frac{nm}{4\lambda^2} \begin{bmatrix} \dfrac{\lambda}{\phi}\left(\dfrac{\lambda^2}{m} + \dfrac{1}{n}\right) & \left(\dfrac{\lambda^2}{m} - \dfrac{1}{n}\right) \\ \left(\dfrac{\lambda^2}{m} - \dfrac{1}{n}\right) & \phi\left(\dfrac{\lambda}{m} + \dfrac{1}{n\lambda}\right) \end{bmatrix},$$

with $|\, \boldsymbol{H}_{\boldsymbol{\psi}}(\boldsymbol{\psi}) \,| = \dfrac{nm}{4\lambda^2}$, so that

$$\pi(\lambda \,|\, \phi) \propto \{h_{22}(\phi, \lambda)\}^{1/2} \propto \frac{(nm\phi)^{1/2}}{\lambda}\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{1/2}.$$

The question now arises as to what constitutes a "natural" sequence $\{\lambda_i(\phi)\}$, over which to define the normalised $\pi_i(\lambda \,|\, \phi)$ required by Definition 9. A natural increasing sequence of subsets of the original parameter space, $\Re^+ \times \Re^+$, for (α, β) would be the sets

$$S_i = \{(\alpha, \beta); \quad 0 < \alpha < i, \quad 0 < \beta < i\}\,, \quad i = 1, 2, \ldots,$$

which transform, in the space of $\lambda \in \Lambda$, into the sequence

$$\Lambda_i(\phi) = \left\{\lambda; \quad \frac{\phi}{i^2} < \lambda < \frac{i^2}{\phi}\right\}\,.$$

We note that unlike in the previous cases we have considered, this does depend on ϕ.

To complete the analysis, it can be shown, after some manipulation, that, for large i,

$$\pi_i(\lambda \,|\, \phi) = \frac{\sqrt{nm}}{i(\sqrt{m} + \sqrt{n})}\phi^{1/2}\lambda^{-1}\left(\frac{1}{m} + \frac{1}{n\lambda}\right)^{1/2}$$

and

$$\pi_i(\phi) = \frac{\sqrt{nm}}{i\,(\sqrt{m} + \sqrt{n})}\int_{\Lambda_i(\phi)}\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{1/2}\lambda^{-1}\log\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{-1/2} d\lambda,$$

which leads to a reference prior relative to the ordered parametrisation (ϕ, λ) given by

$$\pi(\phi, \lambda) \propto \phi^{1/2}\lambda^{-1}\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{1/2}$$

In the original parametrisation, this corresponds to

$$\pi(\alpha, \beta) \propto (n\alpha^2 + m\beta^2)^{1/2},$$

which depends on the sample sizes through the ratio m/n and reduces, in the case $n = m$, to $\pi(\alpha, \beta) \propto (\alpha^2 + \beta^2)^{1/2}$, a form originally proposed by Stein (1982) for this problem, who showed that it provides approximate agreement between Bayesian credible regions and classical confidence intervals for ϕ; see also Efron (1986). For a detailed discussion of this example, and of the consequences of choosing a different sequence $\Lambda_i(\phi)$, see Berger and Bernardo (1989). □

We note that the preceding example serves to illustrate the fact that, in structured models, reference priors may depend explicitly on features of their structure, as the ratio m/n in Example 16. There is, of course, nothing paradoxical in this, since the underlying notion of a reference analysis is a "minimally informative" prior *relative* to information provided by infinite replications of the experiment to be analyzed. In structured experiments, such information typically depends of their structure.

3.5. MULTIPARAMETER PROBLEMS

The approach to the nuisance parameter case considered above was based on the use of an ordered parametrisation whose first and second components were (ϕ, λ), referred to, respectively, as the *parameter of interest* and the nuisance parameter. The reference prior for the *ordered* parametrisation (ϕ, λ) was then constructed by conditioning to give the form $\pi(\lambda \,|\, \phi)\pi(\phi)$.

When the model parameter vector $\boldsymbol{\theta}$ has more than two components, this successive conditioning idea can obviously be extended by considering $\boldsymbol{\theta}$ as an ordered parametrisation, $(\theta_1, \ldots, \theta_m)$, say, and generating, by successive conditioning, a reference prior, *relative to this ordered parametrisation*, of the form

$$\pi(\boldsymbol{\theta}) = \pi(\theta_m \,|\, \theta_1, \ldots, \theta_{m-1}) \cdots \pi(\theta_2 \,|\, \theta_1)\pi(\theta_1).$$

We will limit the discussion here to *regular* models, for which the posterior distribution is asymptotically normal. In order to describe the algorithm for producing the successively conditioned reference prior form in this regular case, we shall first introduce some notation.

Assuming the parametric model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, to be such that the Fisher information matrix

$$\boldsymbol{H}(\boldsymbol{\theta}) = -E_{\boldsymbol{x} \,|\, \boldsymbol{\theta}} \left\{ \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \right\}$$

has full rank, we define $\boldsymbol{S}(\theta) = \boldsymbol{H}^{-1}(\theta)$, define the component vectors

$$\boldsymbol{\theta}^{[j]} = (\theta_1, \ldots, \theta_j), \quad \boldsymbol{\theta}_{[j]} = (\theta_{j+1}, \ldots, \theta_m),$$

and denote by $\boldsymbol{S}_j(\boldsymbol{\theta})$ the corresponding upper left $j \times j$ submatrix of $\boldsymbol{S}(\boldsymbol{\theta})$, and by $h_j(\theta)$ the lower right element of $\boldsymbol{S}_j^{-1}(\boldsymbol{\theta})$.

Finally, we assume that $\Theta = \Theta_1 \times \cdots \times \Theta_m$, with $\theta_i \in \Theta_i$, and, for $i = 1, 2, \ldots$, we denote by $\{\Theta_i^l\}$, $l = 1, 2, \ldots$, an increasing sequence of compact subsets of Θ_i, and define $\Theta_{[j]}^l = \Theta_{j+1}^l \times \cdots \times \Theta_m^l$.

Theorem 22. ***Ordered reference priors under asymptotic normality*.**
With the above notation, and under regularity conditions extending those of Theorem 21 in an obvious way, the reference prior $\pi(\boldsymbol{\theta})$, relative to the ordered parametrisation $(\theta_1, \ldots, \theta_m)$, is given by

$$\pi(\boldsymbol{\theta}) = \lim_{l \to \infty} \frac{\pi^l(\boldsymbol{\theta})}{\pi^l(\boldsymbol{\theta}^*)}, \quad \textit{for some } \boldsymbol{\theta}^* \in \Theta,$$

where $\pi^l(\boldsymbol{\theta})$ is defined by the following recursion:

(i) For $j = m$, and $\theta_m \in \Theta_m^l$,

$$\pi_m^l \left(\theta_{[m-1]} \,|\, \theta^{[m-1]} \right) = \pi_m^l \left(\theta_m \,|\, \theta_1, \ldots, \theta_{m-1} \right) = \frac{\{h_m(\boldsymbol{\theta})\}^{1/2}}{\int_{\Theta_m^l} \{h_m(\boldsymbol{\theta})\}^{1/2} \, d\theta_m}.$$

(ii) For $j = m-1, m-2, \ldots, 2$, and $\theta_j \in \Theta_j^l$,

$$\pi_j^l \left(\boldsymbol{\theta}_{[j-1]} \,|\, \boldsymbol{\theta}^{[j-1]} \right) = \pi_{j+1}^l \left(\boldsymbol{\theta}_{[j]} \,|\, \boldsymbol{\theta}^{[j]} \right) \frac{\exp\left\{ E_j^l \left[\log\{h_j(\boldsymbol{\theta})\}^{1/2} \right] \right\}}{\int_{\Theta_j^l} \exp\left\{ E_j^l \left[\log\{h_j(\boldsymbol{\theta})\}^{1/2} \right] \right\} d\theta_j},$$

where

$$E_j^l\left[\log\{h_j(\boldsymbol{\theta})\}^{1/2}\right] = \int_{\Theta^l_{[j]}} \log\{h_j(\boldsymbol{\theta})\}^{1/2}\, \pi^l_{j+1}\left(\boldsymbol{\theta}_{[j]} \mid \boldsymbol{\theta}^{[j]}\right) d\boldsymbol{\theta}_{[j]}.$$

(iii) For $j = 1$, $\boldsymbol{\theta}_{[0]} = \boldsymbol{\theta}$, with $\boldsymbol{\theta}^{[0]}$ vacuous, and

$$\pi^l(\boldsymbol{\theta}) = \pi^l_1\left(\boldsymbol{\theta}_{[0]}|\boldsymbol{\theta}^{[0]}\right).$$

Proof. This follows closely the development given in Theorem 21. For details see Berger and Bernardo (1992a, 1992b, 1992c).

◁

The derivation of the ordered reference prior is greatly simplified if the $\{h_j(\boldsymbol{\theta})\}$ terms in the above depend only on $\theta^{[j]}$: even greater simplification obtains if $\boldsymbol{H}(\boldsymbol{\theta})$ is block diagonal, particularly, if, for $j = 1, \ldots, m$, the jth term can be factored into a product of a function of θ_j and a function not depending on θ_j.

Corollary 1. *Factorisation*.
If $h_j(\boldsymbol{\theta})$ depends only on $\boldsymbol{\theta}^{[j]}$, $j = 1, \ldots, m$, then

$$\pi^l(\boldsymbol{\theta}) = \prod_{j=1}^{m} \frac{\{h_j(\boldsymbol{\theta})\}^{1/2}}{\int_{\Theta^l_j}\{h_j(\boldsymbol{\theta})\}^{1/2}\, d\theta_j}\,, \quad \boldsymbol{\theta} \in \Theta^l.$$

If $\boldsymbol{H}(\boldsymbol{\theta})$ is block diagonal (i.e., $\theta_1, \ldots, \theta_m$ are mutually orthogonal), with

$$\boldsymbol{H}(\boldsymbol{\theta}) = \begin{pmatrix} h_{11}(\boldsymbol{\theta}) & 0 & \cdots & 0 \\ 0 & h_{22}(\boldsymbol{\theta}) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & h_{mm}(\boldsymbol{\theta}) \end{pmatrix},$$

then $h_j(\boldsymbol{\theta}) = h_{jj}(\boldsymbol{\theta}), j = 1, \ldots, m$. Furthermore, if, in this latter case,

$$\{h_{jj}(\boldsymbol{\theta})\}^{1/2} = f_j(\theta_j)g_j(\boldsymbol{\theta}),$$

where $g_j(\boldsymbol{\theta})$ does not depend on θ_j, and if the Θ^l_j's do not depend on $\boldsymbol{\theta}$, then

$$\pi(\boldsymbol{\theta}) \propto \prod_{j=1}^{m} f_j(\theta_j).$$

Proof. The results follow from the recursion of Theorem 21.

◁

The question obviously arises as to the appropriate ordering to be adopted in any specific problem. At present, no formal theory exists to guide such a choice, but experience with a wide range of examples suggests that—at least for non-hierarchical models, where the parameters may have special forms of interrelationship—the best procedure is to order the components of θ on the basis of their inferential interest.

Example 17. ***Reference analysis for*** m ***normal means***. Let e_n be an experiment which consists in obtaining $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $n \geq 2$, a random sample from the multivariate normal model $N_m(\boldsymbol{x} \mid \boldsymbol{\mu}, \tau \boldsymbol{I}_m)$, $m \geq 1$, for which the Fisher information matrix is easily seen to be

$$\boldsymbol{H}(\boldsymbol{\mu}, \tau) = \begin{pmatrix} \tau \boldsymbol{I}_m & 0 \\ 0 & mn/(2\tau^2) \end{pmatrix}.$$

It follows from Theorem 30 that the reference prior relative to the natural parametrisation $(\mu_1, \ldots, \mu_m, \tau)$, is given by

$$\pi(\mu_1, \ldots, \mu_m, \tau) \propto \tau^{-1}.$$

Clearly, in this example the result does not, in fact, depend on the order in which the parametrisation is taken, since the parameters are all mutually orthogonal.

The reference prior $\pi(\mu_1, \ldots, \mu_m, \tau) \propto \tau^{-1}$ or $\pi(\mu_1, \ldots, \mu_m, \sigma) \propto \sigma^{-1}$ if we parametrise in terms of $\sigma = \tau^{-1/2}$, is thus the appropriate reference form if we are interested in any of the individual parameters. The reference posterior for any μ_j is easily shown to be the Student density

$$\pi(\mu_j \mid x_1, \ldots, x_n) = \text{St}\left(\mu_j \mid \overline{x}_j, (n-1)s^{-2}, m(n-1)\right)$$

$$n\overline{x}_j = \sum_{i=1}^{n} x_{ij}, \qquad nms^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ij} - \overline{x}_j)^2$$

which agrees with the standard argument according to which one degree of freedom should be lost by each of the unknown means.

□

Example 18. ***Multinomial model***. Let $\boldsymbol{x} = \{r_1, \ldots, r_m\}$ be an observation from a multinomial distribution, so that

$$p(r_1, \ldots, r_m \mid \theta_1, \ldots, \theta_m) = \frac{n!}{r_1! \cdots r_m!(n - \Sigma r_i)!} \theta_1^{r_1} \cdots \theta_m^{r_m} (1 - \Sigma\theta_i)^{n - \Sigma r_i},$$

from which the Fisher information matrix

$$\boldsymbol{H}(\theta_1, \ldots, \theta_m) = \frac{n}{1 - \Sigma\theta_i} \begin{bmatrix} \dfrac{1 + \theta_1 - \Sigma\theta_i}{\theta_1} & 1 & \ldots & 1 \\ 1 & \dfrac{1 + \theta_2 - \Sigma\theta_i}{\theta_2} & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & \dfrac{1 + \theta_m - \Sigma\theta_i}{\theta_m} \end{bmatrix}$$

is easily derived, with

$$|\boldsymbol{H}| = n^m \left[\left(1 - \sum_{i=1}^{m} \theta_1\right) \prod_{i=1}^{m} \theta_i \right]^{-1}.$$

In this case, the conditional reference priors derived using Theorem 22 turn out to be proper, and there is no need to consider subset sequences $\{\Theta_i^l\}$. In fact, noting that $H^{-1}(\theta_1,\ldots,\theta_m)$ is given by

$$\frac{1}{n}\begin{bmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \cdots & -\theta_1\theta_m \\ -\theta_1\theta_2 & \theta_2(1-\theta_2) & \cdots & -\theta_2\theta_m \\ \cdots & \cdots & \cdots & \cdots \\ -\theta_1\theta_m & -\theta_2\theta_m & \cdots & \theta_m(1-\theta_m) \end{bmatrix},$$

we see that the conditional asymptotic precisions used in Theorem 22 are easily identified, leading to

$$\pi(\theta_j\,|\,\theta_1,\ldots,\theta_{j-1}) \propto \left(\frac{1-\sum_{i=1}^{j-1}\theta_i}{\theta_j}\right)^{1/2}\left(\frac{1}{1-\sum_{i=1}^{j}\theta_i}\right)^{1/2}, \quad \theta_j \le 1-\sum_{i=1}^{j-1}\theta_i.$$

The required reference prior relative to the ordered parametrisation $(\theta_1,\ldots,\theta_m)$, say, is then given by

$$\begin{aligned}\pi(\theta_1,\ldots,\theta_m) &\propto \pi(\theta_1)\pi(\theta_2\,|\,\theta_1)\cdots\pi(\theta_m\,|\,\theta_1,\ldots,\theta_{m-1})\\ &\propto \theta_1^{-1/2}(1-\theta_1)^{-1/2}\theta_2^{-1/2}(1-\theta_1-\theta_2)^{-1/2}\cdots\theta_m^{-1/2}(1-\theta_1-\cdots-\theta_m)^{-1/2},\end{aligned}$$

and corresponding reference posterior for θ_1 is

$$\pi(\theta_1\,|\,r_1,\ldots,r_m) \propto \int p(r_1,\ldots,r_m\,|\,\theta_1,\ldots,\theta_m)\,\pi(\theta_1,\ldots,\theta_m)\,d\theta_2\ldots d\theta_m,$$

which is proportional to

$$\begin{aligned}\int \theta_1^{r_1-1/2}\cdots\theta_m^{r_m-1/2}&(1-\Sigma\theta_i)^{n-\Sigma r_i}\\ &\times(1-\theta_1)^{-1/2}(1-\theta_1-\theta_2)^{-1/2}\cdots(1-\theta_1-\cdots-\theta_m)^{-1/2}d\theta_2\cdots d\theta_m.\end{aligned}$$

After some algebra, this implies that

$$\pi(\theta_1\,|\,r_1,\ldots,r_m) = \text{Be}\left(\theta_1\,|\,r_1+\tfrac{1}{2}, n-r_1+\tfrac{1}{2}\right),$$

which, as one could expect, coincides with the reference posterior which would have been obtained had we initially collapsed the multinomial analysis to a binomial model and then carried out a reference analysis for the latter. Clearly, by symmetry considerations, the above analysis applies to any θ_i, $i=1,\ldots,m$, after appropriate changes in labelling and it is independent of the particular order in which the parameters are taken. For a detailed discussion of this example see Berger and Bernardo (1992a). Further comments on ordering of parameters are given in Chapter 4.

□

Example 19. ***Normal correlation coefficient***. Let $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ be a random sample from a bivariate normal distribution, $N_2(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\tau})$, where

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad \boldsymbol{\tau}^{-1} = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}.$$

Suppose that the correlation coefficient ρ is the parameter of interest, and consider the ordered parametrisation $\{\rho, \mu_1, \mu_2, \sigma_1, \sigma_2\}$. It is easily seen that

$$\boldsymbol{H}(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2) = (1-\rho^2)^{-1} \begin{bmatrix} \dfrac{1+\rho^2}{1-\rho^2} & 0 & 0 & \dfrac{-\rho}{\sigma_1} & \dfrac{-\rho}{\sigma_2} \\ 0 & \dfrac{1}{\sigma_1^2} & \dfrac{-\rho}{\sigma_1\sigma_2} & 0 & 0 \\ 0 & \dfrac{-\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2} & 0 & 0 \\ \dfrac{-\rho}{\sigma_1} & 0 & 0 & \dfrac{2-\rho^2}{\sigma_1^2} & \dfrac{-\rho^2}{\sigma_1\sigma_2} \\ \dfrac{-\rho}{\sigma_2} & 0 & 0 & \dfrac{-\rho^2}{\sigma_1\sigma_2} & \dfrac{2-\rho^2}{\sigma_2^2} \end{bmatrix},$$

so that

$$\boldsymbol{H}^{-1} = \begin{bmatrix} (1-\rho^2)^2 & 0 & 0 & \dfrac{\sigma_1}{2}\rho(1-\rho^2) & \dfrac{\sigma_2}{2}\rho(1-\rho^2) \\ 0 & \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 & 0 \\ 0 & \rho\sigma_1\sigma_2 & \sigma_2^2 & 0 & 0 \\ \dfrac{\sigma_1}{2}\rho(1-\rho^2) & 0 & 0 & \dfrac{\sigma_1^2}{2} & \rho^2\dfrac{\sigma_1\sigma_2}{2} \\ \dfrac{\sigma_2}{2}\rho(1-\rho^2) & 0 & 0 & \rho^2\dfrac{\sigma_1\sigma_2}{2} & \dfrac{\sigma_2^2}{2} \end{bmatrix}.$$

After some algebra it can be shown that this leads to the reference prior

$$\pi(\rho, \mu_1, \mu_2, \sigma_1, \sigma_2) \propto (1-\rho^2)^{-1}\sigma_1^{-1}\sigma_2^{-1},$$

whatever ordering of the nuisance parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ is taken. This agrees with Lindley's (1965, p. 219) analysis. Furthermore, as one could expect from Fisher's (1915) original analysis, the corresponding reference posterior distribution for ρ

$$\pi(\rho \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \propto \frac{(1-\rho^2)^{(n-3)/2}}{(1-\rho r)^{n-3/2}} F\left(\frac{1}{2}, \frac{1}{2}, n-\frac{1}{2}, \frac{1+\rho r}{2}\right),$$

(where F is the hypergeometric function; see *e.g.*, Abramowitz and Stegun, 1964, Ch. 15). Note that $\pi(\rho \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ only depends on the data through the sample correlation coefficient r, whose sampling distribution only depends on ρ. For a detailed analysis of this example, see Bayarri (1981); further discussion will be provided in Chapter 4.

□

See, also, Hills (1987), Ye and Berger (1991) and Berger and Bernardo (1992b) for derivations of the reference distributions for a variety of other interesting models.

Infinite discrete parameter spaces.

The infinite discrete case presents special problems, due to the non-existence of an asymptotic theory comparable to that of the continuous case. It is, however, often possible to obtain an approximate reference posterior by embedding the discrete parameter space within a continuous one.

Example 20. ***Infinite discrete case***. In the context of capture-recapture problems, suppose it is of interest to make inferences about an integer $\theta \in \{1, 2, \ldots\}$ on the basis of a random sample $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ from

$$p(x|\theta) = \frac{\theta(\theta+1)}{(x+\theta)^2}\,, \quad 0 \le x \le 1$$

For several plausible "diffuse looking" prior distributions for θ one finds that the corresponding posterior virtually ignores the data. Intuitively, this has to be interpreted as suggesting that such priors actually contain a large amount of information about θ compared with that provided by the data. A more careful approach to providing a "non-informative" prior is clearly required. One possibility would be to embed the discrete space $\{1, 2, \ldots\}$ in the continuous space $]0, \infty[$ since, for each $\theta > 0$, $p(x|\theta)$ is still a probability density for x. Then, using Theorem 16, the appropriate reference prior is

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

and it is easily verified that this prior leads to a posterior in which the data are no longer overwhelmed. If the physical conditions of the problem require the use of discrete θ values, one could always use, for example,

$$p(\theta = 1 \,|\, \boldsymbol{z}) = \int_0^{3/2} \pi(\theta \,|\, \boldsymbol{z}) d\theta, \qquad p(\theta = j \,|\, \boldsymbol{z}) = \int_{j-1/2}^{j+1/2} \pi(\theta|\boldsymbol{z}) d\theta, \quad j > 1$$

as an approximate discrete reference posterior.

□

Prediction and Hierarchical Models.

Two classes of problems that are not covered by the methods so far discussed are hierarchical models and prediction problems. The difficulty with these problems is that there are unknowns (typically the unknowns of interest) that have specified distributions. For instance, if one wants to predict y based on $\boldsymbol{z}$ when $(y, \boldsymbol{z})$ has density $p(y, \boldsymbol{z} \,|\, \boldsymbol{\theta})$, the unknown of interest is y, but its distribution is conditionally specified. One needs a reference prior for $\boldsymbol{\theta}$, not y. Likewise, in a hierarchical model with, say, $\mu_1, \mu_2, \ldots, \mu_p$ being $\mathrm{N}(\mu_i \,|\, \mu_0, \lambda)$, the μ_i's may be the parameters of interest but a prior is only needed for the hyperparameters μ_0 and λ.

The obvious way to approach such problems is to integrate out the variables with conditionally known distributions (y in the predictive problem and the $\{\mu_i\}$ in the hierarchical model), and find the reference prior for the remaining parameters based on this marginal model. The difficulty that arises is how to then identify parameters of interest and nuisance parameters to construct the ordering necessary for applying the reference prior method, the real parameters of interest having been integrated out.

In future work, we propose to deal with this difficulty by defining the parameter of interest in the reduced model to be the conditional mean of the original parameter of interest. Thus, in the prediction problem, $E[y|\boldsymbol{\theta}]$ (which will be either $\boldsymbol{\theta}$ or some transformation thereof) will be the parameter of interest, and in the hierarchical model $E[\mu_i \mid \mu_0, \lambda] = \mu_0$ will be defined to be the parameter of interest. This technique has so far worked well in the examples to which it has been applied, but further study is clearly needed.

See Aitchison and Dunsmore (1975) and Geisser (1993) for some examples of non-subjective predictive posterior distributions.

