Handbook of
THE PHILOSOPHY OF SCIENCE

General Editors: DOV M. GABBAY, PAUL THAGARD, AND JOHN WOODS

PHILOSOPHY *of* STATISTICS

Edited by Prasanta S. Bandyopadhyay
and Malcolm R. Forster

N·H

CONTENTS

Introduction

Part I. Probability & Statistics

Part II. Philosophical Controversies about Conditional Probability

Part III. Four Paradigms of Statistics

Classical Statistics Paradigm

Bayesian Paradigm

Likelihood Paradigm

Akaikean Paradigm

Part IV: The Likelihood Principle

Part V: Recent Advances in Model Selection

Part VI: Attempts to Understand Different Aspects of "Randomness"

Part VII: Probabilistic and Statistical Paradoxes

Part VIII: Statistics and Inductive Inference

Part IX: Various Issues about Causal Inference

Part X: Some Philosophical Issues Concerning Statistical Learning Theory

Part XI: Different Approaches to Simplicity Related to Inference and Truth

Part XII: Special Problems in Statistics/Computer Science

Part XIII: An Application of Statistics to Climate Change

Part XIV: Historical Approaches to Probability/Statistics

MODERN BAYESIAN INFERENCE: FOUNDATIONS AND OBJECTIVE METHODS

José M. Bernardo

The field of statistics includes two major paradigms: frequentist and Bayesian. Bayesian methods provide a *complete* paradigm for both statistical inference and decision making under uncertainty. Bayesian methods may be derived from an axiomatic system and provide a *coherent* methodology which makes it possible to incorporate relevant initial information, and which solves many of the difficulties which frequentist methods are known to face. If no prior information is to be assumed, a situation often met in scientific reporting and public decision making, a formal initial prior function must be mathematically derived from the assumed model. This leads to *objective* Bayesian methods, objective in the precise sense that their results, like frequentist results, only depend on the assumed model and the data obtained. The Bayesian paradigm is based on an interpretation of probability as a *rational conditional measure of uncertainty*, which closely matches the sense of the word 'probability' in ordinary language. Statistical inference about a quantity of interest is described as the modification of the uncertainty about its value in the light of evidence, and Bayes' theorem specifies how this modification should precisely be made.

1 INTRODUCTION

Scientific experimental or observational results generally consist of (possibly many) sets of data of the general form $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where the $\mathbf{x}_i$'s are somewhat "homogeneous" (possibly multidimensional) observations $\mathbf{x}_i$. Statistical methods are then typically used to derive conclusions on both the nature of the process which has produced those observations, and on the expected behaviour at future instances of the same process. A central element of *any* statistical analysis is the specification of a *probability model* which is assumed to describe the mechanism which has generated the observed data D as a function of a (possibly multidimensional) parameter (vector) $\omega \in \Omega$, sometimes referred to as the *state of nature*, about whose value only limited information (if any) is available. All derived statistical conclusions are obviously conditional on the assumed probability model.

Unlike most other branches of mathematics, frequentist methods of statistical inference suffer from the lack of an axiomatic basis; as a consequence, their proposed desiderata are often mutually incompatible, and the analysis of the same data may well lead to incompatible results when different, apparently intuitive

Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.
Volume editors: Prasanta S. Bandyopadhyay and Malcolm R. Forster. General Editors: Dov M. Gabbay, Paul Thagard and John Woods.

procedures are tried; see Lindley [1972] and Jaynes [1976] for many instructive examples. In marked contrast, the Bayesian approach to statistical inference is firmly based on axiomatic foundations which provide a unifying logical structure, and guarantee the mutual consistency of the methods proposed. Bayesian methods constitute a *complete* paradigm to statistical inference, a scientific revolution in Kuhn's sense.

Bayesian statistics only require the *mathematics* of probability theory and the *interpretation* of probability which most closely corresponds to the standard use of this word in everyday language: it is no accident that some of the more important seminal books on Bayesian statistics, such as the works of de Laplace [1812], Jeffreys [1939] or de Finetti [1970] are actually entitled "Probability Theory". The practical consequences of adopting the Bayesian paradigm are far reaching. Indeed, Bayesian methods (i) reduce statistical inference to problems in probability theory, thereby minimizing the need for completely new concepts, and (ii) serve to discriminate among conventional, typically frequentist statistical techniques, by either providing a logical justification to some (and making explicit the conditions under which they are valid), or proving the logical inconsistency of others.

The main result from these foundations is the mathematical *need* to describe by means of probability distributions all uncertainties present in the problem. In particular, unknown parameters in probability models *must* have a joint probability distribution which describes the available information about their values; this is often regarded as *the* characteristic element of a Bayesian approach. Notice that (in sharp contrast to conventional statistics) *parameters are treated as random variables* within the Bayesian paradigm. This is not a description of their variability (parameters are typically *fixed unknown* quantities) but a description of the *uncertainty* about their true values.

A most important particular case arises when either no relevant prior information is readily available, or that information is subjective and an "objective" analysis is desired, one that is exclusively based on accepted model assumptions and well-documented public prior information. This is addressed by *reference analysis* which uses information-theoretic concepts to derive formal *reference* prior functions which, when used in Bayes' theorem, lead to posterior distributions encapsulating inferential conclusions on the quantities of interest solely based on the assumed model and the observed data.