4. Discussion and Further References

In this monograph, we have described the motivation, definition, and derivation of reference posterior distributions, we have illustrated the theory with a number of examples, and we have demonstrated some of the properties which may be used to substantiate the claim that they constitute the more promising available method to derive *non-subjective* prior distributions. However, the definition and possible uses of non-subjective priors, which under this and many other labels (such as "conventional, "default", "formal", "neutral", "flat" or "noninformative"), are intended to provide Bayesian solutions which do not require to assess a subjective prior, have always been a rather polemic issue among statisticians. In this final chapter, we provide an overview of some of the main directions followed in this search for Bayesian objectivity, we summarize some of the elements of the discussion, and we provide signposts for those interested in pursuing the subject at a deeper level.

4.1. HISTORICAL DEVELOPMENT

In the early works by Bayes (1763) and Laplace (1814/1952), the definition of a non-informative prior is based on what has now become known as the principle of *insufficient reason*, or the Bayes-Laplace postulate. According to this principle, in the absence of evidence to the contrary, all possibilities should have the same initial probability. This is closely related to the so-called Laplace-Bertrand paradox; see Jaynes (1971) for an interesting Bayesian resolution.

In particular, if an unknown quantity, ϕ, say, can only take a finite number of values, M, say, the non-informative prior suggested by the principle is the discrete uniform distribution $p(\phi) = \{1/M, \ldots, 1/M\}$. This may, at first sight, seem intuitively reasonable, but Example 13 showed that even in simple, finite, discrete cases care can be required in appropriately defining the unknown *quantity of interest*. Moreover, in countably infinite discrete cases, the uniform (now *improper*) prior is known to produce unappealing results. Jeffreys (1939/1961, p. 238) suggested, for the case of the integers, the prior

$$\pi(n) \propto n^{-1}, \quad n = 1, 2, \ldots .$$

Far more recently, Rissanen (1983) used a coding theory argument to motivate the prior

$$\pi(n) \propto \frac{1}{n} \times \frac{1}{\log n} \times \frac{1}{\log \log n} \times \ldots, \quad n = 1, 2, \ldots .$$

However, as indicated in Example 20, embedding the discrete problem within a continuous framework and subsequently discretising the resulting reference prior for the continuous case may produce better results.

If the space, Φ, of ϕ values is a continuum (say, the real line) the principle of insufficient reason has been interpreted as requiring a uniform distribution over Φ. However, a uniform distribution for ϕ implies a non-uniform distribution for any non-linear monotone transformation of ϕ and thus the Bayes-Laplace postulate is inconsistent in the sense that, intuitively, "ignorance about ϕ" should surely imply "equal ignorance" about a one-to-one transformation of ϕ. Specifically, if some procedure yields $p(\phi)$ as a non-informative prior for ϕ and the same procedure yields $p(\zeta)$ as a non-informative prior for a one-to-one transformation $\zeta = \zeta(\phi)$ of ϕ, consistency would seem to demand that $p(\zeta)d\zeta = p(\phi)d\phi$; thus, a procedure for obtaining the "ignorance" prior should presumably be invariant under one-to-one reparametrisation.

Based on these invariance considerations, Jeffreys (1946) proposed as a non-informative prior, with respect to an experiment $e = \{X, \phi, p(x \mid \phi)\}$, involving a parametric model which depends on a single parameter ϕ, the (often improper) density

$$\pi(\phi) \propto h(\phi)^{1/2},$$

where

$$h(\phi) = -\int_X p(x \mid \phi) \frac{\partial^2}{\partial \phi^2} \log p(x \mid \phi)\; dx\ .$$

In effect, Jeffreys noted that the logarithmic divergence locally behaves like the square of a distance, determined by a Riemannian metric, whose natural length element is $h(\phi)^{1/2}$, and that natural length elements of Riemannian metrics are invariant to reparametrisation. In an illuminating paper, Kass (1989) elaborated on this *geometrical* interpretation by arguing that, more generally, natural volume elements generate "uniform" measures on manifolds, in the sense that equal mass is assigned to regions of equal volume, the essential property that makes Lebesgue measure intuitively appealing.

In his work, Jeffreys explored the implications of such a non-informative prior for a large number of inference problems. He found that his *rule* (by definition restricted to a continuous parameter) works well in the one-dimensional case, but can lead to unappealing results (Jeffreys, 1939/1961, p. 182) when one tries to extend it to multiparameter situations.

The procedure proposed by Jeffreys' preferred rule was rather *ad hoc*, in that there are many other procedures (some of which he described) which exhibit the required type of invariance. His intuition as to what is required, however, was rather good. Jeffreys' solution for the one-dimensional continuous case has been widely adopted, and a number of alternative justifications of the procedure have been provided.

Perks (1947) used an argument based on the asymptotic size of confidence regions to propose a non-informative prior of the form

$$\pi(\phi) \propto s(\phi)^{-1}$$

where $s(\phi)$ is the asymptotic standard deviation of the maximum likelihood estimate of ϕ. Under regularity conditions which imply asymptotic normality, this turns out to be equivalent to Jeffreys' rule.

Lindley (1961b) argued that, in practice, one can always replace a continuous range of ϕ by discrete values over a grid whose mesh size, $\delta(\phi)$, say, describes the precision of the measuring process, and that a possible operational interpretation of "ignorance" is a probability distribution which assigns equal probability to all points of this grid. In the continuous case, this implies a prior proportional to $\delta(\phi)^{-1}$. To determine $\delta(\phi)$ in the context of an experiment $e = \{X, \phi, p(x \mid \phi)\}$, Lindley showed that if the quantity can only take the values ϕ or $\phi + \delta(\phi)$, the amount of information that e may be expected to provide about ϕ, if $p(\phi) = p(\phi + \delta(\phi)) = \frac{1}{2}$, is $2\delta^2(\phi)h(\phi)$. This expected information will be independent of ϕ if $\delta(\phi) \propto h(\phi)^{-1/2}$, thus defining an appropriate mesh; arguing as before, this suggests Jeffreys' prior $\pi(\phi) \propto h(\theta)^{1/2}$. Akaike (1978a) used a related argument to justify Jeffreys' prior as "locally impartial".

Welch and Peers (1963) and Welch (1965) discussed conditions under which there is formal mathematical equivalence between one-dimensional Bayesian credible regions and corresponding frequentist confidence intervals. They showed that, under suitable regularity assumptions, one-sided intervals asymptotically coincide if the prior used for the Bayesian analysis is Jeffreys' prior. Peers (1965) later showed that the argument does not extend to several dimensions. Hartigan (1966b) and Peers (1968) discuss two-sided intervals. Tibshirani (1989), Mukerjee and Dey (1993) and Nicolau (1993) extend the analysis to the case where there are nuisance parameters.

Hartigan (1965) reported that the prior density which minimises the bias of the estimator d of ϕ associated with the loss function $l(d, \phi)$ is

$$\pi(\phi) = h(\phi) \left[\frac{\partial^2}{\partial d^2} l(d, \phi) \right]^{-1/2} \bigg|_{d=\phi} .$$

If, in particular, one uses the discrepancy measure

$$l(d, \phi) = \int p(x \mid \phi) \log \frac{p(x \mid \phi)}{p(x \mid d)} \, dx$$

as a natural loss function, this implies that $\pi(\phi) = h(\phi)^{1/2}$, which is, again, Jeffreys' prior.

Good (1969) derived Jeffreys' prior as the "least favourable" initial distribution with respect to a logarithmic scoring rule, in the sense that it minimises the expected score from reporting the true distribution. Since the logarithmic score is proper, and hence is maximised by reporting the true distribution, Jeffreys' prior may technically be described, under suitable regularity conditions, as a minimax solution to the problem of scientific reporting when the utility function is the logarithmic score function. Kashyap (1971) provided a similar, more detailed argument; an axiom system is used to justify the use of an information measure as a payoff function and Jeffreys' prior is shown to be a minimax solution in a —two person— zero sum game, where the statistician chooses the "non-informative" prior and nature chooses the "true" prior.

Hartigan (1971, 1983, Chapter 5) defines a similarity measure for events E, F to be $P(E \cap F)/P(E)P(F)$ and shows that Jeffreys' prior ensures, asymptotically, constant similarly for current and future observations.

Following Jeffreys (1955), Box and Tiao (1973, Section 1.3) argued for selecting a prior by convention to be used as a *standard of reference*. They suggested that the principle of insufficient reason may be sensible in location problems, and proposed as a conventional prior $\pi(\phi)$ for a model parameter ϕ that $\pi(\phi)$ which implies a uniform prior

$$\pi(\zeta) = \pi(\phi) \left| \frac{\partial \zeta}{d\phi} \right|^{-1} \propto c$$

for a function $\zeta = \zeta(\phi)$ such that $p(x \mid \zeta)$ is, at least approximately, a location parameter family; that is, such that, for some functions g and f,

$$p(x \mid \phi) \sim g \left[\zeta(\phi) - f(x) \right] .$$

Using standard asymptotic theory, they showed that, under suitable regularity conditions and for large samples, this will happen if

$$\zeta(\phi) = \int^{\phi} h(\phi)^{1/2} d\phi \, ,$$

i.e., if the non-informative prior is Jeffreys' prior. For a recent reconsideration and elaboration of these ideas, see Kass (1990), who extends the analysis by conditioning on an ancillary statistic.

Unfortunately, although many of the arguments summarised above generalise to the multiparameter continuous case, leading to the so-called multivariate Jeffreys' rule

$$\pi(\boldsymbol{\theta}) \propto |\,\boldsymbol{H}(\boldsymbol{\theta})\,|^{1/2},$$

where

$$[\boldsymbol{H}(\boldsymbol{\theta})]_{ij} = -\int p(x\,|\,\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(x\,|\,\boldsymbol{\theta})\,dx$$

is *Fisher's information matrix*, the results thus obtained typically have intuitively unappealing implications. An example of this, pointed out by Jeffreys himself (Jeffreys, 1939/1961 p. 182) is provided by the simple location-scale problem —described below for the normal case— where the multivariate rule leads to the prior $\pi(\theta, \sigma) \propto \sigma^{-2}$, where θ is the location and σ the scale parameter, rather than to the commonly accepted $\pi(\theta, \sigma) \propto \sigma^{-1}$ (which is also the reference prior, when either θ or σ are the quantities of interest). See, also, Stein (1962).

Example 21. ***Univariate normal model***. Let $\{x_1, \ldots, x_n\}$ be a random sample from $N(x\,|\,\mu, \lambda)$, and consider $\sigma = \lambda^{-1/2}$, the (unknown) standard deviation. In the case of known mean, $\mu = 0$, say, the appropriate (univariate) Jeffreys' prior is $\pi(\sigma) \propto \sigma^{-1}$ and the posterior distribution of σ would be such that $[\Sigma_{i=1}^n x_i^2]/\sigma^2$ is χ_n^2. In the case of unknown mean, if we used the multivariate Jeffreys' prior $\pi(\mu, \sigma) \propto \sigma^{-2}$ the posterior distribution of σ would be such that $[\Sigma_{i=1}^n (x_i - \overline{x})^2]/\sigma^2$ is, again, χ_n^2. This is widely recognised as unacceptable, in that one does not lose any degrees of freedom even though one has lost the knowledge that $\mu = 0$, and conflicts with the use of the widely adopted reference prior $\pi(\mu, \sigma) = \sigma^{-1}$, which implies that $[\Sigma_{i=1}^n (x_i - \overline{x})^2]/\sigma^2$ is χ_{n-1}^2.

□

The kind of problem exemplified above led Jeffreys to the *ad hoc* recommendation, widely adopted in the literature, of independent a priori treatment of location and scale parameters, applying his rule separately to each of the two subgroups of parameters, and then multiplying the resulting forms together to arrive at the overall prior specification. For an illustration of this, see Geisser and Cornfield (1963): for an elaboration of the idea, see Zellner (1986a).

At this point, one may wonder just what has become of the intuition motivating the arguments outlined above. Unfortunately, although the implied information limits are mathematically well-defined in one dimension, in higher dimensions the forms obtained may depend on the path followed to obtain the limit. Similar problems arise with other intuitively appealing desiderata. For example, the Box and Tiao suggestion of a uniform prior following transformation to a parametrisation ensuring data translation generalises, in the multiparameter setting, to the requirement of uniformity following a transformation which ensures that credible regions are of the same size. The problem, of course, is that, in several dimensions, such regions can be of the same size but very different in form.

Jeffreys' original requirement of invariance under reparametrisation remains perhaps the most intuitively convincing. If this is conceded, it follows that, whatever their apparent motivating intuition, approaches which do not have this property should be regarded as unsatisfactory. Such approaches include the use of limiting forms of conjugate priors, as in Haldane (1948), Novick and Hall (1965), Novick (1969), DeGroot (1970, Chapter 10) and Piccinato (1973, 1977), a predictivistic version of the principle of insufficient reason, Geisser (1984), and different forms of information-theoretical arguments, such as those put forward by Zellner (1977, 1991), Geisser (1979) and Torgesen (1981).

Maximising the expected information (as opposed to maximising the expected *missing* information) gives invariant, but unappealing results, producing priors that can have finite support (Berger *et al.*, 1989). Other information-based suggestions are those of Eaton (1982), Spall and Hill (1990) and Rodríguez (1991).

Partially satisfactory results have nevertheless been obtained in multiparameter problems where the parameter space can be considered as a group of transformations of the sample space. Invariance considerations within such a group suggest the use of *relatively invariant* (Hartigan, 1964) priors like the Haar measures. This idea was pioneered by Barnard (1952). Stone (1965) recognised that, in an appropriate sense, it should be possible to approximate the results obtained using a non-informative prior by those obtained using a convenient sequence of proper priors. He went on to show that, if a group structure is present, the corresponding *right* Haar measure is the only prior for which such a desirable convergence is obtained. It is reassuring that, in those one-dimensional problems for which a group of transformations does exist, the right Haar measures coincides with the relevant Jeffreys' prior. For some undesirable consequences of the *left* Haar measure see Bernardo (1978b). Further developments involving Haar measures are provided by Zidek (1969), Villegas (1969, 1971, 1977a, 1977b, 1981), Stone (1970), Florens (1978, 1982), Chang and Villegas (1986) and Chang and Eaves (1990). Dawid (1983b) provides an excellent review of work up to the early 1980's. However, a large group of interesting models do not have any group structure, so that these arguments cannot produce general solutions.

Even when the parameter space may be considered as a group of transformations there is no definitive answer. In such situations, the right Haar measures are the obvious choices and yet even these are open to criticism.

Example 22. ***Standardised mean***. Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $N(x \mid \mu, \lambda)$. The standard prior recommended by group invariance arguments is $\pi(\mu, \sigma) = \sigma^{-1}$ where $\lambda = \sigma^{-2}$. Although this gives adequate results if one wants to make inferences about either μ or σ, it is quite unsatisfactory if inferences about the standardised mean $\phi = \mu/\sigma$ are required. Stone and Dawid (1972) show that the posterior distribution of ϕ obtained from such a prior depends on the data through the statistic

$$t = \frac{\sum_{i=1}^n x_i}{(\sum_{i=1}^n x_i^2)^{1/2}} \,,$$

whose sampling distribution,

$$\begin{aligned} p(t \mid \mu, \sigma) &= p(t \mid \phi) \\ &= e^{-n\phi^2/2} \left\{1 - \frac{t^2}{n}\right\}^{(n-3)/2} \int_0^\infty \omega^{n-1} \exp\left\{-\frac{\omega^2}{2} + t\phi\omega\right\} d\omega, \end{aligned}$$

only depends on ϕ. One would, therefore, expect to be able to "match" the original inferences about ϕ by the use of $p(t \mid \phi)$ together with some appropriate prior for ϕ. However, no such prior exists.

On the other hand, the reference prior relative to the ordered partition (ϕ, σ) is (see Example 15)

$$\pi(\phi, \sigma) = (1 + \tfrac{1}{2}\phi^2)^{-1/2}\sigma^{-1}$$

and the corresponding posterior distribution for ϕ is

$$\pi(\phi \,|\, \boldsymbol{x}) \propto (1 + \tfrac{1}{2}\phi^2)^{-1/2} \left[e^{-n\phi^2/2} \int_0^\infty \omega^{n-1} \exp\{ -\frac{\omega^2}{2} + \lambda\phi\omega \} d\omega \right] \cdot$$

We observe that the factor in square brackets is proportional to $p(t \,|\, \phi)$ and thus the inconsistency disappears.

□

This type of *marginalisation paradox*, further explored by Dawid, Stone and Zidek (1973), appears in a large number of multivariate problems and makes it difficult to believe that, for any given model, a *single* prior may be usefully regarded as "universally" non-informative. Jaynes (1980) disagrees; Dawid *et al.* (1980) contest his argument.

An acceptable general theory for non-informative priors should be able to provide consistent answers to the same inference problem whenever this is posed in different, but equivalent forms. Although this idea has failed to produce a constructive procedure for deriving priors, it may be used to discard those methods which fail to satisfy this rather intuitive requirement.

Example 23. ***Correlation coefficient***. Let $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a random sample from a bivariate normal distribution, and suppose that inferences about the correlation coefficient ρ are required. It may be shown that if the prior is of the form

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \pi(\rho)\sigma_1^{-a}\sigma_2^{-a},$$

which includes all proposed "non-informative" priors for this model that we are aware of, then the posterior distribution of ρ is given by

$$\begin{aligned}\pi\,(\rho \,|\, \boldsymbol{x}, \boldsymbol{y}) &= \pi(\rho \,|\, r) \\ &= \frac{\pi(\rho)(1-\rho^2)^{(n+2a-3)/2}}{(1-\rho r)^{n+a-(5/2)}}\, F\left(\tfrac{1}{2}, \tfrac{1}{2}, n + a - \tfrac{3}{2}, \frac{1+\rho r}{2} \right) ,\end{aligned}$$

where

$$r = \frac{\sum_i x_i y_i - n\overline{x}\,\overline{y}}{[\sum_i (x_i - \overline{x})^2]^{1/2} [\sum_i (y_i - \overline{y})^2]^{1/2}}$$

is the sample correlation coefficient, and F is the hypergeometric function. This posterior distribution only depends on the data through the sample correlation coefficient r; thus, with this form of prior, r is sufficient. On the other hand, the sampling distribution of r is

$$\begin{aligned} p(r \,|\, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &= p(r \,|\, \rho) \\ &= \frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{(1-\rho r)^{n-3/2}} F\left(\tfrac{1}{2}, \tfrac{1}{2}, n - \tfrac{1}{2}, \frac{1+\rho r}{2} \right). \end{aligned}$$

Moreover, using the transformations $\delta = \tanh^{-1}\rho$ and $t = \tanh^{-1} r$, Jeffreys' prior for this univariate model is found to be $\pi(\rho) \propto (1-\rho^2)^{-1}$ (see Lindley, 1965, pp. 215–219).

Hence one would expect to be able to match, using this reduced model, the posterior distribution $\pi(\rho \,|\, r)$ given previously, so that

$$\pi(\rho \,|\, r) \propto p(r \,|\, \rho)(1-\rho^2)^{-1}.$$

Comparison between $\pi(\rho \,|\, r)$ and $p(r \,|\, \rho)$ shows that this is possible if and only if $a = 1$, and $\pi(\rho) = (1-\rho^2)^{-1}$. Hence, to avoid inconsistency the joint reference prior must be of the form

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = (1-\rho^2)^{-1}\sigma_1^{-1}\sigma_2^{-1},$$

which is precisely (see Example 19) the reference prior relative to any ordered parameterisation which begins by ρ, such as $\{\rho, \mu_1, \mu_2, \sigma_1, \sigma_2\}$.

However, it is easily checked that Jeffreys' multivariate prior is

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = (1-\rho^2)^{-3/2}\sigma_1^{-2}\sigma_2^{-2}$$

and that the "two-step" Jeffreys' multivariate prior which separates the location and scale parameters is

$$\pi(\mu, \mu_2)\pi(\sigma_1, \sigma_2, \rho) = (1-\rho^2)^{-3/2}\sigma_1^{-1}\sigma_2^{-1}.$$

For further detailed discussion of this example, see Bayarri (1981).

□

Once again, this example suggests that different non-informative priors may be appropriate *depending on the particular function of interest* or, more generally, on the ordered parameterisation.

Although marginalisation paradoxes disappear when one uses proper priors, to use proper approximations to non-informative priors as an approximate description of "ignorance" does not solve the problem either.

Example 24. ***Stein's paradox***. Let $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a random sample from a multivariate normal distribution $N_k(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{I}_k\}$. Let $\overline{x}_i$ be the mean of the n observations from coordinate i and let $t = \sum_i \overline{x}_i^2$. The universally recommended "non-informative" prior for this model is $\pi(\mu_1, \ldots, \mu_k) = 1$, which may be approximated by the proper density

$$\pi(\mu_1, \ldots, \mu_k) = \prod_{i=1}^{m} N(\mu_i \,|\, 0, \lambda),$$

where λ is very small. However, if inferences about $\phi = \sum_i \mu_i^2$ are desired, the use of this prior overwhelms, for large k, what the data have to say about ϕ. Indeed, with such a prior the posterior distribution of $n\phi$ is a non-central χ^2 distribution with k degrees of freedom and non-centrality parameter nt, so that

$$E[\phi \,|\, \boldsymbol{x}] = t + \frac{k}{n}, \quad V[\phi \,|\, x] = \frac{2}{n}\left[2t + \frac{k}{n}\right],$$

while the sampling distribution of nt is a non-central χ^2 distribution $\chi^2(nt \,|\, k, n\phi)$, with k degrees of freedom and parameter $n\phi$, so that $E[t \,|\, \phi] = \phi + k/n$. Thus, with, say, $k = 100$, $n = 1$ and $t = 200$, we have $E[\phi \,|\, \boldsymbol{x}] \approx 300$, $V[\phi \,|\, x] \approx 32^2$, whereas the unbiased estimator based on the sampling distribution gives $\hat{\phi} = t - k \approx 100$.

However, the asymptotic posterior distribution of ϕ is $N(\phi \,|\, \hat{\phi}, (4\hat{\phi})^{-1})$ and hence, by Theorem 16, the reference posterior for ϕ relative to $p(t \,|\, \phi)$ is

$$\pi(\phi \,|\, \boldsymbol{x}) \propto \pi(\phi)p(t \,|\, \phi) \propto \phi^{-1/2}\chi^2(nt \,|\, k, n\phi)$$

whose mode is close to $\hat{\phi}$. It may be shown that this is also the posterior distribution of ϕ derived from the reference prior relative to the ordered partition $\{\phi, \omega_1, \ldots, \omega_{k-1}\}$, obtained by reparametrising to polar coordinates in the full model. For further details, see Stein (1959), Efron (1973), Bernardo (1979b) and Ferrándiz (1982).

□

Naïve use of apparently "non-informative" prior distributions can lead to posterior distributions whose corresponding credible regions have untenable coverage probabilities, in the sense that, for some region C, the corresponding posterior probabilities $P(C \mid \boldsymbol{z})$ may be completely different from the conditional values $P(C \mid \theta)$ for almost all θ values.

Such a phenomenon is often referred to as *strong inconsistency* (see, for example, Stone, 1976). However, by carefully distinguishing between parameters of interest and nuisance parameters, reference analysis avoids this type of inconsistency. An illuminating example is provided by the reanalysis by Bernardo (1979b, reply to the discussion) of Stone's (1976) *Flatland* example.

Jaynes (1968) introduced a more general formulation of the problem. He allowed for the existence of a certain amount of initial "objective" information and then tried to determine a prior which reflected this initial information, but nothing else (see, also, Csiszár, 1985). Jaynes considered the entropy of a distribution to be the appropriate measure of uncertainty subject to any "objective" information one might have. If no such information exists and ϕ can only take a finite number of values, Jaynes' *maximum entropy* solution reduces to the Bayes-Laplace postulate. His arguments are quite convincing in the finite case; however, if ϕ is continuous, the non-invariant entropy functional, $H\{p(\phi)\} = -\int p(\phi) \log p(\phi) d\phi$, no longer has a sensible interpretation in terms of uncertainty. Jaynes' solution is to introduce a "reference" density $\pi(\phi)$ in order to define an "invariantised" entropy,

$$-\int p(\phi) \log \frac{p(\phi)}{\pi(\phi)} d\phi,$$

and to use the prior which maximises this expression, subject, again, to any initial "objective" information one might have. Unfortunately, $\pi(\phi)$ must itself be a representation of ignorance about ϕ so that no progress has been made. If a convenient group of transformations is present, Jaynes suggests invariance arguments to select the reference density. However, no general procedure is proposed.

Context-specific "non-informative" Bayesian analyses have been produced for specific classes of problems, with no attempt to provide a general theory. These include dynamic models (Pole and West, 1989) and finite population survey sampling (Meeden and Vardeman, 1991).

The quest for non-informative priors could be summarised as follows.

(i) In the finite case, Jaynes' principle of maximising the entropy is convincing, but cannot be extended to the continuous case.

(ii) In one-dimensional continuous regular problems, Jeffreys' prior is appropriate.

(iii) The infinite discrete case can often be handled by suitably embedding the problem within a continuous framework.

(iv) In continuous multiparameter situations there is no hope for a single, unique, "non-informative prior", appropriate for all the inference problems within a given model. To avoid having the prior dominating the posterior for *some* function ϕ of interest, the prior has to depend not only on the model but also on the parameter of interest or, more generally, on some notion of the order of importance of the parameters.

The reference prior theory introduced in Bernardo (1979b) avoids most of the problems encountered with other proposals. It reduces to Jaynes' form in the finite case and to Jeffreys' form in one-dimensional regular continuous problems, avoiding marginalisation paradoxes by insisting that the reference prior be tailored to the parameter of interest. However, subsequent work by Berger and Bernardo (1989) has shown that the heuristic arguments in Bernardo (1979b) required more precise definitions in complicated situations. Moreover, Berger and

Bernardo (1992a, 1992b, 1992c) showed that the partition into parameters of interest and nuisance parameter may not go far enough and that reference priors should be viewed relative to a given ordering—or, more generally, a given ordered grouping—of the parameters. This is the approach that has been described in detail in Chapter 3.

4.2. INTERPRETATION OF NON-SUBJECTIVE PRIORS

A major criticism to the use of non-subjective priors comes from subjectivist Bayesians, who argue that the prior should be an honest expression of the analyst's prior knowledge and not a function of the model, specially if this involves integration over the sample space and hence may violate the likelihood principle.

> . . . why should one's knowledge, or ignorance, of a quantity depend on the experiment being used to determine it? Lindley (1972, p. 71).

In many situations, we would accept this argument. However, as we argued earlier, priors which reflect knowledge of the experiment can sometimes be genuinely appropriate in Bayesian inference, as mentioned the discussion of stopping rules in Section 1.4. Moreover, priors intended to serve as *reference* points with respect to specific experimental setups should naturally depend on the experiment to be analysed

> In general we feel that it is sensible to choose a non-informative prior which expresses ignorance *relative* to information which can be supplied by a particular experiment. If the experiment is changed, then the expression of relative ignorance can be expected to change correspondingly. (Box and Tiao, 1973, p. 46).

Posteriors obtained from actual prior opinions could then be compared with those derived from a reference analysis in order to assess the relative importance of the initial opinions on the final inference. Indeed, from a *foundational* viewpoint, the derivation of a reference posterior should be seen as part of a healthy *sensitivity analysis*, where it is desired to analyze the changes in the posterior of interest induced by changes in the prior: a reference posterior is just an answer to a *what if* question, namely what could be said about the quantity of interest given the data, if one's prior knowledge were dominated by the data. If the experiment is changed the reference prior may be expected to change correspondingly; if subjective prior information is specified, the corresponding posterior could be compared with the reference posterior in order to assess the relative importance on the initial opinions in the final inference. Moreover, from a *pragmatic* point of view, it must be stressed that in the Bayesian analysis of the complex multiparameter models which are now systematically used as a consequence of the availability of numerical MCMC methods (models typically intractable from a frequentist perspective), there is little hope for a detailed assessment of a huge personal multivariate prior; the naïve use of some tractable "noninformative" prior may then hide important unwarranted assumptions which may easily dominate the analysis (see *e.g.*, Casella, 1996, and references therein). Careful, responsible choice of a non-subjective prior is then possibly the best available alternative.

It should also be mentioned here that some Bayesian statisticians would follow Jeffreys (1939/1961) or Jaynes (1996) in a radical non-subjective view: they would claim that subjective priors are useless for scientific inference and so, non-subjective priors are necessary because there is nothing else to do.

4.3. IMPROPER PRIORS

Reference priors may occasionally *proper* probability even when the parameter space Φ is not bounded. (see *e.g.,* Bernardo and Ramón, 1998, Sec. 3 for an interesting example). However, reference priors associated to models with unbounded parameter spaces are typically *improper* in that, in most cases, if Φ is not compact, then $\int_\Phi \pi(\phi)\, d\phi = \infty$. This has often been criticized on the grounds that (i) foundational arguments require the use of a proper prior, and (ii) the use of improper priors may lead to unsatisfactory posteriors.

With respect to the foundational issue, we should point out that the natural axioms do *not* imply that the prior must be proper: they only lead to finite additivity, which is compatible with improper measures. However, the further natural assumption of *conglomerability* leads to σ-additivity and, hence, to proper measures; some signposts to this interesting debate are Heath and Sudderth (1978, 1989), Hartigan (1983), Cifarelli and Regazzini (1987), Seidenfeld (1987), Consonni and Veronese (1989) and Lindley (1996). It must be stressed however that, by definition, non-subjective priors are *not* intended to describe personal beliefs: they are *only* positive functions to be formally used in Bayes theorem to obtain non-subjective *posteriors*, —which indeed *should always be proper* given a minimum sample size—. Uncritical use of a "noninformative" prior may lead to an improper posterior (see *e.g.*, Berger, 1985, p. 187, for a well known example); the precise conditions for an improper prior to lead to a proper posterior are not known, but we are not aware of any example where the reference algorithm has lead to an improper posterior given a sample of minimum size.

It is very important to emphasize here that the use of a proper prior does certainly *not* guarantee a sensible behaviour of the resulting posterior. Indeed, if an improper prior leads to a posterior with undesirable properties, the posterior which would result from a proper approximation to that prior, —say that obtained by truncation of the parameter space—, will still have the same undesirable properties; for instance (see Example 24), the posterior of the sum of the squares of normal means $\phi = \sum_{j=1}^m \mu_j^2$ based on a joint uniform prior on the means $\pi(\mu_1, \ldots, \mu_m) \propto 1$ is extremely unsatisfactory as a non-subjective posterior (Stein, 1959), but so it is the posterior of ϕ based on the *proper* multinormal prior $\pi(\mu_1, \ldots, \mu_m) \propto \prod_i \mathrm{N}(\mu_i | 0, \lambda)$, for small precision λ. Proper or improper, what must pragmatically be required from non-subjective priors is that, for any data set, they lead to sensible, data dominated, posterior distributions.

Finally, the use of improper non-subjective prior distributions have sometimes been criticised on the grounds that they may lead to inadmissible estimates (see, e.g. Stein, 1956). However, sensible non-subjective posteriors should be expressible as a *limit* of some sequence of posteriors derived from proper priors (Stone, 1963, 1965, 1970; Stein, 1965; Akaike, 1980a); this is precisely the procedure used to *define* reference distributions. Regarded as a "baseline" for admissible inferences, reference posterior distributions need not be themselves admissible, but only arbitrarily close to admissible posteriors.

4.4. CALIBRATION

Non-subjective posterior credible intervals are often numerically very close, and sometimes identical, to frequentist confidence intervals based on *sufficient* statistics (for an instructive discussion of how unsatisfactory confidence intervals may be when not based on sufficient statistics see Jaynes, 1976). Indeed, the analysis on the frequentist coverage probabilities of credible intervals derived from non-subjective posteriors, —in an attempt to verify whether or not they are "well calibrated" —, has a very long history, and it does provide some bridges between frequentist and Bayesian inference. References within this topic include Lindley (1958), Welch and Peers (1963), Bartholomew (1965), Peers (1965, 1968), Welch (1965), Hartigan (1966, 1983), DeGroot (1973), Robinson (1975, 1978), Rubin (1984), Stein (1985), Chang and Villegas (1986), Tibshirani (1989), Dawid (1991), Severini (1991, 1993, 1994), Ghosh and Mukerjee (1992, 1993), Efron (1993), Mukerjee and Day (1993), Nicolau (1993), DiCiccio and Stern (1994), Samaniego and Reneau (1994), Datta and Ghosh (1995a) and Datta (1996).

This is a very active research area; indeed, the frequentist coverage probabilities of posterior credible intervals have often been an important element in arguing among competing non-subjective posteriors, as in Stein (1985), Efron (1986), Tibshirani (1989), Berger and Bernardo (1989), Ye and Berger (1991), Liseo (1993), Berger and Yang (1994), Yang and Berger (1994), Ghosh, Carlin and Srivastava (1995) and Sun and Ye (1995). Reference posteriors have consistently been found to have very attractive coverage properties, even for small samples, but no general results have been established.

4.5. FURTHER SIGNPOSTS

The classic books by Jeffreys (1961), Lindley (1965) and Box and Tiao (1973) are a must for anyone interested in non-subjective Bayesian inference; they prove that most “textbook” inference problems have a simple non-subjective Bayesian solution, and one which produces credible intervals which are often, *numerically*, either identical or very close to their frequentist “accepted” counterparts, but much easier to obtain (and to interpret). Zellner (1971) is a textbook on econometrics from a non-subjective Bayesian viewpoint; Geisser (1993) summarizes many results on non-subjective posterior *predictive* distributions.

In Section 4.1 we have outlined the interesting history of the topic, which dates back to Laplace (1812), and has known a long modern revival which began with Jeffreys (1939/1961). In this final section, we simply recall its basic milestones: Jeffreys (1946), Perks (1947), Lindley (1961), Geisser and Cornfield (1963), Welch and Peers (1963), Hartigan (1964, 1965), Novick and Hall (1965), Jaynes (1968, 1971), Good (1969), DeGroot (1970, Ch. 10), Villegas (1971, 1977, 1981) Box and Tiao (1973, Sec. 1.3), Zellner (1977, 1986), Akaike (1978), Bernardo (1979), Geisser (1979, 1984), Rissanen (1983), Tibshirani (1989) and Berger and Bernardo (1989, 1992c). The study of the development of this long quest may be completed with the review paper by Kass and Wasserman (1996), and references therein.

Some recent developments in the definition of non-subjective priors include Eaton (1992), Ghosh and Mukerjee (1992), Mukerjee and Dey (1993), Clarke and Wasserman (1993), George and McCulloch (1993), Ye(1993), Clarke and Barron (1994), Wasserman and Clarke (1995), Datta and Ghosh (1995b, 1995c, 1996), Zellner (1996) and Bernardo (1999). Yang and Berger (1996) is a partial *catalog*, alphabetically ordered by probability model, of many non-subjective priors which have been suggested in the literature. Bernardo (1997a) is a non technical analysis, in a dialog format, on the *foundational* issues involved, and it is followed by a discussion.

For someone specifically interested in reference distributions, the original discussion paper, Bernardo (1979b), is easily read and it is followed by a very lively discussion; Bernardo (1981) extends the theory to general decision problems; Berger and Bernardo (1989, 1992c) contain crucial mathematical extensions. A simple introduction to reference analysis is provided in Bernardo and Ramón (1998).

Papers which contain explicit analysis of specific reference distributions include Bernardo (1977, 1978, 1979, 1980, 1982, 1985), Bayarri (1981, 1985), Ferrándiz (1982, 1985), Sendra (1982), Eaves (1983a, 1983b, 1985), Bernardo and Bayarri (1985), Chang and Villegas (1986), Hills (1987), Mendoza (1987, 1988), Bernardo and Girón (1988), Lindley (1988), Berger and Bernardo (1989, 1992a, 1992b, 1992c), Pole and West (1989), Chang and Eaves (1990), Polson and Wasserman (1990), Ye and Berger (1991), Stephens and Smith (1992), Liseo (1993), Ye (1993, 1994, 1995), Berger and Yang, (1994) Kubokawa and Robert (1994), Yang and Berger (1994, 1996), Datta and Ghosh (1995c), Ghosh, Carlin and Srivastava (1995), Sun and Ye (1995), Ghosal (1996) and Reid (1996).

Appendix: Basic Formulae

Two sets of tables are provided for reference. The first set records the definition and some characteristics of the probability distributions used in this monograph. The second set records the basic elements of the Bayesian reference analysis of some simple models.

A.1. PROBABILITY DISTRIBUTIONS

This section consists of a set of tables which record the notation, parameter range, variable range, definition, and first two moments of the probability distributions (discrete and continuous, univariate and multivariate) used in this monograph.

Univariate Discrete Distributions.

Br$(x \mid \theta)$ *Bernoulli*	
$0 < \theta < 1$	$x = 0, 1$
$p(x) = \theta^x (1-\theta)^{1-x}$	
$E[x] = \theta$	$V[x] = \theta(1-\theta)$
Bi$(x \mid \theta, n)$ *Binomial*	
$0 < \theta < 1$, $n = 1, 2, \ldots$	$x = 0, 1, \ldots, n$
$p(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$	
$E[x] = n\theta$	$V[x] = n\theta(1-\theta)$
Bb$(x \mid \alpha, \beta, n)$ *Binomial-Beta*	
$\alpha > 0$, $\beta > 0$, $n = 1, 2, \ldots$	$x = 0, 1, \ldots, n$
$p(x) = c \binom{n}{x} \Gamma(\alpha + x)\Gamma(\beta + n - x)$	$c = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}$
$E[x] = n\dfrac{\alpha}{\alpha+\beta}$	$V[x] = \dfrac{n\alpha\beta}{(\alpha+\beta)^2}\dfrac{(\alpha+\beta+n)}{(\alpha+\beta+1)}$

Univariate Discrete Distributions (continued).

$\mathrm{Hy}(x \mid N, M, n)$ *Hypergeometric*	
$N = 1, 2, \ldots$ $M = 1, 2, \ldots$ $n = 1, \ldots, N + M$	$x = a, a + 1, \ldots, b$ $a = \max(0, n - M)$ $b = \min(n, N)$
$p(x) = c \binom{N}{x} \binom{M}{n - x}$	$c = \binom{N + M}{n}^{-1}$
$E[x] = n \frac{N}{N + M}$	$V[x] = \frac{nNM}{(N + M)^2} \frac{N + M - n}{N + M - 1}$

$\mathrm{Nb}(x \mid \theta, r)$ *Negative-Binomial*	
$0 < \theta < 1$, $r = 1, 2, \ldots$	$x = 0, 1, 2, \ldots$
$p(x) = c \binom{r + x - 1}{r - 1} (1 - \theta)^x$	$c = \theta^r$
$E[x] = r \frac{1 - \theta}{\theta}$	$V[x] = r \frac{1 - \theta}{\theta^2}$

$\mathrm{Nbb}(x \mid \alpha, \beta, r)$ *Negative-Binomial-Beta*	
$\alpha > 0,\ \beta > 0,\ r = 1, 2 \ldots$	$x = 0, 1, 2, \ldots$
$p(x) = c \binom{r + x - 1}{r - 1} \frac{\Gamma(\beta + x)}{\Gamma(\alpha + \beta + r + x)}$	$c = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + r)}{\Gamma(\alpha)\Gamma(\beta)}$
$E[x] = \frac{r\beta}{\alpha - 1}$	$V[x] = \frac{r\beta}{(\alpha - 1)} \left[\frac{\alpha + \beta + r - 1}{(\alpha - 2)} + \frac{r\beta}{(\alpha - 1)(\alpha - 2)} \right]$

$\mathrm{Pn}(x \mid \lambda)$ *Poisson*	
$\lambda > 0$	$x = 0, 1, 2, \ldots$
$p(x) = c\ \frac{\lambda^x}{x!}$	$c = e^{-\lambda}$
$E[x] = \lambda$	$V[x] = \lambda$

$\mathrm{Pg}(x \mid \alpha, \beta, n)$ *Poisson-Gamma*	
$\alpha > 0,\ \beta > 0,\ \gamma > 0$	$x = 0, 1, 2, \ldots$
$p(x) = c \frac{\Gamma(\alpha + x)}{x!} \frac{\gamma^x}{(\beta + \gamma)^{\alpha + x}}$	$c = \frac{\beta^\alpha}{\Gamma(\alpha)}$
$E[x] = \gamma \frac{\alpha}{\beta}$	$V[x] = \frac{\gamma\alpha}{\beta} \left[1 + \frac{\gamma}{\beta} \right]$

Univariate Continuous Distributions.

$\mathrm{Be}(x \mid \alpha, \beta)$ *Beta*	
$\alpha > 0$, $\beta > 0$	$0 < x < 1$
$p(x) = c\, x^{\alpha-1}(1-x)^{\beta-1}$	$c = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$
$E[x] = \dfrac{\alpha}{\alpha+\beta}$	$V[x] = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
$\mathrm{Un}(x \mid a, b)$ *Uniform*	
$b > a$	$a < x < b$
$p(x) = c$	$c = (b-a)^{-1}$
$E[x] = \frac{1}{2}(a+b)$	$V[x] = \frac{1}{12}(b-a)^2$
$\mathrm{Ga}(x \mid \alpha, \beta)$ *Gamma*	
$\alpha > 0$, $\beta > 0$	$x > 0$
$p(x) = c\, x^{\alpha-1}e^{-\beta x}$	$c = \dfrac{\beta^\alpha}{\Gamma(\alpha)}$
$E[x] = \alpha\beta^{-1}$	$V[x] = \alpha\beta^{-2}$
$\mathrm{Ex}(x \mid \theta)$ *Exponential*	
$\theta > 0$	$x > 0$
$p(x) = c\, e^{-\theta x}$	$c = \theta$
$E[x] = 1/\theta$	$V[x] = 1/\theta^2$
$\mathrm{Gg}(x \mid \alpha, \beta, n)$ *Gamma-Gamma*	
$\alpha > 0$, $\beta > 0$, $n > 0$	$x > 0$
$p(x) = c\, \dfrac{x^{n-1}}{(\beta+x)^{\alpha+n}}$	$c = \dfrac{\beta^\alpha}{\Gamma(\alpha)}\, \dfrac{\Gamma(\alpha+n)}{\Gamma(n)}$
$E[x] = n\dfrac{\beta}{\alpha-1}$	$V[x] = \dfrac{\beta^2(n^2+n(\alpha-1))}{(\alpha-1)^2(\alpha-2)}$
$\chi^2(x \mid \nu) = \chi^2_\nu$ *Chi-squared*	
$\nu > 0$	$x > 0$
$p(x) = c\, x^{(\nu/2)-1}e^{-x/2}$	$c = \dfrac{(1/2)^{\nu/2}}{\Gamma(\nu/2)}$
$E[x] = \nu$	$V[x] = 2\nu$

Univariate Continuous Distributions (continued).

$\chi^2(x \mid \nu, \lambda)$ *Non-central Chi-squared*	
$\nu > 0,\ \lambda > 0$	$x > 0$
$p(x) = \sum_{i=0}^{\infty} \mathrm{Pn}\left(i \,\middle|\, \frac{\lambda}{2}\right) \chi^2(x \mid \nu + 2i)$	
$E[x] = \nu + \lambda$	$V[x] = 2(\nu + 2\lambda)$
$\mathrm{Ig}(x \mid \alpha, \beta)$ *Inverted-Gamma*	
$\alpha > 0,\ \beta > 0$	$x > 0$
$p(x) = c\, x^{-(\alpha+1)} e^{-\beta/x}$	$c = \frac{\beta^\alpha}{\Gamma(\alpha)}$
$E[x] = \frac{\beta}{\alpha - 1}$	$V[x] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
$\chi^{-1}(x \mid \nu)$ *Inverted-Chi-squared*	
$\nu > 0$	$x > 0$
$p(x) = c\ x^{-(\nu/2+1)} e^{-1/2x^2}$	$c = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)}$
$E[x] = \frac{1}{\nu - 2}$	$V[x] = \frac{2}{(\nu-2)^2(\nu-4)}$
$\mathrm{Ga}^{-1/2}(x \mid \alpha, \beta)$ *Square-root Inverted-Gamma*	
$\alpha > 0,\ \beta > 0$	$x > 0$
$p(x) = c\ x^{-(2\alpha+1)} e^{-\beta/x^2}$	$c = \frac{2\beta^\alpha}{\Gamma(\alpha)}$
$E[x] = \frac{\sqrt{\beta}\Gamma(\alpha - 1/2)}{\Gamma(\alpha)}$	$V[x] = \frac{\beta}{\alpha - 1} - E[x]^2$
$\mathrm{Pa}(x \mid \alpha, \beta)$ *Pareto*	
$\alpha > 0,\ \beta > 0$	$\beta \le x < +\infty$
$p(x) = c\ x^{-(\alpha+1)}$	$c = \alpha\beta^\alpha$
$E[x] = \beta\alpha(\alpha-1)^{-1}$	$V[x] = \beta^2\alpha(\alpha-1)^{-2}(\alpha-2)^{-1}$
$\mathrm{Ip}(x \mid \alpha, \beta)$ *Inverted-Pareto*	
$\alpha > 0, \quad \beta > 0$	$0 < x < \beta^{-1}$
$p(x) = c\, x^{\alpha-1}$	$c = \alpha\beta^\alpha$
$E[x] = \beta^{-1}\alpha(\alpha+1)^{-1}$	$V[x] = \beta^{-2}\alpha(\alpha+1)^{-2}(\alpha+2)^{-1}$

Univariate Continuous Distributions (continued).

N$(x \mid \mu, \lambda)$ *Normal*	
$-\infty < \mu < +\infty$, $\lambda > 0$	$-\infty < x < +\infty$
$p(x) = c\ \exp\left\{-\frac{1}{2}\lambda(x-\mu)^2\right\}$	$c = \lambda^{1/2}(2\pi)^{-1/2}$
$E[x] = \mu$	$V[x] = \lambda^{-1}$

St$(x \mid \mu, \lambda, \alpha)$ *Student t*	
$-\infty < \mu < +\infty$, $\lambda > 0$, $\alpha > 0$	$-\infty < x < +\infty$
$p(x) = c\ \left[1 + \alpha^{-1}\lambda(x-\mu)^2\right]^{-(\alpha+1)/2}$	$c = \dfrac{\Gamma\left(\frac{1}{2}(\alpha+1)\right)}{\Gamma(\frac{1}{2}\alpha)} \left(\dfrac{\lambda}{\alpha\pi}\right)^{1/2}$
$E[x] = \mu$	$V[x] = \lambda^{-1}\alpha(\alpha-2)^{-1}$

F$(x \mid \alpha, \beta) = \mathrm{F}_{\alpha,\beta}$ *Snedecor F*	
$\alpha > 0$, $\beta > 0$	$x > 0$
$p(x) = c\ \dfrac{x^{\alpha/2-1}}{(\beta+\alpha x)^{(\alpha+\beta)/2}}$	$c = \dfrac{\Gamma\left(\frac{1}{2}(\alpha+\beta)\right)\alpha^{\alpha/2}\beta^{\beta/2}}{\Gamma(\frac{1}{2}\alpha)\Gamma(\frac{1}{2}\beta)}$
$E[x] = \dfrac{\beta}{\beta-2}$	$V[x] = \dfrac{2\beta^2(\alpha+\beta-2)}{\alpha(\beta-2)^2(\beta-4)}$

Lo$(x \mid \alpha, \beta)$ *Logistic*	
$-\infty < \alpha < +\infty$, $\beta > 0$	$-\infty < x < +\infty$
$p(x) = \beta^{-1}\ \exp\left\{-\beta^{-1}(x-\alpha)\right\}\left[1 + \exp\left\{-\beta^{-1}(x-\alpha)\right\}\right]^{-2}$	
$E[x] = \alpha$	$V[x] = \beta^2\pi^2/3$

Multivariate Discrete Distributions.

Mu$_k(\boldsymbol{x} \mid \boldsymbol{\theta}, n)$ *Multinomial*		
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$		$\boldsymbol{x} = (x_1, \ldots, x_k)$
$0 < \theta_i < 1$, $\sum_{\ell=1}^{k} \theta_\ell \leq 1$		$\sum_{\ell=1}^{k} x_\ell \leq n$
$n = 1, 2, \ldots$		$x_i = 0, 1, 2, \ldots$
$p(\boldsymbol{x}) = \dfrac{n!}{\prod_{\ell=1}^{k+1} x_\ell!} \displaystyle\prod_{\ell=1}^{k+1} \theta^{x_\ell}$,	$\theta_{k+1} = 1 - \displaystyle\sum_{\ell=1}^{k} \theta_\ell$,	$x_{k+1} = n - \displaystyle\sum_{\ell=1}^{k} x_\ell$
$E[x_i] = n\theta_i$	$V[x_i] = n\theta_i(1-\theta_i)$	$C[x_i, x_j] = -n\theta_i\theta_j$

Multivariate Continuous Distributions.