In this article it is assumed that probability distributions may be described through their probability density functions, and no distinction is made between a random quantity and the particular values that it may take. Bold italic roman fonts are used for *observable* random vectors (typically data) and bold italic greek fonts are used for unobservable random vectors (typically parameters); lower case is used for variables and calligraphic upper case for their dominion sets. Moreover, the standard mathematical convention of referring to *functions*, say f and g of $\mathbf{x} \in \mathcal{X}$, respectively by $f(\mathbf{x})$ and $g(\mathbf{x})$, will be used throughout. Thus, $\pi(\theta|D, C)$ and $p(\mathbf{x}|\theta, C)$ respectively represent general *probability densities* of the unknown parameter $\theta \in \Theta$ given data D and conditions C, and of the observable random vector $\mathbf{x} \in \mathcal{X}$ conditional on θ and C. Hence, $\pi(\theta|D, C) \geq 0$, $\int_\Theta \pi(\theta|D, C)d\theta = 1$, and $p(\mathbf{x}|\theta, C) \geq 0$, $\int_\mathcal{X} p(\mathbf{x}|\theta, C)\, d\mathbf{x} = 1$. This admittedly imprecise notation will greatly simplify the exposition. If the random vectors are discrete, these functions naturally become probability mass functions, and integrals over their values become sums. Density functions of specific distributions are denoted by appropriate names. Thus, if x is a random quantity with a normal distribution of mean μ and standard deviation σ, its probability density function will be denoted $\mathbf{N}(x|\mu, \sigma)$.

Bayesian methods make frequent use of the the concept of logarithmic divergence, a very general measure of the goodness of the approximation of a probability density $p(\mathbf{x})$ by another density $\hat{p}(\mathbf{x})$. The Kullback-Leibler, or *logarithmic divergence* of a probability density $\hat{p}(\mathbf{x})$ of the random vector $\mathbf{x} \in \mathcal{X}$ from its true probability density $p(\mathbf{x})$, is defined as $\kappa\{\hat{p}(\mathbf{x})|p(\mathbf{x})\} = \int_\mathcal{X} p(\mathbf{x}) \log\{p(\mathbf{x})/\hat{p}(\mathbf{x})\}\, d\mathbf{x}$. It may be shown that (i) the logarithmic divergence is non-negative (and it is zero if, and only if, $\hat{p}(\mathbf{x}) = p(\mathbf{x})$ almost everywhere), and (ii) that $\kappa\{\hat{p}(\mathbf{x})|p(\mathbf{x})\}$ is invariant under one-to-one transformations of $\mathbf{x}$.

This article contains a brief summary of the mathematical foundations of Bayesian statistical methods (Section 2), an overview of the paradigm (Section 3), a detailed discussion of objective Bayesian methods (Section 4), and a description of useful objective inference summaries, including estimation and hypothesis testing (Section 5).

Good introductions to objective Bayesian statistics include Lindley [1965], Zellner [1971], and Box and Tiao [1973]. For more advanced monographs, see [Berger, 1985; Bernardo and Smith, 1994].

2 FOUNDATIONS

A central element of the Bayesian paradigm is the use of probability distributions to describe all relevant unknown quantities, interpreting the probability of an event as a conditional measure of uncertainty, on a $[0, 1]$ scale, about the occurrence of the event in some specific conditions. The limiting extreme values 0 and 1, which are typically inaccessible in applications, respectively describe impossibility and certainty of the occurrence of the event. This interpretation of probability includes and extends all other probability interpretations. There are two independent arguments which prove the mathematical inevitability of the use of probability distributions to describe uncertainties; these are summarized later in this section.

2.1 Probability as a Rational Measure of Conditional Uncertainty

Bayesian statistics uses the word *probability* in precisely the same sense in which this word is used in everyday language, as a *conditional measure of uncertainty* associated with the occurrence of a particular event, given the available information and the accepted assumptions. Thus, $\Pr(E|C)$ is a measure of (presumably rational) belief in the occurrence of the *event* E under *conditions* C. It is important to stress that probability is *always* a function of two arguments, the event E whose uncertainty is being measured, and the conditions C under which the measurement takes place; "absolute" probabilities do not exist. In typical applications, one is interested in the probability of some event E given the available *data* D, the set of *assumptions* A which one is prepared to make about the mechanism which has generated the data, and the relevant contextual *knowledge* K which might be available. Thus, $\Pr(E|D, A, K)$ is to be interpreted as a measure of (presumably rational) belief in the occurrence of the *event* E, given data D, assumptions A and any other available knowledge K, as a measure of how "likely" is the occurrence of E in these conditions. Sometimes, but certainly not always, the probability of an event under given conditions may be associated with the relative frequency of "similar" events in "similar" conditions. The following examples are intended to illustrate the use of probability as a conditional measure of uncertainty.

Probabilistic diagnosis. A human population is known to contain 0.2% of people infected by a particular virus. A person, *randomly selected* from that population, is subject to a test which is from laboratory data known to yield positive results in 98% of infected people and in 1% of non-infected, so that, if V denotes the event that a person carries the virus and + denotes a positive result, $\Pr(+|V) = 0.98$ and $\Pr(+|\overline{V}) = 0.01$. Suppose that the result of the test turns out to be positive. Clearly, one is then interested in $\Pr(V|+, A, K)$, the *probability* that the person carries the virus, given the positive result, the assumptions A about the probability mechanism generating the test results, and the available knowledge K of the prevalence of the infection in the population under study (described here by $\Pr(V|K) = 0.002$). An elementary exercise in probability algebra, which involves Bayes' theorem in its simplest form (see Section 3), yields $\Pr(V|+, A, K) = 0.164$. Notice that the four probabilities involved in the problem have *the same interpretation*: they are all conditional measures of uncertainty. Besides, $\Pr(V|+, A, K)$ is *both* a measure of the uncertainty associated with the event that the particular person who tested positive is actually infected, *and* an *estimate* of the proportion of people in that population (about 16.4%) that would eventually prove to be infected among those which yielded a positive test. $\triangleleft$

Estimation of a proportion. A survey is conducted to estimate the proportion θ of individuals in a population who share a given property. A random sample of n elements is analyzed, r of which are found to possess that property. One is then typically interested in using the results from the sample to establish regions of $[0, 1]$ where the unknown value of θ may plausibly be expected to lie; this information is provided by *probabilities* of the form $\Pr(a < \theta < b|r, n, A, K)$, a conditional measure of the uncertainty about the event that θ belongs to (a, b) *given* the information provided by the data (r, n), the assumptions A made on the behaviour of the mechanism which has generated the data (a random sample of n Bernoulli trials), and any relevant knowledge K on the values of θ which might be available. For example, after a political survey in which 720 citizens out of a random sample of 1500 have declared their support to a particular political measure, one may conclude that $\Pr(\theta < 0.5 | 720, 1500, A, K) = 0.933$, indicating a probability of about 93% that a referendum on that issue would be lost. Similarly, after a screening test for an infection where 100 people have been tested, none of which has turned out to be infected, one may conclude that $\Pr(\theta < 0.01 | 0, 100, A, K) = 0.844$, or a probability of about 84% that the proportion of infected people is smaller than 1%. $\triangleleft$