$\mathrm{Md}_k(\boldsymbol{x}\mid\boldsymbol{\theta},n)$ *Multinomial-Dirichlet*

$\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{k+1})$	$\boldsymbol{x}=(x_1,\ldots,x_k)$
$\alpha_i>0$	$x_i=0,1,2,\ldots$
$n=1,2,\ldots$	$\sum_{\ell=1}^{n}x_l\le n$
$p(\boldsymbol{x})=c\prod_{\ell=1}^{k+1}\frac{\alpha_\ell^{[x_\ell]}}{x_\ell!}$	$c=\frac{n!}{\left(\sum_{\ell=1}^{k+1}\alpha_\ell\right)^{[n]}}$
$\alpha^{[s]}=\prod_{\ell=1}^{s}(\alpha+\ell-1)$	$x_{k+1}=n-\sum_{\ell=1}^{k}x_\ell$
$E[x_i]=np_i$	$V[x_i]=\frac{n+\sum_{\ell=1}^{k+1}\alpha_\ell}{1+\sum_{\ell=1}^{k+1}\alpha_\ell}\,np_i(1-p_i)$
$p_i=\frac{\alpha_i}{\sum_{\ell=1}^{k+1}\alpha_\ell}$	$C[x_i,x_j]=-\frac{n+\sum_{\ell=1}^{k+1}\alpha_\ell}{1+\sum_{\ell=1}^{k+1}\alpha_\ell}\,np_ip_j$

$\mathrm{Di}_k(\boldsymbol{x}\mid\boldsymbol{\alpha})$ *Dirichlet*

$\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{k+1})$	$\boldsymbol{x}=(x_1,\ldots,x_k)$
$\alpha_i>0$	$0<x_i<1,\quad\sum_{\ell=1}^{k}x_\ell\le 1$
$p(\boldsymbol{x})=c\left(1-\sum_{\ell=1}^{k}x_\ell\right)^{\alpha_{k+1}-1}\prod_{\ell=1}^{k}x_\ell^{\alpha_\ell-1}$	$c=\frac{\Gamma(\sum_{\ell=1}^{k+1}\alpha_\ell)}{\prod_{\ell=1}^{k+1}\Gamma(\alpha_\ell)}$

$$E[x_i]=\frac{\alpha_i}{\sum_{\ell=1}^{k+1}\alpha_\ell}\qquad V[x_i]=\frac{E[x_i](1-E[x_i])}{1+\sum_{\ell=1}^{k+1}\alpha_\ell}\qquad C[x_i,x_j]=\frac{-E[x_i]E[x_j]}{1+\sum_{\ell=1}^{k+1}\alpha_\ell}$$

$\mathrm{Ng}(x,y\mid\mu,\lambda,\alpha,\beta)$ *Normal-Gamma*

$\mu\in\Re,\ \lambda>0,\ \alpha>0,\ \beta>0,$	$x\in\Re,\ y>0$

$p(x,y)=\mathrm{N}(x\mid\mu,\lambda y)\ \mathrm{Ga}(y\mid\alpha,\beta)$

$E[x]=\mu\qquad E[y]=\alpha\beta^{-1}\qquad V[x]=\beta\lambda^{-1}(\alpha-1)^{-1}\qquad V[y]=\alpha\beta^{-2}$

$p(x)=\mathrm{St}(x\mid\mu,\alpha\beta^{-1}\lambda,2\alpha)$

$\mathrm{N}_k(\boldsymbol{x}\mid\boldsymbol{\mu},\boldsymbol{\lambda})$ *Multivariate Normal*

$\boldsymbol{\mu}=(\mu_1,\ldots,\mu_k)\in\Re^k$	$\boldsymbol{x}=(x_1,\ldots,x_k)\in\Re^k$
$\boldsymbol{\lambda}$ symmetric positive-definite	
$p(\boldsymbol{x})=c\ \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\lambda}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$	$c=\ \vert\boldsymbol{\lambda}\vert^{1/2}(2\pi)^{-k/2}$
$E[\boldsymbol{x}]=\boldsymbol{\mu}$	$V[\boldsymbol{x}]=\boldsymbol{\lambda}^{-1}$

Multivariate Continuous Distributions (continued).

$\mathrm{Pa}_2(x, y \mid \alpha, \beta_0, \beta_1)$ *Bilateral Pareto*

$(\beta_0, \beta_1) \in \Re^2,\ \beta_0 < \beta_1,\ \alpha > 0$ $\quad$ $(x, y) \in \Re^2,\ x < \beta_0,\ y > \beta_1$

$p(x, y) = c\ (y - x)^{-(\alpha+2)}$ $\quad$ $c = \alpha(\alpha+1)(\beta_1 - \beta_0)^{\alpha}$

$$E[x] = \frac{\alpha\beta_0 - \beta_1}{\alpha - 1} \quad E[y] = \frac{\alpha\beta_1 - \beta_0}{\alpha - 1} \qquad V[x] = V[y] = \frac{\alpha(\beta_1 - \beta_0)^2}{(\alpha-1)^2(\alpha-2)}$$

$\mathrm{Ng}_k(\boldsymbol{x}, y \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha, \beta)$ *Multivariate Normal-Gamma*

$-\infty < \mu_i < +\infty,\ \alpha > 0,\ \beta > 0$ $\quad$ $(\boldsymbol{x}, y) = (x_1, \ldots, x_k, y)$

$\boldsymbol{\lambda}$ symmetric positive-definite $\quad$ $-\infty < x_i < \infty, \quad y > 0$

$p(\boldsymbol{x}, y) = \mathrm{N}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\lambda} y)\ \mathrm{Ga}(y \mid \alpha, \beta)$

$E[\boldsymbol{x}, y] = (\boldsymbol{\mu},\ \alpha\beta^{-1}), \qquad V[\boldsymbol{x}] = (\alpha - 1)^{-1}\beta\boldsymbol{\lambda}^{-1}, \qquad V[y] = \alpha\beta^{-2}$

$p(\boldsymbol{x}) = \mathrm{St}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}\alpha\beta^{-1}, 2\alpha)$ $\quad$ $p(y) = Ga(y \mid \alpha, \beta)$

$\mathrm{Nw}_k(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\mu}, \lambda, \alpha, \boldsymbol{\beta})$ *Multivariate Normal-Wishart*

$-\infty < \mu_i < +\infty,\ \lambda > 0$ $\quad$ $\boldsymbol{x} = (x_1, \ldots, x_k)$

$2\alpha > k - 1$ $\quad$ $-\infty < x_i < +\infty$

$\boldsymbol{\beta}$ symmetric non-singular $\quad$ $\boldsymbol{y}$ symmetric positive-definite

$p(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{N}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \lambda\boldsymbol{y})\ \mathrm{Wi}_k(\boldsymbol{y} \mid \alpha, \boldsymbol{\beta})$

$E[\boldsymbol{x}, \boldsymbol{y}] = \{\boldsymbol{\mu}, \alpha\boldsymbol{\beta}^{-1}\}$ $\quad$ $V[\boldsymbol{x}] = (\alpha - 1)^{-1}\boldsymbol{\beta}\lambda^{-1}$

$p(\boldsymbol{x}) = \mathrm{St}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \lambda\alpha\boldsymbol{\beta}^{-1}, 2\alpha)$ $\quad$ $p(\boldsymbol{y}) = \mathrm{Wi}_k(\boldsymbol{y} \mid \alpha, \boldsymbol{\beta})$

$\mathrm{St}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}, \alpha)$ *Multivariate Student*

$-\infty < \mu_i < +\infty,\ \alpha > 0$ $\quad$ $\boldsymbol{x} = (x_1, \ldots, x_k)$

$\boldsymbol{\lambda}$ symmetric positive-definite $\quad$ $-\infty < x_i < +\infty$

$$p(\boldsymbol{x}) = c\left[1 + \frac{1}{\alpha}(\boldsymbol{x} - \boldsymbol{\mu})^t\boldsymbol{\lambda}(\boldsymbol{x} - \boldsymbol{\mu})\right]^{-(\alpha+k)/2} \qquad c = \frac{\Gamma\left(\frac{1}{2}(\alpha + k)\right)}{\Gamma(\frac{1}{2}\alpha)(\alpha\pi)^{k/2}}|\boldsymbol{\lambda}|^{1/2}$$

$E[\boldsymbol{x}] = \boldsymbol{\mu}, \quad V[\boldsymbol{x}] = \boldsymbol{\lambda}^{-1}(\alpha - 2)^{-1}\alpha$

$\mathrm{Wi}_k(\boldsymbol{x} \mid \alpha, \boldsymbol{\beta})$ *Wishart*

$2\alpha > k - 1$ $\quad$ $\boldsymbol{x}$ symmetric positive-definite

$\boldsymbol{\beta}$ symmetric non-singular

$$p(\boldsymbol{x}) = c\,|\boldsymbol{x}|^{\alpha-(k+1)/2}\exp\{-\operatorname{tr}(\boldsymbol{\beta}\boldsymbol{x})\} \qquad c = \frac{\pi^{-k(k-1)/4}|\boldsymbol{\beta}|^{\alpha}}{\prod_{\ell=1}^{k}\Gamma(\frac{1}{2}(2\alpha + 1 - \ell))}$$

$E[\boldsymbol{x}] = \alpha\boldsymbol{\beta}^{-1}, \quad E[\boldsymbol{x}^{-1}] = (\alpha - \frac{k+1}{2})^{-1}\boldsymbol{\beta}$

A.2. INFERENTIAL PROCESSES

This section records the basic elements of the Bayesian reference analysis of some commonly used statistical models.

For each of these models, we provide, in separate sections of the table, the following: the model, the sufficient statistic and its sampling distribution; the reference prior(s) and the corresponding reference posterior(s), and the reference posterior predictive for a single future observation.

In the case of uniparameter models this can always be done. We recall, however, from Section 3.4 that, in multiparameter problems, the reference prior is only defined *relative to an ordered parametrisation*. In the univariate normal model $\mathrm{N}(x \,|\, \mu, \lambda)$ (Example 14), the reference prior for the orfdered parametrisation (μ, λ) happens to be the same as that for (λ, μ), namely $\pi(\mu, \lambda) = \pi(\lambda, \mu) \propto \lambda^{-1}$, and we provide the corresponding reference posteriors for μ and λ, together with the reference predictive distribution for a future observation.

In the multinomial, multivariate normal and linear regression models, however, there are very many different reference priors, corresponding to different inference problems, and specified by different ordered parametrisations. These are not reproduced in this Appendix.

Bernoulli model.

$\boldsymbol{z} = \{x_1, \ldots, x_n\}$,	$x_i \in \{0, 1\}$
$p(x_i \,|\, \theta) = \mathrm{Br}(x_i \,|\, \theta)$,	$0 < \theta < 1$
$t(\boldsymbol{z}) = r = \sum_{i=1}^{n} x_i$	
$p(r \,|\, \theta) = \mathrm{Bi}(r \,|\, \theta, n)$	
$\pi(\theta) = \mathrm{Be}(\theta \,|\, \frac{1}{2}, \frac{1}{2})$	
$\pi(\theta \,|\, \boldsymbol{z}) = \mathrm{Be}(\theta \,|\, \frac{1}{2} + r, \frac{1}{2} + n - r)$	
$\pi(x \,|\, \boldsymbol{z}) = \mathrm{Bb}(x \,|\, \frac{1}{2} + r, \frac{1}{2} + n - r, 1)$	

Poisson Model.

$\boldsymbol{z} = \{x_1, \ldots, x_n\}$,	$x_i = 0, 1, 2, \ldots$
$p(x_i \,|\, \lambda) = \mathrm{Pn}(x_i \,|\, \lambda)$,	$\lambda \geq 0$
$t(\boldsymbol{z}) = r = \sum_{i=1}^{n} x_i$	
$p(r \,|\, \lambda) = \mathrm{Pn}(r \,|\, n\lambda)$	
$\pi(\lambda) \propto \lambda^{-1/2}$	
$\pi(\lambda \,|\, \boldsymbol{z}) = \mathrm{Ga}(\lambda \,|\, r + \frac{1}{2}, n)$	
$\pi(x \,|\, \boldsymbol{z}) = \mathrm{Pg}(x \,|\, r + \frac{1}{2}, n, 1)$	

Negative-Binomial model.

$\boldsymbol{z} = (x_1, \ldots, x_n), \qquad x_i = 0, 1, 2, \ldots$
$p(x_i \,|\, \theta) = \text{Nb}(x_i \,|\, \theta, r), \qquad 0 < \theta < 1$

$t(\boldsymbol{z}) = s = \sum_{i=1}^{n} x_i$
$p(s \,|\, \theta) = \text{Nb}(s \,|\, \theta, nr)$

$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$
$\pi(\theta \,|\, \boldsymbol{z}) = \text{Be}(\theta \,|\, nr, s + \frac{1}{2})$
$\pi(x \,|\, \boldsymbol{z}) = \text{Nbb}(x \,|\, nr, s + \frac{1}{2}, r)$

Exponential Model.

$\boldsymbol{z} = \{x_1, \ldots, x_n\}, \qquad 0 < x_i < \infty$
$p(x_i \,|\, \theta) = \text{Ex}(x_i \,|\, \theta), \qquad \theta > 0$

$t(\boldsymbol{z}) = t = \sum_{i=1}^{n} x_i$
$p(t \,|\, \theta) = \text{Ga}(t \,|\, n, \theta)$

$\pi(\theta) \propto \theta^{-1}$
$\pi(\theta \,|\, \boldsymbol{z}) = \text{Ga}(\theta \,|\, n, t)$
$\pi(x \,|\, \boldsymbol{z}) = \text{Gg}(x \,|\, n, t, 1)$

Uniform Model.

$\boldsymbol{z} = \{x_1, \ldots, x_n\}, \qquad 0 < x_i < \theta$
$p(x_i \,|\, \theta) = \text{Un}(x_i \,|\, 0, \theta), \qquad \theta > 0$

$t(\boldsymbol{z}) = t = \max\{x_1, \ldots, x_n\}$
$p(t \,|\, \theta) = \text{Ip}(t \,|\, n, \theta^{-1})$

$\pi(\theta) \propto \theta^{-1}$
$\pi(\theta \,|\, \boldsymbol{z}) = \text{Pa}(\theta \,|\, n, t)$
$\pi(x \,|\, \boldsymbol{z}) = \frac{n}{n+1}\text{Un}(x \,|\, 0, t), \text{ if } x \leq t, \quad \frac{1}{n+1}\text{Pa}(x \,|\, n, t), \text{ if } x > t$

Normal Model (known precision λ).

$\boldsymbol{z} = \{x_1, \ldots, x_n\}, \qquad -\infty < x_i < \infty$

$p(x_i \mid \mu, \lambda) = \mathrm{N}(x_i \mid \mu, \lambda), \qquad -\infty < \mu < \infty$

$t(\boldsymbol{z}) = \bar{x} = n^{-1} \sum_{i=1}^{n} x_i$

$p(\bar{x} \mid \mu, \lambda) = \mathrm{N}(x \mid \mu, n\lambda)$

$\pi(\mu) =$ constant

$\pi(\mu \mid \boldsymbol{z}) = \mathrm{N}(\mu \mid \bar{x}, n\lambda)$

$\pi(x \mid \boldsymbol{z}) = \mathrm{N}(x \mid \bar{x}, \lambda\, n(n+1)^{-1})$

Normal Model (known mean μ).

$\boldsymbol{z} = \{x_1, \ldots, x_n\}, \qquad -\infty < x_i < \infty$

$p(x_i \mid \mu, \lambda) = \mathrm{N}(x_i \mid \mu, \lambda), \qquad \lambda > 0$

$t(\boldsymbol{z}) = t = \sum_{i=1}^{n} (x_i - \mu)^2$

$p(t \mid \mu, \lambda) = \mathrm{Ga}(t \mid \tfrac{1}{2}n, \tfrac{1}{2}\lambda), \qquad p(\lambda t) = \chi^2(\lambda t \mid n)$

$\pi(\lambda) \propto \lambda^{-1}$

$\pi(\lambda \mid \boldsymbol{z}) = \mathrm{Ga}(\lambda \mid \tfrac{1}{2}n, \tfrac{1}{2}t)$

$\pi(x \mid \boldsymbol{z}) = \mathrm{St}(x \mid \mu, nt^{-1}, n)$

Normal Model (both parameters unknown).

$\boldsymbol{z} = \{x_1, \ldots, x_n\}, \qquad -\infty < x_i < \infty$

$p(x_i \mid \mu, \lambda) = \mathrm{N}(x_i \mid \mu, \lambda), \qquad -\infty < \mu < \infty, \quad \lambda > 0$

$t(\boldsymbol{z}) = (\bar{x}, s), \qquad n\bar{x} = \sum_{i=1}^{n} x_i, \qquad ns^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$

$p(\bar{x} \mid \mu, \lambda) = \mathrm{N}(\bar{x} \mid \mu, n\lambda)$

$p(ns^2 \mid \mu, \lambda) = \mathrm{Ga}(ns^2 \mid \tfrac{1}{2}(n-1), \tfrac{1}{2}\lambda), \qquad p(\lambda ns^2) = \chi^2(\lambda ns^2 \mid n-1)$

$\pi(\mu, \lambda) = \pi(\lambda, \mu) \propto \lambda^{-1}, \qquad n > 1$

$\pi(\mu \mid \boldsymbol{z}) = \mathrm{St}(\mu \mid \bar{x}, (n-1)s^{-2}, n-1)$

$\pi(\lambda \mid \boldsymbol{z}) = \mathrm{Ga}\left(\lambda \mid \tfrac{1}{2}(n-1), \tfrac{1}{2}ns^2\right)$

$\pi(x \mid \boldsymbol{z}) = \mathrm{St}\left(x \mid \bar{x}, (n-1)(n+1)^{-1}s^{-2}, n-1\right)$

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions*. New York: Dover.

Aitchison, J. (1964). Bayesian tolerance regions. *J. Roy. Statist. Soc. B* **26**, 161–175.

Aitchison, J. (1966). Expected cover and linear utility tolerance intervals. *J. Roy. Statist. Soc. B* **28**, 57–62.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547–554.

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: University Press.

Aitkin, M. (1991). Posterior Bayes factors. *J. Roy. Statist. Soc. B* **53**, 111–142 (with discussion).

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd. Int. Symp. Information Theory*. Budapest: Akademia Kaido, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19**, 716–727.

Akaike, H. (1978a). A new look at the Bayes procedure. *Biometrika* **65**, 53–59.

Akaike, H. (1978b). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30**, 9–14.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fifting. *Biometrika* **66**, 237–242.

Akaike, H. (1980a). The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J. Roy. Statist. Soc. B* **45**, 46–52.

Akaike, H. (1980b). Likelihood and the Bayes procedure. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 144–166 and 185–203 (with discussion).

Akaike, H. (1983). Information measures and model selection. *Bull. Internat. Statist. Institute* **50**, 277–290.

Akaike, H. (1987). Factor analysis and the AIC. *Psychometrika* **52**, 317–332.

Albert, J. H. (1989). Nuisance parameters and the use of exploratory graphical methods in Bayesian analysis. *Amer. Statist.* **43**, 191–196.

Albert, J. H. (1990a). A Bayesian test for a two-way contingency table using independence priors. *Canadian J. Statist.* **14**, 1583–1590.

Albert, J. H. (1990b). Algorithms for Bayesian computing using Mathematica. *Computing Science and Statistics: Proceedings of the Symposium on the Interface* (C. Page and R. LePage eds.). Berlin: Springer, 286–290.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669–679.

Amaral-Turkman, M. A. and Dunsmore, I. R. (1985). Measures of information in the predictive distribution. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 603–612.

Anscombe, F. J. (1961). Bayesian statistics. *Amer. Statist.* **15**, 21–24.

Anscombe, F. J. (1963). Sequential medical trials. *J. Amer. Statist. Assoc.* **58**, 365–383.

Anscombe, F. J. (1964a). Some remarks on Bayesian statistics. *Human Judgement and Optimality* (Shelly and Bryan, eds.). New York: Wiley, 155–177.

Anscombe, F. J. (1964b). Normal likelihood functions. *Ann. Inst. Statist. Math.* **16**, 1–41.

Bar-lev, S. K. and Reiser, B. (1982). An exponential subfamily which admits UMPU tests based on a single test statistic. *Ann. Statist.* **10**, 979–989.

Barnard, G. A. (1949). Statistical inference. *J. Roy. Statist. Soc. B* **11**, 115–149 (with discussion).

Barnard, G. A. (1951). The theory of information. *J. Roy. Statist. Soc. B* **13**, 46–64.

Barnard, G. A. (1952). The frequency justification of certain sequential tests. *Biometrika* **39**, 155–150.

Barnard, G. A. (1963). Some aspects of the fiducial argument. *J. Roy. Statist. Soc. B* **25**, 111-114.

Barnard, G. A. (1967). The use of the likelihood function in statistical practice. *Proc. Fifth Berkeley Symp.* **1** (L. M. LeCam and J Neyman, eds.). Berkeley: Univ. California Press, 27–40.

Barnard, G. A. (1980a). In discussion of Box (1980). *J. Roy. Statist. Soc. A* **143**, 404–406.

Barnard, G. A. (1980b). Pivotal inference and the Bayesian controversy. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 295–318 (with discussion).

Barnard, G. A. (1995). Pivotal models and the fiducial argument. *Internat. Statist. Rev.* **63**, 309–323; correction 64, 137.

Barnard, G. A., Jenkins, G. M. and Winsten, C. B. (1962). Likelihood inference and time series. *J. Roy. Statist. Soc. A* **125**, 321–372 (with discussion).

Barnard, G. A. and Sprott, D. A. (1968). Likelihood. *Encyclopedia of Statistical Sciences* **9** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 639–644.

Barnett, V. (1973/1982). *Comparative Statistical Inference*. Second edition in 1982, Chichester: Wiley.

Bartholomew, D. J. (1965). A comparison of some Bayesian and frequentist inferences. *Biometrika* **52**, 19–35.

Bartholomew, D. J. (1967). Hypothesis testing when the sample size is treated as a random variable. *J. Roy. Statist. Soc. B* **29**, 53–82.

Bartholomew, D. J. (1971). A comparison of Bayesian and frequentist approaches to inferences with prior knowledge. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 417–434 (with discussion).

Bartholomew, D. J. (1994). Bayes theorem in latent variable modelling. *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.). Chichester: Wiley, (to appear).

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533–534.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā A* **37**, 1–71 (with discussion).

Basu, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72**, 355-366.

Basu, D. (1988). *Statistical Information and Likelihood: a Collection of Critical Essays* (J. K. Ghosh, ed.). Berlin: Springer.

Basu, D. (1992). Learning statistics from counter examples: ancillary statistics. *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.). Berlin: Springer, 217–224.

Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población normal bivariante. *Trab. Estadist.* **32**, 18–31.

Bayarri, M. J. (1985). Bayesian inference on the parameters of a Beta distribution. *Statistics and Decisions* **2**, 17–22.

Bayarri, M. J. (1987). Comment to Berger and Delampady. *Statist. Sci.* **3**, 342–344.

Bayarri, M. J. and DeGroot, M. H. (1988). Gaining weight: a Bayesian approach. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 25–44 (with discussion).

Bayarri, M. J. and DeGroot, M. H. (1989). Optimal reporting of predictions. *J. Amer. Statist. Assoc.* **84**, 214–222.

Bayarri, M. J. and DeGroot, M. H. (1992b). Difficulties and ambiguities in the definition of a likelihood function. *J. It. Statist. Soc.* **1**, 1–15.

Bayarri, M. J., DeGroot, M. H. and Kadane, J. B. (1988). What is the likelihood function? *Statistical Decision Theory and Related Topics IV* **1** (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 3–27.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Published posthumously in *Phil. Trans. Roy. Soc. London* **53**, 370–418 and **54**, 296–325. Reprinted in *Biometrika* **45** (1958), 293–315, with a biographical note by G. A. Barnard. Reproduced in Press (1989), 185–217.

Berger, J. O. (1979). Multivariate estimation with nonsymmetric loss functions. *Optimizing Methods in Statistics* (J. S. Rustagi, ed.). New York: Academic Press.

Berger, J. O. (1985a). *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer.

Berger, J. O. (1985b). In defense of the likelihood principle: axiomatics and coherence. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 33–65, (with discussion).

Berger, J. O. (1986). Bayesian salesmanship. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Amsterdam: North-Holland, 473–488.

Berger, J. (1992). Discussion of Ghosh and Mukerjee. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 205–206.

Berger, J. O. (1993). The present and future of Bayesian multivariate analysis. *Multivariate Analysis: Future Directions* (C. R. Rao, ed.). Amsterdam: North-Holland, 25–53.

Berger, J. O. (1994). A review of recent developments in robust Bayesian analysis. *Test* **3**, 5–124., (with discussion).

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200–207.

Berger, J. O. and Bernardo, J. M. (1992a). Ordered group reference priors with applications to a multinomial problem. *Biometrika* **79**, 25–37.

Berger, J. O. and Bernardo, J. M. (1992b). Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.). Berlin: Springer, 323–340.

Berger, J. O. and Bernardo, J. M. (1992c). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 35–60 (with discussion).

Berger, J. O., Bernardo, J. M. and Mendoza, M. (1989). On priors that maximize expected information. *Recent Developments in Statistics and their Applications* (J. P. Klein and J. C. Lee, eds.). Seoul: Freedom Academy, 1–20.

Berger, J. O. and Berry, D. A. (1988). The relevance of stopping rules in statistical inference. *Statistical Decision Theory and Related Topics IV* **1** (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 29–72 (with discussion).

Berger, J. O., Boukai, B. and Wang, Y. (1997). Unified Bayesian and frequentist testing for a precise hypothesis. *Statist. Sci.* **12**, 133–160.

Berger, J. O., Brown, L. D. and Wolpert, R. L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Ann. Statist.* **22**, 1787–1807.

Berger, J. O. and DasGupta, A. (1991). *Multivariate Estimation, Bayes, Empirical Bayes and Stein Approaches*. Philadelphia, PA: SIAM.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2**, 317–352 (with discussion).

Berger, J. O. and Fan, T. H. (1991). Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions. *Statistics and Decisions* **10**, 99–120.

Berger, J. O. and Jefferys, W. H. (1992) The application of robust Bayesian analysis to hypothesis testing and Occam's razor. *J. It. Statist. Soc.* **1**, 17–32.

Berger, J. O. and Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *Internat. Statist. Rev.* **59**, 337–353.

Berger, J. O. and Mortera, J. (1994). Robust Bayesian hypothesis testing in the presence of nuisance parameters. *J. Statist. Planning and Inference* **40**, 357–373.

Berger, J. O. and Pericchi, L. R. (1995). The intrinsec Bayes factor for linear models. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 25–44 (with discussion).

Berger, J. O. and Pericchi, L. R. (1996). The intrinsec Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.

Berger, J. O. and Robert, C. P. (1990). Subjective hierarchical Bayes estimation of a multivariate normal mean: on the frequentist interface. *Ann. Statist.* **18**, 617–651.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of significance levels and evidence. *J. Amer. Statist. Assoc.* **82**, 112–133 (with discussion).

Berger, J. O. and Srinivasan, C. (1978). Generalized Bayes estimators in multivariate problems. *Ann. Statist.* **6**, 783–801.

Berger, J. O. and Strawderman, W. E. (1986). Choice of hierarchical priors: admissibility of estimation of normal means. *Ann. Statist.* **24**, 931–951.

Berger, J. O. and Wolpert, R. L. (1984/1988). *The Likelihood Principle*. Second edition in 1988, Hayward, CA: IMS.

Berger, J. O. and Yang, R. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory* **10**, 461–482.

Berk, R. H. (1966). Limiting behaviour of the posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37**, 51–58.

Berk, R. H. (1970). Consistency a posteriori. *Ann. Math. Statist.* **41**, 894–906.

Berliner, L. M. (1987). Bayesian control in mixture models. *Technometrics* **29**, 455–460.

Berliner, L. M. and Goel P. K. (1990). Incorporating partial prior information: ranges of posterior probabilities. *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, eds.). Amsterdam: North-Holland, 397–406.

Bermúdez. J. D. (1985). On the asymptotic normality of the posterior distribution of the logistic classification model. *Statistics and Decisions* **2**, 301–308.

Bernardo, J. M. (1977a). Inferences about the ratio of normal means: a Bayesian approach to the Fieller-Creasy problem. *Recent Developments in Statistics* (J. R. Barra, F. Brodeau, G. Romier and B. van Cutsem eds.). Amsterdam: North-Holland, 345–349.

Bernardo, J. M. (1977b). Inferencia Bayesiana sobre el coeficiente de variación: una solución a la paradoja de marginalización. *Trab. Estadist.* **28**, 23-80.

Bernardo, J. M. (1978a). Una medida de la información útil proporcionada por un experimento. *Rev. Acad. Ciencias Madrid* **72**, 419–440.

Bernardo, J. M. (1978b). Unacceptable implications of the left Haar measure in a standard normal theory inference problem *Trab. Estadist.* **29**, 3–9.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.). Brookfield, VT: Edward Elgar, 1995, 229–263.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 605–647 (with discussion).

Bernardo, J. M. (1981a). Reference decisions. *Symposia Mathematica* **25**, 85–94.

Bernardo, J. M. (1981b). *Bioestadística, una Perspectiva Bayesiana*. Barcelona: Vicens-Vives.

Bernardo, J. M. (1982). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trab. Estadist.* **33**, 16–30.

Bernardo, J. M. (1984). Monitoring the 1982 Spanish socialist victory: a Bayesian analysis. *J. Amer. Statist. Assoc.* **79**, 510–515.

Bernardo, J. M. (1985a). Análisis Bayesiano de los contrastes de hipótesis paramétricos. *Trab. Estadist.* **36**, 45–54.

Bernardo, J. M. (1985b). On a famous problem of induction. *Trab. Estadist.* **36**, 24–30.

Bernardo, J. M. (1987). Approximations in statistics from a decision-theoretical viewpoint. *Probability and Bayesian Statistics* (R. Viertl, ed.). London: Plenum, 53–60.

Bernardo, J. M. (1988). Bayesian linear probabilistic classification. *Statistical Decision Theory and Related Topics IV* **1** (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 151–162.

Bernardo, J. M. (1989). Análisis de datos y métodos Bayesianos. *Historia de la Ciencia Estadística* (S. Ríos, ed.). Madrid: Academia de Ciencias, 87–105.

Bernardo, J. M. (1994). Optimal prediction with hierarchical models: Bayesian clustering. *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.). Chichester: Wiley, 67–76.

Bernardo, J. M. (1997a). Noninformative priors do not exist *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).

Bernardo, J. M. (1997b). Comment to 'Exponential and Bayesian conjugate families: review and extensions', by E. Gutiérrez-Peña and A. F. M. Smith. *Test* **6**, 70–71.

Bernardo, J. M. (1999). Nested Hipothesis testing: the Bayesian reference criterion. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 77–106 (with discussion).

Bernardo, J. M. and Bayarri, M. J. (1985). Bayesian model criticism. *Model Choice* (J.-P. Florens, M. Mouchart, J.-P. Raoult and L. Simar, eds.). Brussels: Pub. Fac. Univ. Saint Louis, 43–59.

Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.) (1992). *Bayesian Statistics 4*. Oxford: Oxford University Press.

Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.) (1996). *Bayesian Statistics 5*. Oxford: Oxford University Press.

Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.) (1999). *Bayesian Statistics 6*. Oxford: Oxford University Press. (to appear).

Bernardo, J. M. and Bermúdez, J. D. (1985). The choice of variables in probabilistic classification. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 67–81 (with discussion).

Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds.) (1980). *Bayesian Statistics*. Valencia: University Press.

Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds.) (1985). *Bayesian Statistics 2*. Amsterdam: North-Holland.

Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds.) (1988). *Bayesian Statistics 3*. Oxford: Oxford University Press.

Bernardo, J. M. and Girón F. J. (1988). A Bayesian analysis of simple mixture problems. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 67–88 (with discussion).

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician* **47**, 101–135.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Berry, D. A. (1996). *Statistics, a Bayesian Perspective*. Belmont, CA: Wasdsworth.

Berry, D. A. and Stangl, D. K. (eds.) (1994). *Bayesian Biostatistics*. New York: Marcel Dekker. (to appear).

Bertolino, F., Piccinato, L. and Racugno, W. (1995). Multiple Bayes factor for testing hypothesis. *J. Amer. Statist. Assoc.* **90**, 213–219.

Bickel, P. J. and Blackwell, D. (1967). A note on Bayes estimates. *Ann. Math. Statist.* **38**, 1907–1911.

Bickel, P. J. and Ghosh, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—a Bayesian argument. *Ann. Statist.* **18**, 1070–1090.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57**, 269–306.

Birnbaum, A. (1968). Likelihood. *Internat. Encyclopedia of the Social Sciences* **9**, 299-301.

Birnbaum, A. (1969). Concepts of statistical evidence. *Philosophy Science and Methods*. (S. Morgenbesso, P. Suppes and M. White eds.) New York: St. John's Press.

Birnbaum, A. (1972). More on concepts of statistical evidence. *J. Amer. Statist. Assoc.* **67**, 858–861.

Birnbaum, A. (1978). Likelihood. *International Encyclopedia of Statistics* (W. H. Kruskal and J. M. Tanur, eds.). London: Macmillan, 519–522.

Bjørnstad, J. F. (1990). Predictive likelihood: a review. *Statist. Sci.* **5**, 242–265 (with discussion).

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling. *J. Roy. Statist. Soc. A* **143**, 383–430 (with discussion).

Box, G. E. P. (1983). An apology for ecumenism in statistics. *Science* **151**, 15–84.

Box, G. E. P. (1985). *The Collected Works of G. E. P. Box* (G. C. Tiao, ed.). Pacific Drove, CA: Wadsworth.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. B* **26**, 211–252 (with discussion).

Box, G. E. P. and Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics* **9**, 57–71.

Box, G. E. P., Leonard, T. and Wu, C.-F. (eds.) (1983). *Scientific Inference, Data Analysis and Robustness*. New York: Academic Press.

Box, G. E. P. and Tiao, G. C. (1962). A further look at robustness via Bayes' theorem. *Biometrika* **49**, 419–432.

Box, G. E. P. and Tiao, G. C. (1964). A note on criterion robustness and inference robustness. *Biometrika* **51**, 169–173.

Box, G. E. P. and Tiao, G. C. (1965). Multiparameter problems from a Bayesian point of view. *Ann. Math. Statist.* **36**, 1468–1482.

Box, G. E. P. and Tiao, G. C. (1968a). A Bayesian approach to some outlier problems. *Biometrika* **55**, 119–129.

Box, G. E. P. and Tiao, G. C. (1968b). Bayesian estimation of means for the random effect model. *J. Amer. Statist. Assoc.* **63**, 174–181.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Bridgman, P. W. (1927). *The Logic of Modern Physics*. London: Macmillan.

Brillinger, D. R. (1962). Examples bearing on the definition of fiducial probability, with a bibliography. *Ann. Math. Statist.* **33**, 1349–1355.

Broemeling, L. D. (1985). *Bayesian Analysis of Linear Models*. New York: Marcel Dekker.

Brown, L. D. (1973). Estimation with incompletely specified loss functions. *J. Amer. Statist. Assoc.* **70**, 417–427.

Brown, L. D. (1985). *Foundations of Exponential Families*. Hayward, CA: IMS.

Brown, P. J., Le, N. D. and Zidek, J. V. (1994). Inference for a covariance matrix. *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.). Chichester: Wiley, 77–92.

Brown, R. V. (1993). Impersonal probability as an ideal assessment. *J. Risk and Uncertainty* **7**, 215–235.

Brunk, H. D. (1991). Fully coherent inference. *Ann. Statist.* **19**, 830–849.

Buehler, R. J. (1959). Some validity criteria for statistical inference. *Ann. Math. Statist.* **30**, 845–863.

Buehler, R. J. (1971). Measuring information and uncertainty. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 330–351 (with discussion).

Buehler, R. J. and Feddersen, A. P. (1963). Note on a conditional property of Student's t. *Ann. Math. Statist.* **34**, 1098–1100.

Butler, R. W. (1986). Predictive likelihood inference with applications. *J. Roy. Statist. Soc. B* **48**, 1–38 (with discussion).

Carlin, B. P. and Gelfand, A. E. (1991). An iterative Monte Carlo method for nonconjugate Bayesian analysis. *Statist. Computing* **1**, 119-128.

Casella, G. (1992). Conditional inference for confidence sets. *Current Issues in Statistical Inference: Essays in Honor of D. Basu*. (M. Ghosh and P. K. Pathak eds.). Hayward, CA: IMS.

Casella, G. (1996). Statistical inference and Monte Carlo algorithms. *Test* **5** 249–334, (with discussion).

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–135, (with discussion).

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Pacific Drove, CA: Wadsworth.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46**, 167–174.

Casella, G., Hwang, J. T. G. and Robert, C. P. (1993). A paradox in decision theoretic interval estimation. *Statistica Sinica* **3**, 141–155.

Chang, T. and Eaves, D. M. (1990). Reference priors for the orbit of a group model. *Ann. Statist.* **18**, 1595–1614.

Chang, T. and Villegas, C. (1986). On a theorem of Stein relating Bayesian and classical inferences in group models. *Canadian J. Statist.* **14**, 289–296.

Cifarelli, D. M. (1987). Recent contributions to Bayesian statistics. *Italian Contributions to the Methodology of Statistics* (A. Naddeo, ed.). Padova: Cleub, 483–516.

Cifarelli, D. M. and Muliere, P. (1989). *Statistica Bayesiana*. Pavia: G. Iuculano.

Cifarelli, D. M. and Regazzini, E. (1982). Some considerations about mathematical statistics teaching methodology suggested by the concept of exchangeability. *Exchangeability in Probability and Statistics*. (G. Koch and F. Spizzichino, eds.). Amsterdam: North-Holland, 185–205.

Cifarelli, D. M. and Regazzini, E. (1987). Priors for exponential families which maximize the association between past and future observations. *Probability and Bayesian Statistics* (R. Viertl, ed.). London: Plenum, 83–95.

Clarke, B. (1996). Implications of reference priors for prior infrmation and for sample size. *J. Amer. Statist. Assoc.* **91**, 173–184.

Clarke, B. and Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Planning and Inference* **41**, 37–60.

Clarke, B. and Sun, D. (1997). Reference priors under the chi-square distance. *Sankhyā A* **59**, 215–231.

Clarke, B. and Wasserman, L. (1993). Non-informative priors and nuisance parameters. *J. Amer. Statist. Assoc.* **88**, 1427–1432.

Consonni, G. and Veronese, P. (1987). Coherent distributions and Lindley's paradox. *Probability and Bayesian Statistics* (R. Viertl, ed.). London: Plenum, 111–120.

Consonni, G. and Veronese, P. (1989a). Some remarks on the use of improper priors for the analysis of exponential regression problems. *Biometrika* **76**, 101–106.

Consonni, G. and Veronese, P. (1989b). A note on coherent invariant distributions as non-informative priors for exponential and location scale families. *Comm. Statist. Theory and Methods* **18**, 2883-2887.

Consonni, G. and Veronese, P. (1992a). Bayes factors for linear models and improper priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 587–594.

Consonni, G. and Veronese, P. (1992b). Conjugate priors for exponential families having quadratic variance functions. *J. Amer. Statist. Assoc.* **87**, 1123–1127.

Cornfield, J. (1969). The Bayesian outlook and its applications. *Biometrics* **25**, 617–657.

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357–372.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D. R. (1988). Conditional and asymptotic inference. *Sankhyā A* **50**, 314–337.

Cox, D. R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5**, 169–174.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B* **49**, 1–39 (with discussion).

Cox, D. R. and Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* **79**, 408–411.

Cox R. T. (1946). Probability, frequency and expectation. *Amer. J. Physics* **14**, 1–13.

Cox R. T. (1961). *The Algebra of Probable Inference*. Baltimore: Johns Hopkins.

Creasy, M. A. (1959). Limits for the ratio of the means. *J. Roy. Statist. Soc. B* **16**, 186–189.

Crowder, M. J. (1988). Asymptotic expansions of posterior expectations, distributions and densities for stochastic processes. *Ann. Inst. Statist. Math.* **40**, 297–309.

Csiszár, I. (1985). An extended maximum entropy principle and a Bayesian justification. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 83–98, (with discussion).

Daboni, L. and Wedlin, A. (1982). *Statistica. Un'Introduzione all'Impostazione Neo-Bayesiana*. Torino: UTET.

DasGupta, A. (1991). Diameter and volume minimizing confidence sets in Bayes and classical problems. *Ann. Statist.* **19**, 1225–1243.

Datta, G. S. (1996). On priors providing frequentist validity for Bayesian inference of multiple parametric functions. *Biometrika* **83**, 287–298.

Datta, G. S. and Ghosh, J. K. (1995). On priors providing a frequentist validity for Bayesian inference. *Biometrika* **82**, 37–45.

Datta, G. S. and Ghosh, J. K. (1995c). Noninformative priors for maximal invariant parameter in group models. *Test* **4**, 95–114.

Datta, G. S. and Ghosh, M. (1995a). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.

Datta, G. S. and Ghosh, M. (1995b). Hierarchical Bayes estimators of the error variance in one-way ANOVA models. *J. Statist. Planning and Inference* **45**, 399–411.

Datta, G. S. and Ghosh, M. (1996). On the invariance of noninformative priors. *Ann. Statist.* **24**, 141–159.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–332.

Dawid, A. P. (1970). On the limiting normality of posterior distributions. *Proc. Camb. Phil. Soc.* **67**, 625–633.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 644–666.

Dawid, A. P. (1977). Invariant distributions and analysis of variance models. *Biometrika* **64**, 291–297.

Dawid, A. P. (1979a). Conditional independence in statistical theory. *J. Roy. Statist. Soc. B* **41**, 1–31, (with discussion).

Dawid, A. P. (1979b). Some misleading arguments involving conditional independence. *J. Roy. Statist. Soc. B* **41**, 249–252.

Dawid, A. P. (1980). A Bayesian look at nuisance parameters. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 167–203 (with discussion).

Dawid, A. P. (1983a). Statistical inference. *Encyclopedia of Statistical Sciences* **4** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 89–105.

Dawid, A. P. (1983b). Invariant prior distributions. *Encyclopedia of Statistical Sciences* **4** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 228–236.

Dawid, A. P. (1984). Statistical theory, the prequential approach. *J. Roy. Statist. Soc. A* **147**, 278–292.

Dawid, A. P. (1986b). A Bayesian view of statistical modelling. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Amsterdam: North-Holland. 391–404.

Dawid, A. P. (1988a). The infinite regress and its conjugate analysis. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 95–110 (with discussion).

Dawid, A. P. (1988b). Symmetry models and hypotheses for structured data layouts. *J. Roy. Statist. Soc. B* **50**, 1–34 (with discussion).

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *J. Roy. Statist. Soc. B* **53**, 79—109 (with discussion).

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 109–125 (with discussion).

Dawid, A. P. and Stone, M. (1972). Expectation consistency of inverse probability distributions. *Biometrika* **59**, 486–489.

Dawid, A. P. and Stone, M. (1973). Expectation consistency and generalised Bayes inference. *Ann. Statist.* **1**, 478–485.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189–233 (with discussion).

Dawid, A. P., Stone, M. and Zidek, J. V. (1980). Comment on Jaynes (1980). *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 79–87.

Dawid, A. P., Stone, M. and Zidek, J. V. (1996). Critique of E. T. Jaynes' 'Paradoxes of Probability Teory'. *Tech. Rep.* **172**, University College London, UK..

de Finetti, B. (1937/1964). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68. Reprinted in 1980 as 'Foresight; its logical laws, its subjective sources' in *Studies in Subjective Probability* (H. E. Kyburg and H. E Smokler, eds.). New York: Dover, 93–158.

de Finetti, B. (1962). Does it make sense to speak of 'Good Probability Appraisers'? *The Scientist Speculates: An Anthology of Partly-Baked Ideas* (I. J. Good, ed.). New York: Wiley, 257–364. Reprinted in 1972, *Probability, Induction and Statistics* New York: Wiley, 19–23.

de Finetti, B, (1974). Bayesianism: its unifying role for both the foundations and applications of statistics *Internat. Statist. Rev.* **42**, 117–130. Reprinted in de Finetti (1993), 205–228 (in italian) and 467–490 (in english) .

de la Horra, J. (1986). Convergencia del vector de probabilidad a posterior bajo una distribución predictiva. *Trab. Estadist.* **1**, 3–11.

de la Horra, J. (1987). Generalized estimators: a Bayesian decision theoretic view. *Statistics and Decisions* **5**, 347–352.

de la Horra, J. (1988). Parametric estimation with L_1 distance. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 579–583.

de la Horra, J. (1992). Using the prior mean of a nuisance parameter. *Test* **1**, 31–38.

de la Horra, J. and Fernández, C. (1994). Bayesian robustness of credible regions in the presence of nuisance parameters. *Comm. Statist. Theory and Methods* **23**, (to appear).

DeGroot, M. H. (1962). Uncertainty, information and sequential experiments. *Ann. Math. Statist.* **33**, 404–419.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

DeGroot, M. H. (1973). Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. *J. Amer. Statist. Assoc.* **68**, 966–969.

DeGroot, M. H. (1980). Improving predictive distributions. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 385–395 and 415–429 (with discussion).

DeGroot, M. H. (1986). Changes in utility and information. *Recent Developments in the Foundations of Utility and Risk Theory* (L. Daboni *et al.* . eds.). Dordrecht: Reidel.

DeGroot, M. H. (1987). *Probability and Statistics*. Reading, MA: Addison-Wesley.

DeGroot, M. H. and Rao, M. M. (1963). Bayes estimation with convex loss. *Ann. Math. Statist.* **34**, 839–846.

DeGroot, M. H. and Rao, M. M. (1966). Multidimensional information inequalities and prediction. *Multivariate Statistics* (P. R. Krishnaiah, ed.). New York: Academic Press, 287–313.

Delampady, M. (1989). Lower bounds on Bayes factors for interval null hypotheses. *J. Amer. Statist. Assoc.* **84**,120–124.

Delampady, M. and Berger, J. O. (1990). Lower bounds on Bayes factors for multinomial and chi-squared tests of fit. *Ann. Statist.* **18**, 1295–1316.

Dempster, A. P. (1968). A generalization of Bayesian inference. *J. Roy. Statist. Soc. B* **30**, 205–247 (with discussion).

Dempster, A. P. (1975). A subjective look at robustness. *Internat. Statist. Rev.* **46**, 349–374.

Dempster, A. P. (1985). Probability, evidence and judgement. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 119–132 (with discussion).

DeRobertis, L. and Hartigan, J. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **9**, 235–244.

de Vos, A. F. (1993). A fair comparison between regression models of different dimensions. *Tech. Rep.*, The Free University, Amsterdam, Holland..

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Berlin: Springer.

De Waal, D. J. and Groenewald, P. C. N. (1989). On measuring the amount of information from the data in a Bayesian analysis. *South African Statist. J.* **23**, 23–61 (with discussion).

Diaconis, P. (1988). Recent progress on de Finetti's notion of exchangeability. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 111-125 (with discussion).

Diaconis, P. and Freedman, D. (1986a). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–67, (with discussion).

Diaconis, P. and Freedman, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* **14**, 68–87.

DiCiccio, T. J., Field, C. A. and Frase, D. A. S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77**, 77–95.

DiCiccio, T. J. and Stern, S. E. (1994). Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihood. *J. Roy. Statist. Soc. B* **56**, 397–408.

Dickey, J. M. (1968). Three multidimensional integral identities with Bayesian applications. *Ann. Math. Statist.* **39**, 1615–1627.