Measurement of a physical constant. A team of scientists, intending to establish the unknown value of a physical constant μ, obtain data $D = \{x_1, \ldots, x_n\}$ which are considered to be measurements of μ subject to error. The probabilities of interest are then typically of the form $\Pr(a < \mu < b | x_1, \ldots, x_n, A, K)$, the *probability* that the unknown value of μ (fixed in nature, but unknown to the scientists) lies within an interval (a, b) given the information provided by the data D, the assumptions A made on the behaviour of the measurement mechanism, and whatever knowledge K might be available on the value of the constant μ. Again, those probabilities are conditional measures of uncertainty which describe the (necessarily probabilistic) conclusions of the scientists on the true value of μ, given available information and accepted assumptions. For example, after a classroom experiment to measure the gravitational field with a pendulum, a student may report (in m/sec^2) something like $\Pr(9.788 < g < 9.829 | D, A, K) = 0.95$, meaning that, under accepted knowledge K and assumptions A, the *observed* data D indicate that the true value of g lies within 9.788 and 9.829 with probability 0.95, a conditional uncertainty measure on a [0,1] scale. This is naturally compatible with the fact that the value of the gravitational field at the laboratory may well be known with high precision from available literature or from precise previous experiments, but the student may have been instructed *not* to use that information as part of the accepted knowledge K. Under some conditions, it is also true that if the same *procedure* were actually used by many other students with similarly obtained data sets, their reported intervals would actually cover the true value of g in approximately 95% of the cases, thus providing a frequentist *calibration* of the student's probability statement. $\triangleleft$

Prediction. An experiment is made to count the number r of times that an event E takes place in each of n replications of a well defined situation; it is observed that E does take place r_i times in replication i, and it is desired to forecast the number of times r that E will take place in a similar future situation. This is a *prediction* problem on the value of an *observable* (discrete) quantity r, given the information provided by data D, accepted assumptions A on the probability mechanism which generates the r_i's, and any relevant available knowledge K. Computation of the probabilities $\{\Pr(r | r_1, \ldots, r_n, A, K)\}$, for $r = 0, 1, \ldots$, is thus required. For example, the quality assurance engineer of a firm which produces automobile restraint systems may report something like $\Pr(r = 0 | r_1 = \ldots = r_{10} = 0, A, K) = 0.953$, after observing that the entire production of airbags in each of $n = 10$ consecutive months has yielded no complaints from their clients. This should be regarded as a measure, on a $[0, 1]$ scale, of the conditional uncertainty, given observed data, accepted assumptions and contextual knowledge, associated with the event that no airbag complaint will come from next month's production and, if conditions remain constant, this is also an estimate of the proportion of months expected to share this desirable property.

A similar problem may naturally be posed with continuous observables. For instance, after measuring some continuous magnitude in each of n randomly chosen elements within a population, it may be desired to forecast the proportion of items in the whole population whose magnitude satisfies some precise specifications. As an example, after measuring the breaking strengths $\{x_1, \ldots, x_{10}\}$ of 10 randomly chosen safety belt webbings to verify whether or not they satisfy the requirement of remaining above 26 kN, the quality assurance engineer may report something like $\Pr(x > 26 | x_1, \ldots, x_{10}, A, K) = 0.9987$. This should be regarded as a measure, on a $[0, 1]$ scale, of the conditional uncertainty (given observed data, accepted assumptions and contextual knowledge) associated with the event that a randomly chosen safety belt webbing will support no less than 26 kN. If production conditions remain constant, it will also be an estimate of the proportion of safety belts which will conform to this particular specification.

Often, additional information of future observations is provided by related covariates. For instance, after observing the outputs $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ which correspond to a sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of different production conditions, it may be desired to forecast the output $\mathbf{y}$ which would correspond to a particular set $\mathbf{x}$ of production conditions. For instance, the viscosity of commercial condensed milk is required to be within specified values a and b; after measuring the viscosities $\{y_1, \ldots, y_n\}$ which correspond to samples of condensed milk produced under different physical conditions $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, production engineers will require probabilities of the form $\Pr(a < y < b | \mathbf{x}, (y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n), A, K)$. This is a conditional measure of the uncertainty (always given observed data, accepted assumptions and contextual knowledge) associated with the event that condensed milk produced under conditions $\mathbf{x}$ will actually satisfy the required viscosity specifications. ◁

2.2 Statistical Inference and Decision Theory

Decision theory not only provides a precise methodology to deal with decision problems under uncertainty, but its solid axiomatic basis also provides a powerful reinforcement to the logical force of the Bayesian approach. We now summarize the basic argument.

A decision problem exists whenever there are two or more possible courses of action; let $\mathcal{A}$ be the class of possible actions. Moreover, for each $a \in \mathcal{A}$, let Θ_a be the set of *relevant events* which may affect the result of choosing a, and let

$c(a,\theta) \in \mathcal{C}_a$, $\theta \in \Theta_a$, be the *consequence* of having chosen action a when event θ takes place. The class of pairs $\{(\Theta_a, \mathcal{C}_a), a \in \mathcal{A}\}$ describes the *structure* of the decision problem. Without loss of generality, it may be assumed that the possible actions are mutually exclusive, for otherwise one would work with the appropriate Cartesian product.

Different sets of principles have been proposed to capture a minimum collection of logical rules that could sensibly be required for "rational" decision-making. These all consist of axioms with a strong intuitive appeal; examples include the *transitivity* of preferences (if $a_1 > a_2$ given C, and $a_2 > a_3$ given C, then $a_1 > a_3$ given C), and the *sure-thing principle* (if $a_1 > a_2$ given C and E, and $a_1 > a_2$ given C and not E, then $a_1 > a_2$ given C). Notice that these rules are not intended as a description of actual human decision-making, but as a *normative* set of principles to be followed by someone who aspires to achieve coherent decision-making.