Dickey, J. M. (1969). Smoothing by cheating. *Ann. Math. Statist.* **40**, 1477–1482.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameteters. *Ann. Math. Statist.* **42**, 204–223.

Dickey, J. M. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *J. Roy. Statist. Soc. B* **35**, 285–305. Reprinted in 1974 in *Studies in Bayesian Econometrics and Statistics: in Honor of Leonard J. Savage* (S. E. Fienberg and A. Zellner, eds.). Amsterdam: North-Holland, 485–511.

Dickey, J. M. (1974). Bayesian alternatives to the F test and least-squares estimate in normal linear model. *Studies in Bayesian Econometrics and Statistics: in Honor of Leonard J. Savage* (S. E. Fienberg and A. Zellner, eds.). Amsterdam: North-Holland, 515–554.

Dickey, J. M. (1976). Approximate posterior distributions. *J. Amer. Statist. Assoc.* **71**, 680–689.

Dickey, J. M. (1977). Is the tail area useful as an approximate Bayes factor? *J. Amer. Statist. Assoc.* **72**, 138–142.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. B* **56**, 363–375.

du Plessis, J. L., van der Merwe, A. J. and Groenewald, P. C. N. (1995). Reference priors for the multivariate calibration problem. *South African Statist. J.* **29**, 155-168.

Dudley, R. M. and Haughton, D. (1997). Information criteria for multiple data sets and restricted parameters. *Statistica Sinica* **7**, 265–284.

Dunsmore, I. R. (1966). A Bayesian approach to classification. *J. Roy. Statist. Soc. B* **28**, 568–577.

Dunsmore, I. R. (1968). A Bayesian approach to calibration. *J. Roy. Statist. Soc. B* **30**, 396–405.

Dunsmore, I. R. (1969). Regulation and optimization. *J. Roy. Statist. Soc. B* **31**, 160–170.

Durbin, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *J. Amer. Statist. Assoc.* **65**, 395–398.

Dupuis, J. and Robert, C. P. (1997). Model choice in qualitative regression models. *Tech. Rep.* **9717**, CREST-INSEE, France.

Eaton, M. L. (1982). A method for evaluating improper prior distributions. *Statistical Decision Theory and Related Topics III* **1** (S. S. Gupta and J. O. Berger, eds.). New York: Academic Press,

Eaton, M. L. (1992). A statistical diptych: admissible inferences, recurrence of symmetric Markov chains. *Ann. Statist.* **20**, 1147–1179.

Eaves, D. M. (1983a). On Bayesian nonlinear regression with an enzyme example. *Biometrika* **70**, 373-379.

Eaves, D. M. (1983b). Minimally informative prior analysis of a non-linear model. *The Statistician* **32**, 117.

Eaves, D. M. (1985). On maximizing the missing information about a hypothesis. *J. Roy. Statist. Soc. B* **47**, 263-266.

Edwards, A. W. F. (1972/1992). *Likelihood*. Cambridge: University Press. Second edition in 1992. Baltimore: John Hopkins University Press.

Edwards, A. W. F. (1974). The history of likelihood. *Internat. Statist. Rev.* **42**, 9–15.

Edwards, W. L., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242. Reprinted in *Robustness of Bayesian Analysis* (J. B. Kadane, ed.). Amsterdam: North-Holland, 1984, 1–62. Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.). Brookfield, VT: Edward Elgar, 1995, 140–189.

Efron, B. (1973). In discussion of Dawid, Stone and Zidek (1973). *J. Roy. Statist. Soc. B* **35**, 219.

Efron, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40**, 1–11 (with discussion).

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26.

Efron, B. and Morris, C. N. (1972). Empirical Bayes estimators on vector observations—an extension of Stein's method. *Biometrika* **59**, 335–347.

Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalisations. *J. Amer. Statist. Assoc.* **70**, 311–319.

Efstathiou, M. (1996). Some Aspects of Approximation and Computation for Bayesian Inference. Ph.D. Thesis, Imperial College London, UK.

Erickson, G. J. and Smith, C. R. (eds.) (1988). *Maximum Entropy and Bayesian Methods in Science and Engineering*. (2 volumes). Dordrecht: Kluwer.

Farrell, R. H. (1964). Estimators of a location parameter in the absolutely continuous case. *Ann. Math. Statist.* **35**, 949–998.

Farrell, R. H. (1968). Towards a theory of generalized Bayes tests. *Ann. Math. Statist.* **39**, 1–22.

Ferrándiz, J. R. (1982). Una solución Bayesiana a la paradoja de Stein. *Trab. Estadist.* **33**, 31–46.

Ferrándiz, J. R. (1985). Bayesian inference on Mahalanobis distance: an alternative approach to Bayesian model testing. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 645–654.

Ferrándiz, J. R. and Sendra, M. (1982). *Tablas de Bioestadística*. Valencia: University Press.

Fieller, E. C. (1954). Some problems in interval estimation. *J. Roy. Statist. Soc. B* **16**, 186–194 (with discussion).

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507–521.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London A* **222**, 309–368. Reprinted in *Breakthroughs in Statistics* **1** (S. Kotz and N. L. Johnson, eds.). Berlin: Springer, 1991, 11–44.

Fisher, R. A. (1925). Theory of statistical information. *Proc. Camb. Phil. Soc.* **22**, 700–725.

Fisher, R. A. (1930). Inverse probability. *Proc. Camb. Phil. Soc.* **26**, 528–535.

Fisher, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proc. Roy. Soc. A* **139**, 343–348.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Ann. Eugenics* **6**, 391–398.

Fisher, R. A. (1939). A note on fiducial inference. *Ann. Statist.* **10**, 383–388.

Fisher, R. A. (1956/1973). *Statistical Methods and Scientific Inference*. Third edition in 1973. Edinburgh: Oliver and Boyd. Reprinted in 1990 whithin *Statistical Methods, Experimental Design, and Scientific Inference* (J. H. Bennet, ed.). Oxford: Oxford University Press.

Florens, J.-P. (1978). Mesures à priori et invariance dans une expérience Bayésienne. *Pub. Inst. Statist. Univ. Paris* **23**, 29–55.

Florens, J.-P. (1982). Expériences Bayésiennes invariantes. *Ann. Inst. M. Poincaré* **18**, 309–317.

Florens, J.-P. and Mouchart, M. (1985). Model selection: some remarks from a Bayesian viewpoint. *Model Choice* (Florens, J.-P., Mouchart, M., Raoult J.-P. and Simar, L., eds.). Brussels: Pub. Fac. Univ. Saint Louis, 27–44.

Florens, J.-P. and Mouchart, M. (1993). Bayesian testing and testing Bayesians. *Handbook of Statistics*, (G. S. Maddala and C. R. Rao, eds.), Amsterdam: North-Holland, Ch. 11.

Florens, J.-P., Mouchart, M., Raoult J.-P. and Simar, L. (eds.) (1985). *Model Choice*. Brussels: Pub. Fac. Univ. Saint Louis.

Fougère, P. T. (ed.) (1990). *Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer.

Fraser, D. A. S. (1961). On fiducial inference. *Ann. Math. Statist.* **32**, 661–676.

Fraser, D. A. S. (1963). On the sufficiency and likelihood principles. *J. Amer. Statist. Assoc.* **58**, 641–647.

Fraser, D. A. S. (1968). *The Structure of Inference*. New York: Wiley.

Fraser, D. A. S. (1972). Bayes, likelihood or structural. *Ann. Math. Statist.* **43**, 777–790.

Fraser, D. A. S. (1972). Events, information processing and the structured model. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 32–55 (with discussion).

Fraser, D. A. S. (1974). Comparison of inference philosophies. *Information, Inference and decision* (G. Menges, ed.). Dordrecht: Reidel, 77–98.

Fraser, D. A. S. (1979). *Inference and Linear Models*. New York: McGraw-Hill.

Fraser, D. A. S. and McDunnough, P. (1989). Further remarks on the asymptotic normality of likelihood and conditional analysis. *Canadian J. Statist.* **12**, 183–190.

Fraser, D. A. S. and Reid, N. (1989). Adjustments to profile likelihood. *Biometrika* **76**, 477–488.

Fraser, D. A. S., Monette, G., and Ng, K. W. (1985). Marginalization, likelihood and structural models, *Multivariate Analysis* **6** (P. R. Krishnaiah, ed.). Amsterdam: North-Holland, 209–217.

Freedman, D. A. (1963b). On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386–1403.

Freedman, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *Ann. Math. Statist.* **36**, 454–456.

Freedman, D. A. and Diaconis, P. (1983). On inconsistent Bayes estimates in the discrete case. *Ann. Statist.* **11**, 1109–1118.

Freeman, P. R. and Smith, A. F. M. (eds.) (1994). *Aspects of Uncertainty: a Tribute to D. V. Lindley*. Chichester: Wiley.

Fu, J. C. and Kass, R. E. (1988). The exponential rate of convergence of posterior distributions. *Ann. Inst. Statist. Math.* **40**, 683–691.

Gamerman, D. and Migon, H. S. (1993). Dynamic hierarchical models. *J. Roy. Statist. Soc. B* **55**, 629–642.

Gatsonis, C. A. (1984). Deriving posterior distributions for a location parameter: a decision-theoretic approach. *Ann. Statist.* **12**, 958–970.

Gatsonis, C. A., Hodges, J. S., Kass, R. E. and Singpurwalla, N. (eds.) (1993). *Case Studies in Bayesian Statistics*. Berlin: Springer.

Gatsonis, C. A., Hodges, J. S., Kass, R. E. and Singpurwalla, N. (eds.) (1995). *Case Studies in Bayesian Statistics II*. Berlin: Springer.

Gatsonis, C. A., Hodges, J. S., Kass, R. E. and Singpurwalla, N. (eds.) (1997). *Case Studies in Bayesian Statistics III*. Berlin: Springer.

Geisser, S. (1964). Posterior odds for multivariate normal classification. *J. Roy. Statist. Soc. B* **26**, 69–76.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.* **36**, 150–159.

Geisser, S. (1966). Predictive discrimination. *Multivariate Analysis* (P. R. Krishnaiah, ed.). New York: Academic Press, 149–163.

Geisser, S. (1971). The inferential use of predictive distributions. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 456–469.

Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika* **61**, 101–107.

Geisser, S. (1975). The predictive sample reuse method, with applications. *J. Amer. Statist. Assoc.* **70**, 320–328.

Geisser, S. (1977). Comment to Wilkinson (1977). *J. Roy. Statist. Soc. B* **39**, 155–156.

Geisser, S. (1979). In discussion of Bernardo (1979b). *J. Roy. Statist. Soc. B* **41**, 136–137.

Geisser, S. (1980a). The contributions of Sir Harold Jeffreys to Bayesian inference. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 13–20.

Geisser, S. (1980b). A predictivist primer. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 363–381.

Geisser, S. (1982). Bayesian discrimination. *Handbook of Statistics 2. Classification* (P. R. Krishnaiah and L. N. Kanal eds.). Amsterdam: North-Holland, 101–120.

Geisser, S. (1984). On prior distributions for binary trials. *J. Amer. Statist. Assoc.* **38**, 244–251 (with discussion).

Geisser, S. (1985). On the prediction of observables: a selective update. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 203–230 (with discussion).

Geisser, S. (1986). Predictive analysis. *Encyclopedia of Statistical Sciences* **7** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 158–170.

Geisser, S. (1987). Influential observations, diagnostics and discordancy tests. *Appl. Statist.* **14**, 133–142.

Geisser, S. (1988). The future of statistics in retrospect. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 147–158 (with discussion).

Geisser, S. (1992). Bayesian perturbation diagnostics and robustness. *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.). Berlin: Springer, 289–302.

Geisser, S. (1993). *Predictive Inference: an Introduction*. London: Chapman and Hall.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. B* **25**, 368–376.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74**, 153–160.

Geisser, S., Hodges, J. S., Press, S. J. and Zellner, A. (eds.) (1990). *Bayesian and Likelihood methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*. Amsterdam: North-Holland.

Gelfand, A. E. and Desu, A. (1968). Predictive zero-mean uniform discrimination. *Biometrika* **55**, 519–524.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. B* **56**, 501–514.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 147–167 (with discussion).

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal models using Gibbs sampling. *J. Amer. Statist. Assoc.* **85**, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

Gelfand, A. E., Smith, A. F. M. and Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.* **87**, 523–532.

Gelman, A., Carlin, J. B., Stern, H. and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall

George, E. I., Makov, U. E. and Smith, A. F. M. (1993). Conjugate likelihood distributions. *Scandinavian J. Statist.* **20**, 147–156.

George, E. I. and McCulloch, R. (1993a). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

George, E. I. and McCulloch, R. (1993b). On obtaining invariant prior distributions. *J. Statist. Planning and Inference* **37**, 169–179.

Ghosal, S. (1996). Reference priors in multiparameter nonregular cases. *Test* **5**, 159–186.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1996). Non-Informative priors via sieves and packing numbers. *Advances in Decision Theory and Applications.* (S. Panchpakesan and N. Balakrishnan, eds.) Boston: Birkhouser, 119–132.

Ghosh, J. K. (1994). *Higher Order Asymptotics*. Hayward, CA: IMS.

Ghosh, J. K., Ghosal, S. and Samanta, T. (1994). Stability and convergence of the posterior in non-regular problems. *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer,

Ghosh, J. K. and Mukerjee, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multiparameter case. *J. Multivariate Analysis* **38**, 385–393.

Ghosh, J. K. and Mukerjee, R. (1992a). Non-informative priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 195–210 (with discussion).

Ghosh, J. K. and Mukerjee, R. (1992b). Bayesian and frequentist Bartlett corrections for likelihood ratio tests. *J. Roy. Statist. Soc. B* **54**, 867–875.

Ghosh, J. K. and Mukerjee, R. (1993). On priors that match posterior and frequentist distribution functions. *Canadian J. Statist.* **21**, 89–96.

Ghosh, M. (1991). Hierarchical and empirical Bayes sequential estimation. *Handbook of Statistics* **8**. *Statistical Methods in Biological and Medical Sciences* (C. R. Rao and R. Chakraborty, eds.). Amsterdam: North-Holland, 441-458.

Ghosh, M. (1992a). Hierarchical and empirical Bayes multivariate estimation. *Current Issues in Statistical Inference: Essays in Honor of D. Basu*. (M. Ghosh and P. K. Pathak eds.). Hayward, CA: IMS, 151–177.

Ghosh, M. (1992b). Constrained Bayes estimation with application. *J. Amer. Statist. Assoc.* **87**, 533–540.

Ghosh, M., Carlin, B. P. and Srivastava, M. S. (1995). Probability matching priors for linear calibration. *Test* **4**, 333–357.

Ghosh, M. and Pathak, P. K. (eds.) (1992). *Current Issues in Statistical Inference: Essays in Honor of D. Basu*. Hayward, CA: IMS.

Ghosh, M. and Yang, M.-Ch. (1996). Non-informative priors for the two sample normal problem. *Test* **5**, 145–157.

Gilio, A. and Scozzafava, R. (1985). Vague distributions in Bayesian testing of a null hypothesis. *Metron* **43**, 167–174.

Girón, F. J., Martínez, M. L. and Imlahi, L. (1998). A characterization of the Behrens-Fisher distribution with applications to Bayesian inference. *C. R. Acad. Sci. Paris* , (to appear).

Gleser, L. J. and Hwang, J. T. (1987). The non-existence of $100(1-\alpha)\%$ confidence sets of finite expected diameters in error-in-variable and related models. *Ann. Statist.* **15**, 1351–1362.

Godambe, V. P. and Sprott, D. A. (eds.) (1971). *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Goel, P. K. (1983). Information measures and Bayesian hierarchical models. *J. Amer. Statist. Assoc.* **78**, 408–410.

Goel, P. K. and DeGroot, M. H. (1979). Comparison of experiments and information measures. *Ann. Statist.* **7**, 1066–1077.

Goel, P. K. and DeGroot, M. H. (1980). Only normal distributions have linear posterior expectations in linear regression. *J. Amer. Statist. Assoc.* **75**, 895–900.

Goel, P. K. and DeGroot, M. H. (1981). Information about hyperparameters in hierarchical models. *J. Amer. Statist. Assoc.* **76**, 140–147.

Goel, P. K. and Iyengar, N. S. (eds.) (1992). *Bayesian Analysis in Statistics and Econometrics*. Berlin: Springer

Goel, P. K. and Zellner, A. (eds.) (1986). *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland.

Goldstein, M. (1981). Revising previsions: a geometric interpretation. *J. Roy. Statist. Soc. B* **43**, 105–130.

Goldstein, M. (1985). Temporal coherence. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 231–248 (with discussion).

Goldstein, M. (1986c). Prevision. *Encyclopedia of Statistical Sciences* **7** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 175–176.

Goldstein, M. (1991). Belief transforms and the comparison of hypothesis. *Ann. Statist.* **19**, 2067–2089.

Goldstein, M. and Smith, A. F. M. (1974). Ridge-type estimators for regression analysis. *J. Roy. Statist. Soc. B* **36**, 284–319.

Gómez, E. and Gómez-Villegas, M. A. (1990). Three methods for constructing reference distributions. *Rev. Mat. Univ. Complutense de Madrid* **3**, 153–162.

Gomez,E. Gomez-Villegas, M. A. and Marin, J .M. (1998). A multivariate generalization of the ower exponential family of distributions. *Comm. Statist. Theory and Methods* **27**, 589–600.

Gómez-Villegas, M. A. and Gómez, E. (1992). Bayes factors in testing precise hypotheses. *Comm. Statist. A* **21**, 1707–1715.

Gómez-Villegas, M. A. and Maín, P. (1992). The influence of prior and likelihood tail behaviour on the posterior distribution. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 661–667.

Gómez-Villegas, M. A. and Gómez, E. (1992). Bayes factor in testing precise hypothesis. *Comm. Statist. Theory and Methods* **21**, 1707–1715.

Gómez-Villegas, M. A. and Sanz, L. (1998). Reconciling Bayesian and frequentist evidence in the point null testing problem. *Test* **7**, 207–216.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. London : Griffin; New York: Hafner Press.

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. B* **14**, 107–114.

Good, I. J. (1959). Kinds of probability. *Science* **127**, 443–447.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power and the utility of experiments. *J. Roy. Statist. Soc. B* **22**, 319–331.

Good, I. J. (1962). Subjective probability on the measure of a non-measurable set. *Logic Methodology and Philosophy of Science* (E. Nagel, P. Suppes and A. Tarski, eds.). Stanford: University Press, 319–329.

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34**, 911–934.

Good, I. J. (1965). *The Estimation of Probabilities. An Essay on Modern Bayesian Methods*. Cambridge, Mass: The MIT Press.

Good, I. J. (1966). A derivation of the probabilistic explanation of information. *J. Roy. Statist. Soc. B* **28**, 578–581.

Good, I. J. (1967). A Bayesian test for multinomial distributions. *J. Roy. Statist. Soc. B* **29**, 399–431.

Good, I. J. (1968). Utility of a distribution. *Nature* **219**, 1392.

Good, I. J. (1969). What is the use of a distribution? *Multivariate Analysis* **2** (P. R. Krishnaiah, ed.). New York: Academic Press, 183–203.

Good, I. J. (1971). The probabilistic explication of information, evidence, surprise, causality, explanation and utility. Twenty seven principles of rationality. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 108–141 (with discussion).

Good, I. J. (1976). The Bayesian influence, or how to sweep subjectivism under the carpet. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* **2** (W. L. Harper and C. A. Hooker eds.). Dordrecht: Reidel, 119–168.

Good, I. J. (1980a). The contributions of Jeffreys to Bayesian statistics. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 21–34.

Good, I. J. (1980b). Some history of the hierarchical Bayesian mehodology. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 489–519, (with discussion).

Good, I. J. (1982). Degrees of belief. *Encyclopedia of Statistical Sciences* **2** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 287–292.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*. Minneapolis: Univ. Minnesota Press.

Good, I. J. (1985). Weight of Evidence: a brief survey. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 249–270 (with discussion).

Good, I. J. (1987). Hierarchical Bayesian and empirical Bayesian methods. *Amer. Statist.* **41**, (with discussion).

Good, I. J. (1988a). Statistical evidence. *Encyclopedia of Statistical Sciences* **8** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 651–656.

Good, I. J. (1988b). The interface between statistics and philosophy of science. *Statist. Sci.* **3**, 386–398 (with discussion).

Good, I. J. (1992). The Bayes/non-Bayes compromise: a brief review. *J. Amer. Statist. Assoc.* **87**, 597–606.

Good, I. J. and Crook, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711-720.

Good, I. J. and Gaskins, R. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58**, 255–277.

Goutis, C. and Casella, G. (1991). Improved invariant confidence intervals for a normal variance. *Ann. Statist.* **19**, 2019–2031.

Goutis, C. and Robert, C. P. (1997). Selection between hypotheses using estimation criteria. *Ann. Econom. Stat.* **46**, 1–22.

Goutis, C. and Robert, C. P. (1998). Model choice in generalized linear models: a Bayesian approach via Kullback–Leibler projections. *Biometrika* **85**, 29–37.

Grandy, W. T. and Schick, L. H. (eds.) (1991). *Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer.

Grundy, P. M. (1956). Fiducial distributions and prior distributions: an example in which the former cannot be associated with the later. *J. Roy. Statist. Soc. B* **18**, 217–221.

Gu, C. (1992). Penalized likelihood regression: a Bayesian analysis. *Statistica Sinica* **2**, 255-264.

Gûnel, E, and Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61**, 545–557.

Gupta, S. S. and Berger, J. O. (eds.) (1988). *Statistical Decision Theory and Related Topics IV* **1**. Berlin: Springer.

Gupta, S. S. and Berger, J. O. (eds.) (1994). *Statistical Decision Theory and Related Topics V*. Berlin: Springer. (to appear).

Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 669–674.

Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian*. London: Griffin.

Guttman, I. and Peña, D. (1988). Outliers and influence. Evaluation by posteriors of parameters in the linear model. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 631–640.

Guttman, I. and Peña, D. (1993). A Bayesian look at the question of diagnostics. *Statistica Sinica* **3**, 367–390.

Haldane, J. B. S. (1931). A note on inverse probability. *Proc. Camb. Phil. Soc.* **28**, 55–61.

Haldane, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika* **35**, 297–303.

Harrison P. J. and West, M. (1987). Practical Bayesian forecasting. *The Statistician* **36**, 115–125.

Hartigan, J. A. (1964). Invariant prior distributions. *Ann. Math. Statist.* **35**, 836–845.

Hartigan, J. A. (1965). The asymptotically unbiased prior distribution. *Ann. Math. Statist.* **36**, 1137–1152.

Hartigan, J. A. (1966a). Estimation by ranking parameters. *J. Roy. Statist. Soc. B* **28**, 32–44.

Hartigan, J. A. (1966b). Note on the confidence prior of Welch and Peers. *J. Roy. Statist. Soc. B* **28**, 55-56.

Hartigan, J. A. (1967). The likelihood and invariance principles. *J. Roy. Statist. Soc. B* **29**, 533–539.

Hartigan, J. A. (1969). Use of subsample values as typical values. *J. Amer. Statist. Assoc.* **104**, 1003–1317.

Hartigan, J. A. (1971). Similarity and probability. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 305–313 (with discussion).

Hartigan, J. A. (1975). Necessary and sufficient conditions for asymptotic normality of a statistic and its subsample values. *Ann. Statist.* **3**, 573–580.

Hartigan, J. A. (1983). *Bayes Theory*. Berlin: Springer.

Hartigan, J. A. (1996). Locally uniform prior distributions. *Ann. Statist.* **24**, 160-173.

Heath, D. L. and Sudderth, W. D. (1978). On finitely additive priors, coherence and extended admissibility. *Ann. Statist.* **6**, 333–345.

Heath, D. L. and Sudderth, W. D. (1989). Coherent inference from improper priors and from finitely additive priors. *Ann. Statist.* **17**, 907–919.

Heyde, C. C. and Johnstone, I. M. (1979). On asymptotic posterior normality for stochastic processes. *J. Roy. Statist. Soc. B* **41**, 184–189.

Hill, B. M. (1965). Inference about variance components in the one-way model. *J. Appl. Statist.* **60**, 806–825.