There are naturally different options for the set of acceptable principles (see e.g. Ramsey 1926; Savage, 1954; DeGroot, 1970; Bernardo and Smith, 1994, Ch. 2 and references therein), but all of them lead basically to the same conclusions, namely:

(i) Preferences among consequences should be measured with a real-valued bounded *utility* function $U(c) = U(a,\theta)$ which specifies, on some numerical scale, their desirability.

(ii) The uncertainty of relevant events should be measured with a set of *probability* distributions $\{(\pi(\theta|C,a), \theta \in \Theta_a), a \in \mathcal{A}\}$ describing their plausibility given the conditions C under which the decision must be taken.

(iii) The desirability of the available actions is measured by their corresponding *expected utility*

$$(1) \quad \overline{U}(a|C) = \int_{\Theta_a} U(a,\theta)\, \pi(\theta|C,a)\, d\theta, a \in \mathcal{A}.$$

It is often convenient to work in terms of the non-negative *loss* function defined by

$$(2) \quad L(a,\theta) = \sup_{a \in \mathcal{A}} \{U(a,\theta)\} - U(a,\theta),$$

which directly measures, as a function of θ, the "penalty" for choosing a wrong action. The relative undesirability of available actions $a \in \mathcal{A}$ is then measured by their *expected loss*

$$(3) \quad \overline{L}(a|C) = \int_{\Theta_a} L(a,\theta)\, \pi(\theta|C,a)\, d\theta, \quad a \in \mathcal{A}.$$

Notice that, in particular, the argument described above establishes the need to quantify the uncertainty about all relevant unknown quantities (the actual values of the θ's), and specifies that this quantification *must* have the mathematical structure of probability distributions. These probabilities are conditional on the circumstances C under which the decision is to be taken, which typically, but not necessarily, include the results D of some relevant experimental or observational data.

It has been argued that the development described above (which is not questioned when decisions have to be made) does not apply to problems of statistical inference, where no specific decision making is envisaged. However, there are two powerful counterarguments to this. Indeed, (i) a problem of statistical inference is typically considered worth analyzing because it *may* eventually help make sensible decisions; a lump of arsenic is poisonous because it *may* kill someone, not because it has actually killed someone [Ramsey, 1926], and (ii) it has been shown [Bernardo, 1979a] that statistical inference on θ actually *has* the mathematical structure of a decision problem, where the class of alternatives is the functional space

$$(4) \quad \mathcal{A} = \left\{ \pi(\theta|D); \quad \pi(\theta|D) > 0, \int_{\Theta} \pi(\theta|D)\, d\theta = 1 \right\}$$

of the conditional probability distributions of θ given the data, and the utility function is a measure of the amount of information about θ which the data may be expected to provide.

2.3 Exchangeability and Representation Theorem

Available data often take the form of a set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of "homogeneous" (possibly multidimensional) observations, in the precise sense that only their *values* matter and not the *order* in which they appear. Formally, this is captured by the notion of *exchangeability.* The set of random vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is exchangeable if their joint distribution is invariant under permutations. An infinite sequence $\{\mathbf{x}_j\}$ of random vectors is exchangeable if all its finite subsequences are exchangeable. Notice that, in particular, any random sample from any model is exchangeable in this sense. The concept of exchangeability, introduced by de Finetti [1937], is central to modern statistical thinking. Indeed, the general *representation theorem* implies that if a set of observations is assumed to be a subset of an exchangeable sequence, then it constitutes *a random sample* from some probability model $\{p(\mathbf{x}|\omega), \omega \in \boldsymbol{\Omega}\}$, $\mathbf{x} \in \mathcal{X}$, described in terms of (labeled by) some *parameter vector* ω; furthermore this parameter ω is *defined* as the limit (as $n \to \infty$) of some function of the observations. Available information about the value of ω in prevailing conditions C is *necessarily* described by *some* probability distribution $\pi(\omega|C)$.

For example, in the case of a sequence $\{x_1, x_2, \ldots\}$ of dichotomous exchangeable random quantities $x_j \in \{0, 1\}$, de Finetti's representation theorem establishes that the joint distribution of $(x_1, \ldots, x_n)$ has an *integral representation* of the form

$$(5) \quad p(x_1, \ldots, x_n|C) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}\, \pi(\theta|C)\, d\theta, \quad \theta = \lim_{n\to\infty} \frac{r}{n}\,,$$

where $r = \sum x_j$ is the number of positive trials. This is nothing but the joint distribution of a set of (conditionally) independent Bernoulli trials with parameter θ, over which some probability distribution $\pi(\theta|C)$ is therefore proven to exist. More generally, for sequences of arbitrary random quantities $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$, exchangeability leads to integral representations of the form

$$(6) \quad p(\mathbf{x}_1, \ldots, \mathbf{x}_n|C) = \int_{\boldsymbol{\Omega}} \prod_{i=1}^{n} p(\mathbf{x}_i|\omega)\, \pi(\omega|C)\, d\omega,$$

where $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$ denotes some probability *model*, ω is the limit as $n \to \infty$ of some function $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of the observations, and $\pi(\omega|C)$ is some probability distribution over $\boldsymbol{\Omega}$. This formulation includes "nonparametric" (distribution free) modelling, where ω may index, for instance, all continuous probability distributions on $\mathcal{X}$. Notice that $\pi(\omega|C)$ does *not* describe a possible variability of ω (since ω will typically be a fixed *unknown* vector), but a description on the uncertainty associated with its actual value.