Hill, B. M. (1977). Exact and approximate Bayesian solutions for inference about variance components and multivariate inadmissibility. *New Developments in the Applications of Bayesian Methods* (A. Aykaç and C. Brumat, eds.). Amsterdam: North-Holland, 29–152.

Hill, B. M. (1968). Posterior distributions of percentiles: Bayes' theorem for sampling from a finite population. *J. Amer. Statist. Assoc.* **63**, 677–691.

Hill, B. M. (1969). Foundations of the theory of least squares. *J. Roy. Statist. Soc. B* **31**, 89–97.

Hill, B. M. (1974). On coherence, inadmissibility and inference about many parameters in the theory of least squares. *Studies in Bayesian Econometrics and Statistics: in Honor of Leonard J. Savage* (S. E. Fienberg and A. Zellner, eds.). Amsterdam: North-Holland, 555–584.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163–1174.

Hill, B. M. (1980). On finite additivity, non-conglomerability, and statistical paradoxes. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 39–66 (with discussion).

Hill, B. M. (1986). Some subjective Bayesian considerations in the selection of models. *Econometric Reviews* **4**, 191–288.

Hill, B. M. (1987). The validity of the likelihood principle. *Amer. Statist.* **41**, 95–100.

Hill, B. M. (1990). A theory of Bayesian data analysis. *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, eds.). Amsterdam: North-Holland, 49–73.

Hills, S. E. (1987). Reference priors and identifiability problems in non-linear models. *The Statistician* **36**, 235–240.

Hills, S. E. and Smith, A. F. M. (1992). Parametrization issues in Bayesian inference. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 227–246 (with discussion).

Hills, S. E. and Smith, A. F. M. (1993). Diagnostic plots for improved parametrisation in Bayesian inference. *Biometrika* **80**, 61–74.

Hinkley, D. V. (1979). Predictive likelihood. *Ann. Statist.* **7**, 718–728.

Hipp, C. (1974). Sufficient statistics and exponential families. *Ann. Statist.* **2**, 1283–1292.

Hoadley, B. (1970). A Bayesian look at inverse regression. *J. Amer. Statist. Assoc.* **65**, 356–369.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91**, 1461–1473.

Hodges, J. S. (1990). Can/may Bayesians use pure tests of significance? *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, eds.). Amsterdam: North-Holland, 75–90.

Hodges, J. S. (1992). Who knows what alternative lurks in the hearts of significance tests? *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 247–266 (with discussion).

Hwang, J. T. (1985). Universal domination and stochastic domination: decision theory under a broad class of loss functions. *Ann. Statist.* **13**, 295–314.

Hwang, J. T. (1988). Stochastic and universal domination. *Encyclopedia of Statistical Sciences* **8** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 781–784.

Hwang, J. T., Casella, G, Robert, C. Wells, M. and Farrell, R. (1992). Estimation of accuracy in testing. *Ann. Statist.* **20**, 490–509.

Ibragimov, I. A. and Hasminski, R. Z. (1973). On the information in a sample about a parameter. *Proc. 2nd Internat. Symp. Information Theory*. (B. N. Petrov and F. Csaki, eds.), Budapest: Akademiaikiadó, 295–309.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys' prior. *J. Amer. Statist. Assoc.* **86**, 981–986.

Irony, T. Z. (1992). Bayesian estimation for discrete distributions. *J. Appl. Statist.* **19**, 533–549.

Isaacs, G. L., Christ, D. E., Novick, M. R. and Jackson, P. H. (1974). *Tables for Bayesian Statisticians*. Ames, IO: Iowa University Press.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Trans. Systems, Science and Cybernetics* **4**, 227–291.

Jaynes, E. T. (1971). The well posed problem. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 342–356 (with discussion).

Jaynes, E. T. (1976). Confidence intervals vs. Bayesian intervals. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* **2** (W. L. Harper and C. A. Hooker eds.). Dordrecht: Reidel, 175–257 (with discussion).

Jaynes, E. T. (1980a). Marginalization and prior probabilities. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 43–87 (with discussion).

Jaynes, E. T. (1980). Discussion to the session on hypothesis testing. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 618–629. Reprinted in *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. (R. D. Rosenkranz, ed.). Dordrecht: Kluwer(1983), 378–400.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proc. of the IEEE* **70**, 939–952.

Jaynes, E. T. (1983). *Papers on Probability, Statistics and Statistical Physics*. (R. D. Rosenkrantz, ed.). Dordrecht: Kluwer.

Jaynes, E. T. (1985). Highly informative priors. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 329–359 (with discussion).

Jaynes, E. T. (1994). *Probability Theory: The Logic of Science*. Posted in the Web at <http://ftp://bayes.wustl.edu/pub/Jaynes/book.probability.theory/>.

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *J. Scientific Exploration* **4**, 153–169.

Jefferys, W. H and Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *Amer. Scientist.* **80**, 64–82.

Jeffreys, H. (1931/1973). *Scientific Inference*. Cambridge: University Press. Third edition in 1973, Cambridge: University Press.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.* **31**, 203–222.

Jeffreys, H. (1939/1961). *Theory of Probability*. Oxford: Oxford University Press. Third edition in 1961, Oxford: Oxford University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A* **186**, 453–461.

Jeffreys, H. (1955). The present position in probability theory. *Brit. J. Philos. Sci.* **5**, 275–289.

Jeffreys, H. (1980). Some general points in probability theory. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 451–453.

Jeffreys, H. and Jeffreys, B. S. (1946/1972). *Methods of Mathematical Physics*. Cambridge: University Press. Third edition in 1972, Cambridge: University Press.

Johnson, N. L. and Kotz, S. (1969). *Discrete Distributions*. New York: Wiley.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1970/1995). *Continuous Univariate Distributions (2 vols.),* (2nd. ed.) New York: Wiley.

Johnson, N. L. and Kotz, S. (1972). *Continuous Multivariate Distributions*. New York: Wiley.

Johnson, R. A. (1967). An asymptotic expansion for posterior distributions. *Ann. Math. Statist.* **38**, 1899–1906.

Johnson, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41**, 851–864.

Johnson, R. A. and Ladalla, J. N. (1979). The large-sample behaviour of posterior distributions with sampling from muitiparameter exponential family models and allied results. *Sankhyā B* **41**, 169–215.

Joshi, V. M. (1983). Likelihood principle. *Encyclopedia of Statistical Sciences* **4** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 644–647.

Justice, J. M. (ed.) (1987). *Maximum Entropy and Bayesian Methods in Applied Statistics*. Cambridge: University Press.

Kadane, J. B. and Dickey, J. M. (1980). Bayesian decision theory and the simplification of models. *Evaluation of Econometric Methods* (J. Kmenta and J. Ramsey, eds.), New York: Academic Press, 245–268.

Kadane, J. B. and O'Hagan, A. (1995). Using finitely additive probability: uniform distributions on the natural numbers. *J. Amer. Statist. Assoc.* **95**, 626–631.

Kadane, J. B., Schervish, M. J. and Seidenfeld, T. (1986). Statistical implications of finitely additive probability. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Amsterdam: North-Holland, 59–76.

Kadane, J. B. and Seidenfeld, T. (1990). Randomization in a Bayesian perspective. *J. Statist. Planning and Inference* **25**, 329–345.

Kalbfleish, J. G. (1971). Likelihood methods in prediction. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 372–392 (with discussion).

Kalbfleish, J. G. and Sprott, D. A. (1970). Application of likelihood methods to models involving large number of parameters. *J. Roy. Statist. Soc. B* **32**, 175–208 (with discussion).

Kalbfleish, J. G. and Sprott, D. A. (1973). Marginal and conditional likelihoods. *Sankhyā A* **35**, 311–328.

Kappenman, R. F., Geisser, S. and Antle, C. E. (1970). Bayesian and fiducial solutions to the Fieller-Creasy problem. *Sankhyā B* **32**, 331–340.

Kashyap, R. L. (1971). Prior probability and uncertainty. *IEEE Trans. Information Theory* **14**, 641–650.

Kashyap, R. L. (1974). Minimax estimation with divergence loss function. *Information Sciences* **7**, 341–364.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statist. Sci.* **4**, 188–234.

Kass, R. E. (1990). Data-translated likelihood and Jeffreys' rule. *Biometrika* **77**, 107–114.

Kass, R., Carlin, B., Carriquiry, A., Catsonis, C., Gelman, A., Verdinelli, I. and West, M. (eds.) (1999). *Case Studies in Bayesian Statistics IV*. Berlin: Springer.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Kass, R. E. and Slate E. H. (1992). Reparametrization and diagnostics of posterior non-normality. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 289–305 (with discussion).

Kass, R. E., Tierney, L. and Kadane, J. B. (1988). Asymptotics in Bayesian computation. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 261–278, (with discussion).

Kass, R. E., Tierney, L. and Kadane, J. B. (1989a). The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, eds.). Amsterdam: North-Holland, 473–488.

Kass, R. E., Tierney, L. and Kadane, J. B. (1989b). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76**, 663–674.

Kass, R. E., Tierney, L. and Kadane, J. B. (1991). Laplace's method in Bayesian analysis. *Statistical Multiple Integration* (N. Flournoy and R. K. Tsutakawa eds.). Providence: RI: ASA, 89-99.

Kass, R. E. and Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. Roy. Statist. Soc. B* **54**, 129–144.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypothesis and its relationship to the Schwartz criterion. *J. Amer. Statist. Assoc.* **90**, 928–934.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343–1370.

Kubokawa, T. and Robert, C. P. (1994). New perspectives in linear calibration. *J. Multivariate Analysis* **51**, 178–200.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University Press.

Kullback, S. (1959/1968). *Information Theory and Statistics*. New York: Wiley. Second edition in 1968, New York: Dover. Reprinted in 1978, Gloucester, MA: Peter Smith.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Lane, D. A. and Sudderth, W. D. (1983). Coherent and continuous inference. *Ann. Statist.* **11**, 114–120.

Lane, D. A. and Sudderth, W. D. (1984). Coherent predictive inference. *Sankhyā A* **46**, 166–185.

Laplace, P. S. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier. Reprinted as *Oeuvres Complètes de Laplace* **7**, 1878–1912. Paris: Gauthier-Villars.

Laplace, P. S. (1814/1952). *Essai Philosophique sur les Probabilitiés*. Paris: Courcier. The 5th edition (1825) was the last revised by Laplace. English translation in 1952 as *Philosophical Essay on Probabilities*. New York: Dover.

Lavine, M. (1992b). Sensitivity in Bayesian statistics: the prior and the likelihood. *J. Amer. Statist. Assoc.* **86**, 396–399.

Lavine, M. (1994). An approach to evaluating sensitivity in Bayesian regression analysis. *J. Statist. Planning and Inference* ,

Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *Canadian J. Statist.* **20**, 451–461.

Leamer, E. E. (1978). *Specification Searches: Ad hoc Inference with Nonexperimental Data*. New York: Wiley.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Pub. Statist.* **1**, 277–329.

LeCam, L. (1956). On the asymptotic theory of estimation and testing hypothesis. *Proc. Third Berkeley Symp.* **1** (J. Neyman and E. L. Scott, eds.). Berkeley: Univ. California Press, 129–156.

LeCam, L. (1958). Les propietés asymptotiques de solutions de Bayes. *Pub. Inst. Statist. Univ. Paris* **7**, 17–35.

LeCam, L. (1966). Likelihood functions for large number of independent observations. *Research Papers in Statistics. Festschrift for J. Neyman* (F. N. David, ed.). New York: Wiley, 167–187.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41**, 802–828.

LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Berlin: Springer.

Lecoutre, B. (1984). *L'Analyse Bayésienne des Comparaisons*. Lille: Presses Universitaires.

Lee, P. M. (1964). On the axioms of information theory. *Ann. Math. Statist.* **35**, 415–418.

Lee, P. M. (1989). *Bayesian Statistics: an Introduction*. London: Edward Arnold.

Lehmann, E. L. (1959/1983). *Theory of Point Estimation*. Second edition in 1983, New York: Wiley. Reprinted in 1991, Belmont, CA: Wadsworth.

Lehmann, E. L. (1959/1986). *Testing Statistical Hypotheses*. Second edition in 1986, New York: Wiley. Reprinted in 1991, Belmont, CA: Wadsworth.

Lehmann, E .L. (1990). Model specification: The views of Fisher and Neyman, and later developments. *Statist. Sci.* **5**, 160–168.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*. Rotterdam: University Press.

Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78**, 531–543.

Leonard, T. (1972). Bayesian methods for binomial data. *Biometrika* **59**, 581–589.

Leonard, T. (1973). A Bayesian method for histograms. *Biometrika* **60**, 297–308.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc. B* **37**, 23–37.

Leonard, T. (1977). A Bayesian approach to some multinomial estimation and pretesting problems. *J. Amer. Statist. Assoc.* **72**, 869–874.

Leonard, T. and Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *Ann. Statist.* **20**, 1669-1696.

Leonard, T. and Hsu, J. S. J. (1994). The Bayesian analysis of categorical data: a selective review. *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.). Chichester: Wiley, (to appear).

Leonard, T., Hsu, J. S. J. and Tsui, K.-W. (1989). Bayesian marginal inference. *J. Amer. Statist. Assoc.* **84**, 1051–1058.

Leonard, T. and Ord, K. (1976). An investigation of the F test procedure as an estimation short-cut. *J. Roy. Statist. Soc. B* **38**, 95–98.

Levine, R. D. and Tribus, M. (eds.) (1978). *The Maximum Entropy Formalism*. Cambridge, MA: The MIT Press.

Lindley, D. V. (1953). Statistical inference. *J. Roy. Statist. Soc. B* **15**, 30–76.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Lindley, D. V. (1958). Fiducial distribution and Bayes' Theorem. *J. Roy. Statist. Soc. B* **20**, 102–107.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *Proc. Fourth Berkeley Symp.* **1** (J. Neyman and E. L. Scott, eds.). Berkeley: Univ. California Press, 453–468.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *Ann. Math. Statist.* **35**, 1622-1643.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: University Press.

Lindley, D. V. (1969). Review of Fraser (1968). *Biometrika* **56**, 453–456.

Lindley, D. V. (1971). The estimation of many parameters. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 435–453 (with discussion).

Lindley, D. V. (1971/1985). *Making Decisions*. Second edition in 1985, Chichester: Wiley.

Lindley, D. V. (1972). *Bayesian Statistics, a Review*. Philadelphia, PA: SIAM.

Lindley, D. V. (1976). Bayesian Statistics. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* **2** (W, L. Harper and C. A. Hooker, eds.), Dordrecht: Reidel, 353–363.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika* **44**, 187–192.

Lindley, D. V. (1978). The Bayesian approach. *Scandinavian J. Statist.* **5**, 1–26.

Lindley, D. V. (1980a). Jeffreys's contribution to modern statistical thought. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 35–39.

Lindley, D. V. (1982a). Scoring rules and the inevitability of probability. *Internat. Statist. Rev.* **50**, 1–26 (with discussion).

Lindley, D. V. (1982b). Bayesian inference. *Encyclopedia of Statistical Sciences* **1** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 197–204.

Lindley, D. V. (1982c). Coherence. *Encyclopedia of Statistical Sciences* **2** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 29–31.

Lindley, D. V. (1982d). The improvement of probability judgements. *J. Roy. Statist. Soc. A* **145**, 117–126.

Lindley, D. V. (1984). The next 50 years. *J. Roy. Statist. Soc. A* **147**, 359–367.

Lindley, D. V. (1988). Statistical inference concerning Hardy-Weinberg equilibrium. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 307–326 (with discussion).

Lindley, D. V. (1990). The present position in Bayesian Statistics. *Statist. Sci.* **5**, 44–89 (with discussion).

Lindley, D. V. (1992). Is our view of Bayesian statistics too narrow? *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 1–15 (with discussion).

Lindley, D. V. (1993). On the presentation of evidence. *Math. Scientist* **18**, 60–63.

Lindley, D. V. (1997). Some comments on Bayes Factors. *J. Statist. Planning and Inference* **61**, 181–189.

Lindley, D. V. and Novick, M. R. (1981). The role of exchangeability in inference. *Ann. Statist.* **9**, 45–58.

Lindley, D. V. and Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.* **30**, 112–119.

Lindley, D. V. and Scott, W. F. (1985). *New Cambridge Elementary Statistical Tables*. Cambridge: University Press.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. B* **34**, 1–41 (with discussion).

Lindley, D. V., Tversky, A. and Brown, R. V. (1979). On the reconciliation of probability assessments. *J. Roy. Statist. Soc. A* **142**, 146–180.

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295–304.

Maín, P. (1988). Prior and posterior tail comparisons. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 669–675.

McCarthy, J. (1956). Measurements of the value of information. *Proc. Nat. Acad. Sci. USA* **42**, 654–655.

McCullagh P. and Tibshirani, R. (1990). A simple method for the adjustement of profile likelihoods. *J. Roy. Statist. Soc. B* **52**, 325–344.

McCulloch, R. E. (1989). Local model influence. *J. Amer. Statist. Assoc.* **84**, 473–478.

McCulloch, R. E. and Rossi, P. E. (1992). Bayes factors for non-linear hypothesis and likelihood distributions. *Biometrika* **79**, 663–676.

Meeden, G. (1990). Admissible contour credible sets. *Statistics and Decisions* **8**, 1–10.

Meeden, G. and Isaacson, D. (1977). Approximate behavior of the posterior distribution for a large observation. *Ann. Statist.* **5**, 899–908.

Meeden, G. and Vardeman, S. (1991). A non-informative Bayesian approach to interval estimation in finite population sampling. *J. Amer. Statist. Assoc.* **86**, 972–986.

Meinhold, R. and Singpurwalla, N. D. (1983). Understanding the Kalman filter. *Amer. Statist.* **37**, 123–127.

Mendel, M. B. (1992). Bayesian parametric models for lifetimes. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 697–705.

Mendoza, M. (1987). A Bayesian analysis of a generalized slope ratio bioassay. *Probability and Bayesian Statistics* (R. Viertl, ed.). London: Plenum, 357–364.

Mendoza, M. (1988). Inferences about the ratio of linear combinations of the coefficients in a multiple regression problem. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 705–711.

Mendoza, M. (1994). Asymptotic posterior normality under transformations. *Test* **3**, 173–180.

Mitchell, T. J. and Beauchamp, T. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83**, 1023–1035 (with discussion).

Moreno, E. and Cano, J. A. (1989). Testing a point null hypothesis: asymptotic robust Bayesian analysis with respect to priors given on a sub-sigma field. *Internat. Statist. Rev.* **57**, 221-232.

Morris, C. N. (1988). Approximating posterior distributions and posterior moments. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 327–344 (with discussion).

Mortera, J. (1986). Bayesian forecasting. *Metron* **44**, 277-296.

Mouchart, M. and Simar, L. (1980). Least squares approximation in Bayesian analysis. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 207–222 and 237–245 (with discussion).

Mukerjee, R. and Dey, D. K. (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: Higher order asymptotics. *Biometrika* **80**, 499–505.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalised linear models. *J. Roy. Statist. Soc. A* **135**, 370–384.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap.*J. Roy. Statist. Soc. B* **56**, 3–48 (with discussion).

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypothesis. *Phil. Trans. Roy. Soc. London A* **231**, 289–337.

Neyman, J. and Pearson, E. S. (1967). *Joint Statistical Papers*. Cambridge: University Press.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Nicolau, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J. Roy. Statist. Soc. B* **55**, 377–390.

Novick, M. R. (1969). Multiparameter Bayesian indifference procedures. *J. Roy. Statist. Soc. B* **31**, 29–64.

Novick, M. R. and Hall, W. K. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.* **60**, 1104–1117.

O'Hagan, A. (1981). A moment of indecision. *Biometrika* **68**, 329–330.

O'Hagan, A. (1988a). *Probability: Methods and Measurements*. London: Chapman and Hall.

O'Hagan, A. (1988b). Modelling with heavy tails. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 345–359 (with discussion).

O'Hagan, A. (1994a). *Kendall's Advanced Theory of Statistics* **2B**: *Bayesian Inference*. London: Edward Arnold

O'Hagan, A. (1995). Fractional Bayes factors for model comparison.*J. Roy. Statist. Soc. B* **57**, 99–138 (with discussion).

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factor. *Test* **6**, 101–118.

O'Hagan, A. and Berger, J. O. (1988). Ranges of posterior probabilities for quasimodal priors with specified quantiles. *J. Amer. Statist. Assoc.* **83**, 503–508.

O'Hagan, A. and Le, H. (1994). Conflicting information and a class of bivariate heavy-tailed distributions. *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.). Chichester: Wiley, 311-327.

Osteyee, D. D. B. and Good, I. J. (1974). *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection*. Berlin: Springer.

Pack, D. J. (1986a). Posterior distributions. Posterior probabilities. *Encyclopedia of Statistical Sciences* **7** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 121–124.

Pack, D. J. (1986b). Prior distributions. *Encyclopedia of Statistical Sciences* **7** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 194–196.

Parenti, G. (ed.) (1978). *I Fondamenti dell'Inferenza Statistica*. Florence: Università degli Studi.

Pearn, W. L. and Chen K. S. (1996). A Bayesian-like estimator of C_{pk}. *Comm. Statist. Theory and Methods* , (to appear).

Pearson, E. S. (1978). *The History of Statistics in the 17th and 18th Centuries*. London: Macmillan.

Peers, H. W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. B* **27**, 9–16.

Peers, H. W. (1968). Confidence properties of Bayesian interval estimates. *J. Roy. Statist. Soc. B* **30**, 535–544.

Pereira, C. A. de B. and Lindley, D. V. (1987). Examples questioning the use of partial likelihood. *The Statistician* **37**, 15–20.

Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika* **68**, 35–43.

Pericchi, L. R. (1984). An alternative to the standard Bayesian procedure for discrimination between normal linear models. *Biometrika* **71**, 576–586.

Pericchi, L. R. and Nazaret, W. A. (1988). On being imprecise at the higher levels of a hierarchical linear model. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 361–375 (with discussion).

Pericchi, L. R. and Walley, P. (1991). Robust Bayesian credible intervals and prior ignorance. *Internat. Statist. Rev.* **59**, 1–23.

Perks, W. (1947). Some observations on inverse probability, including a new indifference rule. *J. Inst. Actuaries* **73**, 285–334 (with discussion).

Pettit, L. I. (1986). Diagnostics in Bayesian model choice. *The Statistician* **35**, 183–190.

Pettit, L. I. (1992). Bayes factors for outlier models using the device of imaginary observations. *J. Amer. Statist. Assoc.* **87**, 541–545.

Pettit, L. I. and Young, K. S. (1990). Measuring the effect of observations on Bayes factors. *Biometrika* **77**, 455–466.

Pham-Gia, T. and Turkkan, N. (1992). Sample size determination in Bayesian analysis. *The Statistician* **41**, 389–404.

Philippe , A. and Robert, C.(1994). A note on the confidence properties of reference priors for the calibration model. *Test* **7**, 147–160.

Piccinato, L. (1973). Un metodo per determinare distribuzioni iniziali relativamente non-informative. *Metron* **31**, 124–156.

Piccinato, L. (1977). Predictive distributions and non-informative priors. *Trans. 7th. Prague Conf. Information Theory* (M. Uldrich, ed.). Prague: Czech. Acad. Sciences, 399–407.

Piccinato, L. (1992). Critical issues in different inferential paradigms. *J. It. Statist. Soc.* **2**, 251–274.

Pierce, D. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.* **1**, 241–250.

Pitman E. J. G. (1939). Location and scale parameters. *Biometrika* **36**, 391–421.

Plante, A. (1971). Counter-example and likelihood. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 357–371 (with discussion).

Plante, A. (1984). A reexamination of Stein's antifiducial example. *Canad. J. Statist.* **12**, 135–141.

Plante, A. (1991). An inclusion-consistent solution to the problem of absurd confidence statements. *Canad. J. Statist.* **19**, 389–397.

Poirier, D. J. (1985). Bayesian hypothesis testing in linear models with continuously induced conjugate priors across hypotheses. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 711–722.

Pole, A. and West, M. (1989). Reference analysis of the dynamic linear model. *J. Time Series Analysis* **10**, 131–147.