Under appropriate conditioning, exchangeability is a very general assumption, a powerful extension of the traditional concept of a *random sample.* Indeed, many statistical analyses directly assume data (or subsets of the data) to be a random sample of conditionally independent observations from some probability model, so that $p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\omega) = \prod_{i=1}^{n} p(\mathbf{x}_i|\omega)$; but *any* random sample is exchangeable, since $\prod_{i=1}^{n} p(\mathbf{x}_i|\omega)$ is obviously invariant under permutations. Notice that the observations in a random sample are only independent *conditional* on the parameter value ω; as nicely put by Lindley, the mantra that the observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in a random sample are independent is ridiculous when they are used to infer $\mathbf{x}_{n+1}$. Notice also that, under exchangeability, the general representation theorem provides an *existence theorem* for a probability distribution $\pi(\omega|C)$ on the parameter space Ω, and that this is an argument which only depends on mathematical probability theory.

Another important consequence of exchangeability is that it provides a formal *definition* of the parameter ω which labels the model as the limit, as $n \to \infty$, of *some* function $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of the observations; the function f obviously depends both on the assumed model and the chosen parametrization. For instance, in the case of a sequence of Bernoulli trials, the parameter θ is *defined* as the limit, as $n \to \infty$, of the relative frequency r/n. It follows that, under exchangeability, the sentence "the true value of ω" has a well-defined meaning, if only asymptotically verifiable. Moreover, if two different models have parameters which are functionally related by their definition, then the corresponding posterior distributions may be meaningfully compared, for they refer to functionally related quantities. For instance, if a finite subset $\{x_1, \ldots, x_n\}$ of an exchangeable sequence of integer observations is assumed to be a random sample from a Poisson distribution $\mathrm{Po}(x|\lambda)$, so that $\mathrm{E}[x|\lambda] = \lambda$, then λ is *defined* as $\lim_{n\to\infty}\{\bar{x}_n\}$, where $\bar{x}_n = \sum_j x_j/n$; similarly, if for some fixed non-zero integer r, the same data are assumed to be a random sample for a negative binomial $\mathrm{Nb}(x|r, \theta)$, so that $\mathrm{E}[x|\theta, r] = r(1-\theta)/\theta$, then θ is *defined* as $\lim_{n\to\infty}\{r/(\bar{x}_n + r)\}$. It follows that $\theta \equiv r/(\lambda + r)$ and, hence,

θ and $r/(\lambda + r)$ may be treated as the *same* (unknown) quantity whenever this might be needed as, for example, when comparing the relative merits of these alternative probability models.

3 THE BAYESIAN PARADIGM

The statistical analysis of some observed data D typically begins with some informal *descriptive* evaluation, which is used to suggest a tentative, formal *probability model* $\{p(D|\omega),\ \omega \in \Omega\}$ assumed to represent, for some (unknown) value of ω, the probabilistic mechanism which has generated the observed data D. The arguments outlined in Section 2 establish the logical need to assess a *prior* probability distribution $\pi(\omega|K)$ over the parameter space Ω, describing the available knowledge K about the value of ω prior to the data being observed. It then follows from standard probability theory that, if the probability model is correct, all available information about the value of ω after the data D have been observed is contained in the corresponding *posterior* distribution whose probability density, $\pi(\omega|D, A, K)$, is immediately obtained from Bayes' theorem,

$$(7) \quad \pi(\omega|D, A, K) = \frac{p(D|\omega)\, \pi(\omega|K)}{\int_\Omega p(D|\omega)\, \pi(\omega|K)\, d\omega}\,,$$

where A stands for the assumptions made on the probability model. It is this systematic use of Bayes' theorem to incorporate the information provided by the data that justifies the adjective *Bayesian* by which the paradigm is usually known. It is obvious from Bayes' theorem that any value of ω with zero prior density will have zero posterior density. Thus, it is typically assumed (by appropriate restriction, if necessary, of the *parameter space* Ω) that prior distributions are *strictly positive* (as Savage put it, keep the mind open, or at least ajar). To simplify the presentation, the accepted assumptions A and the available knowledge K are often omitted from the notation, but the fact that *all* statements about ω given D are *also* conditional to A and K should always be kept in mind.

EXAMPLE 1 Bayesian inference with a finite parameter space. Let $p(D|\theta)$, $\theta \in \{\theta_1, \ldots, \theta_m\}$, be the probability mechanism which is assumed to have generated the observed data D, so that θ may only take a *finite* number of values. Using the finite form of Bayes' theorem, and omitting the prevailing conditions from the notation, the posterior probability of θ_i after data D have been observed is

$$(8) \quad \Pr(\theta_i|D) = \frac{p(D|\theta_i)\ \Pr(\theta_i)}{\sum_{j=1}^m p(D|\theta_j)\ \Pr(\theta_j)}\,, \quad i = 1, \ldots, m.$$

For any prior distribution $p(\theta) = \{\Pr(\theta_1), \ldots, \Pr(\theta_m)\}$ describing available knowledge about the value of θ, $\Pr(\theta_i|D)$ measures how likely should θ_i be judged, given both the initial knowledge described by the prior distribution, and the information provided by the data D.

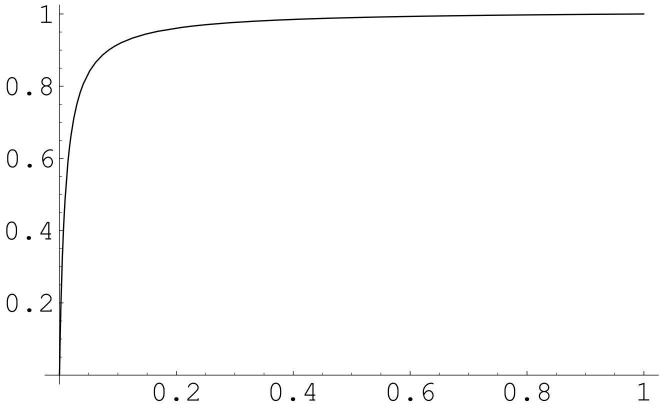

Figure 1. Posterior probability of infection $\Pr(V|+)$ given a positive test, as a function of the prior probability of infection $\Pr(V)$

An important, frequent application of this simple technique is provided by probabilistic diagnosis. For example, consider the simple situation where a particular test designed to detect a virus is known from laboratory research to give a positive result in 98% of infected people and in 1% of non-infected. Then, the posterior probability that a person who tested positive is infected is given by $\Pr(V|+) = (0.98\,p)/\{0.98\,p + 0.01\,(1-p)\}$ as a function of $p = \Pr(V)$, the prior probability of a person being infected (the *prevalence* of the infection in the population under study). Figure 1 shows $\Pr(V|+)$ as a function of $\Pr(V)$.