Pole, A., West, M. and Harrison P. J. (1994). *Applied Bayesian Forecasting and Time Series Analysis* (with computer software). London: Chapman and Hall. (to appear).

Polson, N. G. (1991). A representation of the posterior mean for a location model. *Biometrika* **78**, 426–430.

Polson, N. G. (1992a). In discussion of Ghosh and Mukerjee (1992). *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 203–205.

Polson, N. G. (1992b). On the expected amount of information from a non-linear model. *J. Roy. Statist. Soc. B* **54**, 889–895.

Polson, N. G. and Tiao, G. C., (eds.) (1995). *Bayesian Inference*. Brookfield, VT: Edward Elgar.

Polson, N. G. and Wasserman, L. (1990). Prior distributions for the bivariate binomial. *Biometrika* **77**, 901–904.

Poskitt, D. S. (1987). Precision, complexity and Bayesian model determination. *J. Roy. Statist. Soc. B* **49**, 199–208.

Pratt, J. W. (1961). Length of confidence intervals. *J. Amer. Statist. Assoc.* **56**, 549–567.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *J. Roy. Statist. Soc. B* **27**, 169–203.

Press, S. J. (1972/1982). *Applied Multivariate Analysis: using Bayesian and Frequentist Methods of Inference*. Second edition in 1982, Melbourne, FL: Krieger.

Press, S. J. (1985). Multivariate Analysis (Bayesian). *Encyclopedia of Statistical Sciences* **6** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 16–20.

Press, S. J. (1989). *Bayesian Statistics*. New York: Wiley.

Press, S. J. and Zellner, A. (1978). Posterior distribution for the multiple correlation coefficient with fixed regressors. *J. Econometrics* **8**, 307–321.

Rabena, M. (1998). Deriving reference decisions. *Test* **7**, 161–177.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University.

Raftery, A. E. and Schweder, T. (1993). Inference about the ratio of two parameters, with applications to whale censusing. *Amer. Statist.* **47**, 259–264.

Raftery, A. E. (1996). Hypothesis testing and model selection. *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.). London: Chapman and Hall, 163–187.

Rao, C. R. and Mukerjee, R. (1995). On posterior credible sets based on the score statistic. *Statistica Sinica* **5**, 781–791.

Reid, N. (1996). Likelihood and Bayesian approximation methods. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 349–366 (with discussion).

Renyi, A. (1955). On a new axiomatic theory of probability. *Acta Math. Acad. Sci. Hungaricae* **6**, 285–335.

Renyi, A. (1961). On measures of entropy and information. *Proc. Fourth Berkeley Symp.* **1** (J. Neyman and E. L. Scott, eds.). Berkeley: Univ. California Press, 547–561.

Renyi, A. (1962/1970). *Wahrscheinlichkeitsrechnung*. Berlin: Deutscher Verlag der Wissenschaften. English translation in 1970 as *Probability Theory*. San Francisco, CA: Holden-Day.

Renyi, A. (1964). On the amount of information concerning an unknown parameter in a sequence of observations. *Pub. Math. Inst. Hung. Acad. Sci.* **9**, 617–624.

Renyi, A. (1966). On the amount of missing information and the Neyman-Pearson lemma. *Research Papers in Statistics. Festschrift for J. Neyman* (F. N. David, ed.). New York: Wiley, 281–288.

Renyi, A. (1967). On some basic problems of statistics from the point of view of information theory. *Proc. Fifth Berkeley Symp.* **1** (L. M. LeCam and J Neyman, eds.). Berkeley: Univ. California Press, 531–543.

Ripley, B. D. (1987). *Stochastic Simulation*. Chichester: Wiley.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416–431.

Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc. B* **49**, 223-239 and 252-265 (with discussion).

Robert, C. P. (1992). *L'Analyse Statistique Bayésienne*. Paris: Economica.

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica* **3**, 603–608.

Robert, C.P. (1994) *The Bayesian Choice*. Berlin: Springer.

Robert, C. P. (1996). Intrinsic loss functions. *Theory and Decision* **40**, 191–214.

Robert, C. P. and Caron N. (1996). Noninformative Bayesian testing and neutral Bayes factors. *Test* **5**, 411–437.

Robert, C. P., Hwang, J. T. G. and Strawderman, W. E. (1993). Is Pitman closeness a reasonable criterion? *J. Amer. Statist. Assoc.* **88**, 57–76 (with discussion).

Roberts, H. V. (1965). Probabilistic prediction. *J. Amer. Statist. Assoc.* **60**, 50–62.

Roberts, H. V. (1967). Informative stopping rules and inferences about population size. *J. Amer. Statist. Assoc.* **62**, 763–775.

Roberts, H. V. (1974). Reporting of Bayesian studies. *Studies in Bayesian Econometrics and Statistics: in Honor of Leonard J. Savage* (S. E. Fienberg and A. Zellner, eds.). Amsterdam: North-Holland, 465–483.

Roberts, H. V. (1978). Bayesian inference. *International Encyclopedia of Statistics* (W. H. Kruskal, and J. M. Tanur, eds.). London: Macmillan, 9–16.

Robinson, G. K. (1975). Some counter-examples to the theory of confidence intervals. *Biometrika* **62**, 155–161.

Robinson, G. K. (1978). On the necessity of Bayesian inference and the construction of measures of nearness to Bayesian form. *Biometrika* **65**, 49–52.

Robinson, G. K. (1979a). Conditional properties of statistical procedures. *Ann. Statist.* **7**, 742–755.

Robinson, G. K. (1979b). Conditional properties of statistical procedures for location and scale parameters. *Ann. Statist.* **7**, 756–771.

Rodríguez, C. C. (1991). From Euclid to entropy. *Maximum Entropy and Bayesian Methods* (W. T. Grandy and L. H. Schick eds.). Dordrecht: Kluwer, 343–348.

Rousseau, J. (1997). *Etude des Propietés Asymptotiques des Estimateurs de Bayes*. Ph.D. Thesis, Université Paris VI, France..

Royall, R. M. (1992). The elusive concept of statistical evidence. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 405–418 (with discussion).

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9**, 130–134.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151–1172.

Rubin, H. (1971). A decision-theoretic approach to the problem of testing a null hypothesis. *Statistical Decision Theory and Related Topics* (S. S. Gupta and J. Yackel, eds.). New York: Academic Press, 103–108.

Rubin, H. (1988a). Some results on robustness in testing. *Statistical Decision Theory and Related Topics IV* **1** (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 271–278.

Rubin, H. and Sethuraman, J. (1966). Bayes risk efficiency. *Sankhyā A* **27**, 347–356.

Rueda, R. (1992). A Bayesian alternative to parametric hypothesis testing. *Test* **1**, 61-67.

Sacks, J. (1963). Generalized Bayes solutions in estimation problems. *Ann. Math. Statist.* **34**, 787–794.

Samaniego, F. J. and Reneau, D. M. (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *J. Amer. Statist. Assoc.* **89**, 847–957.

San Martini, A. and Spezzaferri F. (1984). A predictive model selection criterion. *J. Roy. Statist. Soc. B* **46**, 296–303.

Sansó, B. and Pericchi, L. R. (1992). Near ignorance classes of log-concave priors for the location model. *Test* **1**, 39–46.

Särndal C.-E. (1970). A class of explicata for 'information' and 'weight of evidence'. *Internat. Statist. Rev.* **38**, 223–235.

Savage, L. J. (1954/1972). *The Foundations of Statistics.* New York: Wiley. Second edition in 1972, New York: Dover.

Savage, L. J. (1962) (with others). *The Foundations of Statistical Inference: a Discussion*. London: Methuen.

Savage, L. J. (1961). The foundations of statistics reconsidered. *Proc. Fourth Berkeley Symp.* **1** (J. Neyman and E. L. Scott, eds.). Berkeley: Univ. California Press, 575–586. Reprinted in 1980 in *Studies in Subjective Probability* (H. E. Kyburg and H. E Smokler, eds.). New York: Dover, 175–188.

Savage, L. J. (1970). Reading suggestions for the foundations of statistics. *Amer. Statist.* **24**, 23–27.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**, 781–801. Reprinted in 1974 in *Studies in Bayesian Econometrics and Statistics: in Honor of Leonard J. Savage* (S. E. Fienberg and A. Zellner, eds.). Amsterdam: North-Holland, 111–156.

Savage, L. J. (1981). *The Writings of Leonard Jimmie Savage: a Memorial Collection*. Washington: ASA/IMS.

Schwartz, L. (1965). On Bayes procedures. *Z. Wahr.* **4**, 10–26.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Seidenfeld, T. (1987). Entropy and uncertainty. *Foundations of Statistical Inference*. (I. B. MacNeill and G. J. Umphrey eds.). Dordrecht: Reidel, 259–287.

Seidenfeld, T. (1979a). *Philosophical Problems of Statistical Inference*. Dordrecht: Reidel.

Seidenfeld, T. (1979b). Why I am not an objective Bayesian. 11, 413–

Seidenfeld, T. (1992). R. A. Fisher's fiducial argument and Bayes' theorem. *Statist. Sci.* **7**, 358–368.

Sendra, M. (1982). Distribución final de referencia para el problema de Fieller-Creasy. *Trab. Estadist.* **33**, 55–72.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Severini, T. A. (1991). On the relationship between Bayesian and non-Bayesian interval estimates. *J. Roy. Statist. Soc. B* **53**, 611–618.

Severini, T. A. (1993). Bayesian interval estimates which are also confidence intervals. *J. Roy. Statist. Soc. B* **53**, 611-618.

Severini, T. A. (1994). Approximately Bayesian inference. *J. Amer. Statist. Assoc.* **89**, 242–249.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton: University Press.

Shafer, G. (1982a). Belief functions and parametric models. *J. Roy. Statist. Soc. B* **44**, 322–352 (with discussion).

Shafer, G. (1982b). Lindley's paradox. *J. Amer. Statist. Assoc.* **77**, 325–351 (with discussion).

Shafer, G. (1986). Savage revisited. *Statist. Sci.* **1**, 435–462 (with discussion).

Shafer, G. (1990). The unity and diversity of probability. *Statist. Sci.* **5**, 463–501 (with discussion).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423 and 623–656. Reprinted in *The Mathematical Theory of Communication* (Shannon, C. E. and Weaver, W., 1949). Urbana, IL.: Univ. Illinois Press.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. B* **13**, 238–241.

Skilling, J. (ed.) (1989). *Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer.

Smith, A. F. M. (1973a). Bayes estimates in one-way and two way models. *Biometrika* **60**, 319–330.

Smith, A. F. M. (1973b). A general Bayesian linear model. *J. Roy. Statist. Soc. B* **35**, 67–75.

Smith, A. F. M. (1977). In discussion of Wilkinson (1977). *J. Roy. Statist. Soc. B* **39**, 145–147.

Smith, A. F. M. (1978). In discussion of Tanner (1978). *J. Roy. Statist. Soc. A* **141**, 50–51.

Smith, A. F. M. (1981). On random sequences with centred spherical symmetry. *J. Roy. Statist. Soc. B* **43**, 208–209.

Smith, A. F. M. (1984). Bayesian Statistics. Present position and potential developments: some personal views. *J. Roy. Statist. Soc. A* **147**. 245–259 (with discussion).

Smith, A. F. M. (1986). Some Bayesian thoughts on modeling and model choice. *The Statistician* **35**, 97–102.

Smith, A. F. M. (1988). What should be Bayesian about Bayesian software? *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 429–435 (with discussion).

Smith, A. F. M. (1991). Bayesian computational methods. *Phil. Trans. Roy. Soc. London A* **337**, 369–386.

Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *Amer. Statist.* **46**, 84–88.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. B* **55**, 3–23 (with discussion).

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. B* **42**, 213–220.

Smith, C. A. B. (1965). Personal probability and statistical analysis. *J. Roy. Statist. Soc. A* **128**, 469–499.

Smith, C. R. and Erickson, J. G. (eds.) (1987). *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*. Dordrecht: Reidel.

Smith, C. R. and Grandy, W. T. (eds.) (1985). *Maximum Entropy and Bayesian Methods in Inverse Problems*. Dordrecht: Reidel.

Soofi, E. S. (1994). Capturing the intangible concept of information. *J. Amer. Statist. Assoc.* **89**, 1243–1254.

Spall, J. C. and Hill, S. D. (1990). Least informative Bayesian prior distributions for finite samples based on information theory. *IEEE Trans. Automatic Control* **35**, 580–583.

Spiegelhalter, D. J. (1980). An omnibus test for normality for small samples. *Biometrika* **67**, 493–496.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. B* **44**, 377–387.

Stein, C. (1951). A property of some tests of composite hypotheses. *Ann. Math. Statist.* **22**, 475–476.

Stein, C. (1956). Inadmissibility of the usual estimation of the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp.* **1** (J. Neyman and E. L. Scott, eds.). Berkeley: Univ. California Press, 197–206.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877–880.

Stein, C. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. B* **24**, 265–296 (with discussion).

Stein, C. (1965). Approximation of improper prior measures by proper probability measures. *Bernoulli, Bayes, Laplace Festschrift.* (J. Neyman and L. LeCam, eds.). Berlin: Springer, 217–240.

Stein, C. (1982). On the coverage probability of confidence sets based on a prior distribution.*Tech. Rep.* **180**, Stanford University, USA.

Stein, C. (1985). On the coverage probability of confidence sets based on a prior distribution. Sequential Methods in Statistics (R. Zielisnsi, ed.). Warsaw: Polish Scientific Pub., 485—514.

Stephens, D. A. and Smith, A. F. M. (1992). Sampling-resampling techniques for the computation of posterior densities in normal means problems. *Test* **1**, 1–18.

Stewart, L. (1979). Multiparameter univariate Bayesian analysis. *J. Amer. Statist. Assoc.* **74**, 684–693.

Stigler, S. M. (1982). Thomas Bayes' Bayesian inference. *J. Roy. Statist. Soc. A* **145**, 250–258.

Stone, M. (1959). Application of a measure of information to the design and comparison of experiments. *Ann. Math. Statist.* **30**, 55–70.

Stone, M. (1963). The posterior t distribution. *Ann. Math. Statist.* **34**, 568–573.

Stone, M. (1965). Right Haar measures for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.* **36**, 440–453.

Stone, M. (1969). The role of experimental randomization in Bayesian statistics: Finite sampling and two Bayesians. *Biometrika* **56**, 681–683.

Stone, M. (1970). Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.* **41**, 1939–1953.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B* **36**, 11–147 (with discussion).

Stone, M. (1976). Strong inconsistency from uniform priors. *J. Amer. Statist. Assoc.* **71**, 114–125 (with discussion).

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. B* **39**, 44–47.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *J. Roy. Statist. Soc. B* **41**, 276–278.

Stone, M. (1982). Review and analysis of some inconsistencies related to improper distributions and finite additivity. *Logic, Methodology and Philosophy of Science* (L. J. Cohen, J. Løs, H. Pfeiffer and K. P. Podewski, eds.). Amsterdam: North-Holland, 413–426.

Stone, M. and Dawid, A. P. (1972). Un-Bayesian implications of improper Bayesian inference in routine statistical problems. *Biometrika* **59**, 369–375.

Sudderth, W. D. (1980). Finitely additive priors, coherence and the marginalization paradox. *J. Roy. Statist. Soc. B* **42**, 339–341.

Sun, D. (1994). Integrable expansions for posterior distributions for a two-parameter exponential family. *Ann. Statist.* **22**, 1808–1830.

Sun, D. (1996). *NSBayes*. This is an electronic mailing list on Non-Subjective Bayesian methods. For further information see <http://www.stat.missouri.edu/ bayes>.

Sun, D. (1997). A note on noninformative priors for Weibull distributions. *J. Statist. Planning and Inference* **61**, 319–338.

Sun, D. and Berger, J. O. (1998). Reference priors under partial information. *Biometrika* **85** appear

Sun, D., Ghosh, M. and Basu, A. P. (1997). Bayesian analysis for a stress-strength system via noninformative priors. *Canadian J. Statist.* (to appear).

Sun, D. and Ye, K. (1995). Reference prior Bayesian analysis for normal mean products. *J. Amer. Statist. Assoc.* **90**, 589–597.

Sun, D. and Ye, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika* **83**, 55-65.

Sweeting, T. J. (1984). On the choice of prior distributions for the Box-Cox transformed linear model. *Biometrika* **71**, 127–134.

Sweeting, T. J. (1985). Consistent prior distributions for transformed models. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 755–762.

Sweeting, T. J. (1985). Discussion of Cox and Reid (1987). *J. Roy. Statist. Soc. B* **49**, 20–21.

Sweeting, T. J. (1992). On asymptotic posterior normality in the multiparameter case. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 825–835.

Sweeting, T. J. and Adekola, A. D. (1987). Asymptotic posterior normality for stochastic processes revisited *J. Roy. Statist. Soc. B* **49**, 215–222.

Tanner, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. Berlin: Springer.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**, 582–548 (with discussion).

Thatcher, A. R. (1964). Relationships between Bayesian and confidence limits for prediction *J. Roy. Statist. Soc. B* **26**, 126–210.

Thorburn, D. (1986). A Bayesian approach to density estimation. *Biometrika* **73**, 65–75.

Tiao, G. C. and Box, G. E. P. (1974). Some comments on ‘Bayes’ estimators. *Studies in Bayesian Econometrics and Statistics: in Honor of Leonard J. Savage* (S. E. Fienberg and A. Zellner, eds.). Amsterdam: North-Holland, 620–626.

Tiao, G. C. and Zellner, A. (1964). On the Bayesian estimation of multivariate regression. *J. Roy. Statist. Soc. B* **26**, 277–285.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–608.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.

Torgesen, E. N. (1981). Measures of information based on comparison with total information and with total ignorance. *Ann. Statist.* **9**, 638–657.

Tribus, M. (1962). The use of the maximum entropy estimate in reliability engineering. *Recent Developments in Decision and Information Processes* (R. E. Machol and P. Gray, eds.). London: Macmillan, 102–140.

Venegas, F. (1990). On regularity and optimality conditions for maximum entropy priors. *Rev. Bras. Probab. Estatis.* **4**, 105–136.

Verdinelli, I. and Wasserman, L. (1995a). Bayes Factors, nuisance parameters and imprecise tests. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 765–771.

Verdinelli, I. and Wasserman, L. (1995b). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Amer. Statist. Assoc.* **90**, 614–618.

Viertl, R. (ed.) (1987). *Probability and Bayesian Statistics*. London: Plenum.

Villegas, C. (1969). On the a priori distribution of the covariance matrix. *Ann. Math. Statist.* **40**, 1098–1099.

Villegas, C. (1971). On Haar priors. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 409–414 (with discussion).

Villegas, C. (1977a). On the representation of ignorance. *J. Amer. Statist. Assoc.* **72**, 651–654.

Villegas, C. (1977b). Inner statistical inference. *J. Amer. Statist. Assoc.* **72**, 453–458.

Villegas, C. (1981). Inner statistical inference II. *Ann. Statist.* **9**, 768–776.

Villegas, C. (1990). Bayesian inference in models with euclidean structures. *J. Amer. Statist. Assoc.* **85**, 1159–1164.

Wahba, G. (1978). Improper priors, spline smoothing and the problems of guarding against model errors in regression. *J. Roy. Statist. Soc. B* **40**, 364–372.

Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *J. Roy. Statist. Soc. B* **45**, 133–150.

Wahba, G. (1988). Partial and interaction spline models. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 479–491 (with discussion).

Wakefield, J. C., Gelfand, A. E. and Smith, A. F. M. (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing* **1**, 129–133.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypothesis. *Ann. Math. Statist.* **10**, 299–326.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *J. Roy. Statist. Soc. B* **31**, 80–88.

Wallace, C. S. and Freeman, P. R. (1987). Estimation and inference by compact coding. *J. Roy. Statist. Soc. B* **49**, 240–260 (with discussion).

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Wasserman, L. (1989). A robust Bayesian interpretation of likelihood regions. *Ann. Statist.* **17**, 1387–1393.

Wasserman, L. (1990a). Belief functions and statistical inference. *Canadian J. Statist.* **18**, 183–196.

Wasserman, L. (1990b). Prior envelopes based on belief functions. *Ann. Statist.* **18**, 454–464.

Wasserman L. (1991). An inferential interpretation of default priors. *Tech. Rep.* **516**, Carnegie Mellon University, USA.

Wasserman, L. (1992b). Invariance properties of density ratio priors. *Ann. Statist.* **20**, 2177–2182.

Wasserman L. (1995). The conflict between improper priors and robustness. *J. Statist. Planning and Inference* **52**, 1–15.

Wasserman, L. and Clarke, B. (1995). Information tradeoff. *Test* **4**, 19–38.

Welch, B. L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location. *Ann. Math. Statist.* **10**, 58–69.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika* **34**, 28–35.

Welch, B. L. (1965). On comparisons between confidence point procedures in the case of a single parameter. *J. Roy. Statist. Soc. B* **27**, 1–8.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *J. Roy. Statist. Soc. B* **25**, 318–329.

West, M. (1992). Modelling with mixtures. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 503–524 (with discussion).

West, M. and Harrison, P. J. (1989). *Bayesian Forecasting and Dynamic Models*. Berlin: Springer.

Wilkinson, G. N. (1977). On resolving the controversy in statistical inference. *J. Roy. Statist. Soc. B* **39**, 119–171 (with discussion).

Willing, R. (1988). Information contained in nuisance parameters. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 801–805.

Wright, D. E. (1986). A note on the construction of highest posterior density intervals. *Appl. Statist.* **35**, 49–53.

Wrinch, D. H. and Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Phil. Magazine 6*, **42**, 363–390; **45**, 368–374.

Yang, R. (1995). Invariance of the reference prior under reparametrization. *Test* **4**, 83–94.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22**, 1195–1211.

Yang, R. and Berger, J. O. (1996). A catalog of noninformative priors. *Tech. Rep.*, Purdue University, USA..

Yang, R and Chen, M.-H. (1995). Bayesian analysis of random coefficient regression models using noninformative priors. *J. Multivariate Analysis* **55**, 283–311.

Ye, K. (1993). Reference priors when the stopping rule depends on the parameter of interest. *J. Amer. Statist. Assoc.* **88**, 360–363.

Ye, K. (1994). Bayesian reference prior analysis on the ratio of variances for the balanced one-way random effect model. *J. Statist. Planning and Inference* **41**, 267–280.

Ye, K. (1995). Selection of the reference priors for a balanced random effects model. *Test* **4**, 1–17.

Ye, K. and Berger, J. O. (1991). Non-informative priors for inferences in exponential regression models. *Biometrika* **78**, 645–656.

Yfantis, E. A. and Flatman, G. T. (1991). On confidence interval for product of three means of three normally distributed populations. *J. Chemometrics* **5**, 309–319.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley. Reprinted in 1987, Melbourne, FL: Krieger.

Zellner, A. (1977). Maximal data information prior distributions. *New Developments in the Applications of Bayesian Methods* (A. Aykaç and C. Brumat, eds.). Amsterdam: North-Holland, 211–232.

Zellner, A. (ed.) (1980). *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*. Amsterdam: North-Holland.

Zellner, A. (1984). Posterior odds ratios for regression hypothesis: general considerations and some specific results. *Basic Issues in Econometrics* (A. Zellner, ed.). Chicago: University Press, 275–305.

Zellner, A. (1985). Bayesian econometrics. *Econometrica* **53**, 253–269.

Zellner, A. (1986a). On assessing prior distibutions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Amsterdam: North-Holland, 233–243.

Zellner, A. (1986b). Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Statist. Assoc.* **81**, 446–451.

Zellner, A. (1988). Optimal information processing and Bayes' theorem. *Amer. Statist.* **42**, 278–284 (with discussion).

Zellner, A. (1991). Bayesian methods and entropy in economics and econometrics. *Maximum Entropy and Bayesian Methods* (W. T. Grandy and L. H. Schick eds.). Dordrecht: Kluwer, 17–31.

Zellner, A. (1996). Models, prior information and Bayesian analysis. *J. Econometrics* **75**, 51–68.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypothesis. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 585–603 and 618–647 (with discussion).

Zidek, J. (1969). A representation of Bayes invariant procedures in terms of Haar measure. *Ann. Inst. Statist. Math.* **21**, 291–308.