As one would expect, the posterior probability is only zero if the prior probability is zero (so that it is *known* that the population is free of infection) and it is only one if the prior probability is one (so that it is *known* that the population is universally infected). Notice that if the infection is rare, then the posterior probability of a randomly chosen person being infected will be relatively low even if the test is positive. Indeed, for say $\Pr(V) = 0.002$, one finds $\Pr(V|+) = 0.164$, so that in a population where only 0.2% of individuals are infected, only 16.4% of those testing positive within a random sample will actually prove to be infected: most positives would actually be *false* positives.

In this section, we describe in some detail the learning process described by Bayes' theorem, discuss its implementation in the presence of nuisance parameters, show how it can be used to forecast the value of future observations, and analyze its large sample behaviour.

3.1 The Learning Process

In the Bayesian paradigm, the process of learning from the data is systematically implemented by making use of Bayes' theorem to combine the available prior information with the information provided by the data to produce the required posterior distribution. Computation of posterior densities is often facilitated by noting that Bayes' theorem may be simply expressed as

$$(9) \quad \pi(\omega|D) \propto p(D|\omega)\, \pi(\omega),$$

(where $\propto$ stands for 'proportional to' and where, for simplicity, the accepted assumptions A and the available knowledge K have been omitted from the notation), since the missing proportionality constant $[\int_\Omega p(D|\omega)\, \pi(\omega)\, d\omega]^{-1}$ may always be deduced from the fact that $\pi(\omega|D)$, a probability density, must integrate to one. Hence, to identify the form of a posterior distribution it suffices to identify a *kernel* of the corresponding probability density, that is a function $k(\omega)$ such that $\pi(\omega|D) = c(D)\, k(\omega)$ for some $c(D)$ which does not involve ω. In the examples which follow, this technique will often be used.

An *improper prior function* is defined as a positive function $\pi(\omega)$ such that $\int_\Omega \pi(\omega)\, d\omega$ is not finite. Equation (9), the formal expression of Bayes' theorem, remains technically valid if $\pi(\omega)$ is an improper prior function provided that $\int_\Omega p(D|\omega)\, \pi(\omega)\, d\omega < \infty$, thus leading to a well defined *proper* posterior density $\pi(\omega|D) \propto p(D|\omega)\, \pi(\omega)$. In particular, as will later be justified (Section 4) it also remains philosophically valid if $\pi(\omega)$ is an appropriately chosen *reference* (typically improper) prior function.

Considered as a function of ω, $l(\omega, D) = p(D|\omega)$ is often referred to as the *likelihood function*. Thus, Bayes' theorem is simply expressed in words by the statement that *the posterior is proportional to the likelihood times the prior*. It follows from equation (9) that, provided the *same* prior $\pi(\omega)$ is used, two different data sets D_1 and D_2, with possibly different probability models $p_1(D_1|\omega)$ and $p_2(D_2|\omega)$ but yielding *proportional* likelihood functions, will produce identical posterior distributions for ω. This immediate consequence of Bayes theorem has been proposed as a principle on its own, the *likelihood principle*, and it is seen by many as an obvious requirement for reasonable statistical inference. In particular, for any given prior $\pi(\omega)$, the posterior distribution does not depend on the set of possible data values, or the *sample space*. Notice, however, that the likelihood principle only applies to inferences about the parameter vector ω once the data have been obtained. Consideration of the sample space is essential, for instance, in model criticism, in the design of experiments, in the derivation of predictive distributions, and in the construction of objective Bayesian procedures.

Naturally, the terms prior and posterior are only *relative* to a particular set of data. As one would expect from the coherence induced by probability theory, if data $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are sequentially presented, the final result will be the same whether data are globally or sequentially processed. Indeed, $\pi(\omega|\mathbf{x}_1, \ldots, \mathbf{x}_{i+1}) \propto p(\mathbf{x}_{i+1}|\omega)\, \pi(\omega|\mathbf{x}_1, \ldots, \mathbf{x}_i)$, for $i = 1, \ldots, n-1$, so that the "posterior" at a given stage becomes the "prior" at the next.

In most situations, the posterior distribution is "sharper" than the prior so that, in most cases, the density $\pi(\omega|\mathbf{x}_1, \ldots, \mathbf{x}_{i+1})$ will be more concentrated around the true value of ω than $\pi(\omega|\mathbf{x}_1, \ldots, \mathbf{x}_i)$. However, this is not always the case: occasionally, a “surprising” observation will increase, rather than decrease, the uncertainty about the value of ω. For instance, in probabilistic diagnosis, a sharp posterior probability distribution (over the possible causes $\{\omega_1, \ldots, \omega_k\}$ of a syndrome) describing, a “clear” diagnosis of disease ω_i (that is, a posterior with a large probability for ω_i) would typically update to a less concentrated posterior probability distribution over $\{\omega_1, \ldots, \omega_k\}$ if a new clinical analysis yielded data which were unlikely under ω_i.

For a given probability model, one may find that a particular function of the data $\mathbf{t} = \mathbf{t}(D)$ is a *sufficient* statistic in the sense that, given the model, $\mathbf{t}(D)$ contains all information about ω which is available in D. Formally, $\mathbf{t} = \mathbf{t}(D)$ is sufficient if (and only if) there exist nonnegative functions f and g such that the likelihood function may be factorized in the form $p(D|\omega) = f(\omega, \mathbf{t})g(D)$. A sufficient statistic always exists, for $\mathbf{t}(D) = D$ is obviously sufficient; however, a much simpler sufficient statistic, with a fixed dimensionality which is independent of the sample size, often exists. In fact this is known to be the case whenever the probability model belongs to the *generalized exponential family*, which includes many of the more frequently used probability models. It is easily established that if $\mathbf{t}$ is sufficient, the posterior distribution of ω only depends on the data D through $\mathbf{t}(D)$, and may be directly computed in terms of $p(\mathbf{t}|\omega)$, so that, $\pi(\omega|D) = p(\omega|\mathbf{t}) \propto p(\mathbf{t}|\omega)\,\pi(\omega)$.

Naturally, for fixed data and model assumptions, different priors lead to different posteriors. Indeed, Bayes’ theorem may be described as a data-driven probability transformation machine which maps prior distributions (describing prior knowledge) into posterior distributions (representing combined prior and data knowledge). It is important to analyze whether or not sensible changes in the prior would induce noticeable changes in the posterior. Posterior distributions based on reference “noninformative” priors play a central role in this *sensitivity analysis* context. Investigation of the sensitivity of the posterior to changes in the prior is an important ingredient of the comprehensive analysis of the sensitivity of the final results to *all* accepted assumptions which any responsible statistical study should contain.

EXAMPLE 2 Inference on a binomial parameter. If the data D consist of n Bernoulli observations with parameter θ which contain r positive trials, then $p(D|\theta, n) = \theta^r(1-\theta)^{n-r}$, so that $\mathbf{t}(D) = \{r, n\}$ is sufficient. Suppose that prior knowledge about θ is described by a Beta distribution $\mathrm{Be}(\theta|\alpha, \beta)$, so that $\pi(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$. Using Bayes’ theorem, the posterior density of θ is $\pi(\theta|r, n, \alpha, \beta) \propto \theta^r(1-\theta)^{n-r}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}$, the Beta distribution $\mathrm{Be}(\theta|r+\alpha, n-r+\beta)$.

Suppose, for example, that in the light of precedent surveys, available information on the proportion θ of citizens who would vote for a particular political measure in a referendum is described by a Beta distribution $\mathrm{Be}(\theta|50, 50)$, so that it is judged to be equally likely that the referendum would be won or lost, and it is judged that the probability that either side wins less than 60% of the vote is 0.95.

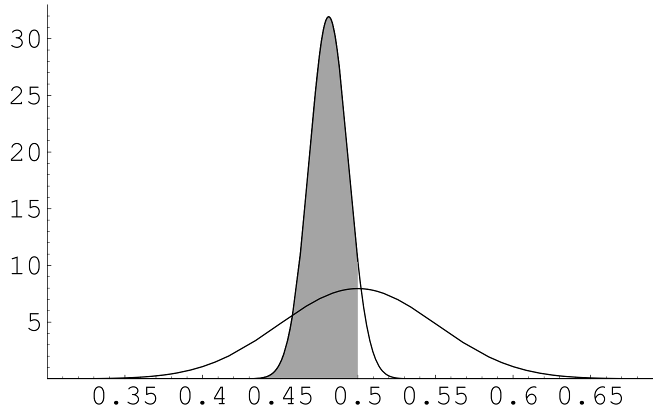

Figure 2. Prior and posterior densities of the proportion θ of citizens that would vote in favour of a referendum

A random survey of size 1500 is then conducted, where only 720 citizens declare to be in favour of the proposed measure. Using the results above, the corresponding posterior distribution is then $\text{Be}(\theta|770, 830)$. These prior and posterior densities are plotted in Figure 2; it may be appreciated that, as one would expect, the effect of the data is to drastically reduce the initial uncertainty on the value of θ and, hence, on the referendum outcome. More precisely, $\Pr(\theta < 0.5|720, 1500, H, K) = 0.933$ (shaded region in Figure 2) so that, after the information from the survey has been included, the probability that the referendum will be lost should be judged to be about 93%.

The general situation where the vector of interest is not the whole parameter vector ω, but some function $\theta = \theta(\omega)$ of possibly lower dimension than ω, will now be considered. Let D be some observed data, let $\{p(D|\omega), \omega \in \mathbf{\Omega}\}$ be a probability model assumed to describe the probability mechanism which has generated D, let $\pi(\omega)$ be a probability distribution describing any available information on the value of ω, and let $\theta = \theta(\omega) \in \Theta$ be a function of the original parameters over whose value inferences based on the data D are required. Any valid conclusion on the value of the *vector of interest* θ will then be contained in its posterior probability distribution $\pi(\theta|D)$ which is conditional on the observed data D and will naturally also depend, although not explicitly shown in the notation, on the assumed model $\{p(D|\omega), \omega \in \mathbf{\Omega}\}$, and on the available prior information encapsulated by $\pi(\omega)$. The required posterior distribution $p(\theta|D)$ is found by standard use of probability calculus. Indeed, by Bayes' theorem, $\pi(\omega|D) \propto p(D|\omega)\, \pi(\omega)$. Moreover, let $\lambda = \lambda(\omega) \in \Lambda$ be some other function of the original parameters such that $\psi = \{\theta, \lambda\}$ is a one-to-one transformation of ω, and let $J(\omega) = (\partial\psi/\partial\omega)$ be the corresponding Jacobian matrix. Naturally, the introduction of λ is not necessary if $\theta(\omega)$ is a one-to-one transformation of ω. Using standard change-of-variable probability techniques, the posterior density of ψ is

$$(10)\quad \pi(\psi|D) = \pi(\theta, \lambda|D) = \left[\frac{\pi(\omega|D)}{|J(\omega)|}\right]_{\omega=\omega(\psi)}$$

and the required posterior of θ is the appropriate *marginal* density, obtained by integration over the *nuisance parameter* λ,

$$(11)\quad \pi(\theta|D) = \int_\Lambda \pi(\theta, \lambda|D)\, d\lambda.$$

Notice that elimination of unwanted nuisance parameters, a simple integration within the Bayesian paradigm is, however, a difficult (often polemic) problem for frequentist statistics.

Sometimes, the range of possible values of ω is effectively restricted by contextual considerations. If ω is known to belong to $\Omega_c \subset \Omega$, the prior distribution is only positive in Ω_c and, using Bayes' theorem, it is immediately found that the restricted posterior is

$$(12)\quad \pi(\omega|D, \omega \in \Omega_c) = \frac{\pi(\omega|D)}{\int_{\Omega_c} \pi(\omega|D)}, \quad \omega \in \Omega_c,$$

and obviously vanishes if $\omega \notin \Omega_c$. Thus, to incorporate a restriction on the possible values of the parameters, it suffices to *renormalize* the unrestricted posterior distribution to the set $\Omega_c \subset \Omega$ of parameter values which satisfy the required condition. Incorporation of known constraints on the parameter values, a simple renormalization within the Bayesian pardigm, is another very difficult problem for conventional statistics. For further details on the elimination of nuisance parameters see [Liseo, 2005].

EXAMPLE 3 Inference on normal parameters. Let $D = \{x_1, \dots x_n\}$ be a random sample from a normal distribution $N(x|\mu, \sigma)$. The corresponding likelihood function is immediately found to be proportional to $\sigma^{-n} \exp[-n\{s^2 + (\bar{x} - \mu)^2\}/(2\sigma^2)]$, with $n\bar{x} = \sum_i x_i$, and $ns^2 = \sum_i (x_i - \bar{x})^2$. It may be shown (see Section 4) that absence of initial information on the value of both μ and σ may formally be described by a joint prior function which is uniform in both μ and $\log(\sigma)$, that is, by the (improper) prior function $\pi(\mu, \sigma) = \sigma^{-1}$. Using Bayes' theorem, the corresponding joint posterior is

$$(13)\quad \pi(\mu, \sigma|D) \propto \sigma^{-(n+1)} \exp[-n\{s^2 + (\bar{x} - \mu)^2\}/(2\sigma^2)].$$

Thus, using the Gamma integral in terms of $\lambda = \sigma^{-2}$ to integrate out σ,

$$(14)\quad \pi(\mu|D) \propto \int_0^\infty \sigma^{-(n+1)} \exp\left[-\frac{n}{2\sigma^2}[s^2 + (\bar{x} - \mu)^2]\right] d\sigma \propto [s^2 + (\bar{x} - \mu)^2]^{-n/2},$$

which is recognized as a kernel of the Student density $\mathrm{St}(\mu|\bar{x}, s/\sqrt{n-1}, n-1)$. Similarly, integrating out μ,

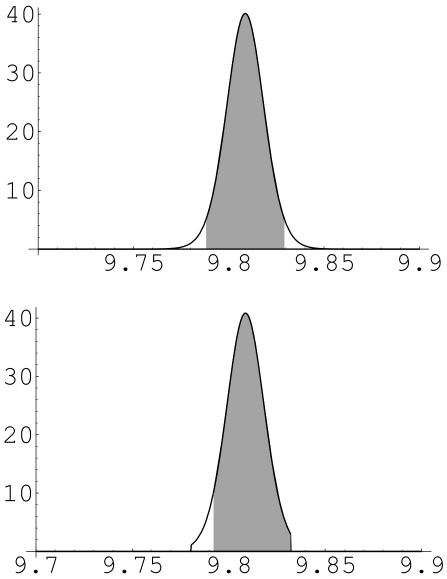

Figure 3. Posterior density $\pi(g|m, s, n)$ of the value g of the gravitational field, given $n = 20$ normal measurements with mean $m = 9.8087$ and standard deviation $s = 0.0428$, (a) with no additional information, and (b) with g restricted to $G_c = \{g;\ 9.7803 < g < 9.8322\}$. Shaded areas represent 95%-credible regions of g

$$(15)\quad \pi(\sigma|D) \propto \int_{-\infty}^{\infty} \sigma^{-(n+1)} \exp\left[-\frac{n}{2\sigma^2}[s^2 + (\bar{x} - \mu)^2]\right] d\mu \propto \sigma^{-n} \exp\left[-\frac{ns^2}{2\sigma^2}\right].$$

Changing variables to the precision $\lambda = \sigma^{-2}$ results in $\pi(\lambda|D) \propto \lambda^{(n-3)/2} e^{ns^2\lambda/2}$, a kernel of the Gamma density $\text{Ga}(\lambda|(n-1)/2, ns^2/2)$. In terms of the standard deviation σ this becomes $\pi(\sigma|D) = p(\lambda|D)|\partial\lambda/\partial\sigma| = 2\sigma^{-3}\text{Ga}(\sigma^{-2}|(n-1)/2, ns^2/2)$, a square-root inverted gamma density.

A frequent example of this scenario is provided by laboratory measurements made in conditions where central limit conditions apply, so that (assuming no experimental bias) those measurements may be treated as a random sample from a normal distribution centered at the quantity μ which is being measured, and with some (unknown) standard deviation σ. Suppose, for example, that in an elementary physics classroom experiment to measure the gravitational field g with a pendulum, a student has obtained $n = 20$ measurements of g yielding (in m/sec^2) a mean $\bar{x} = 9.8087$, and a standard deviation $s = 0.0428$. Using no other information, the corresponding posterior distribution is $\pi(g|D) = \text{St}(g|9.8087, 0.0098, 19)$ represented in Figure 3(a). In particular, $\Pr(9.788 < g < 9.829|D) = 0.95$, so that, with the information provided by this experiment, the gravitational field at the location of the laboratory may be expected to lie between 9.788 and 9.829 with probability 0.95.

Formally, the posterior distribution of g should be restricted to $g > 0$; however, as immediately obvious from Figure 3a, this would not have any appreciable effect, due to the fact that the likelihood function is actually concentrated on positive g values.

Suppose now that the student is further instructed to incorporate into the analysis the fact that the value of the gravitational field g at the laboratory is known to lie between 9.7803 m/sec^2 (average value at the Equator) and 9.8322 m/sec^2 (average value at the poles). The updated posterior distribution will the be

$$(16)\quad \pi(g|D, g \in G_c) = \frac{\mathrm{St}(g|m, s/\sqrt{n-1}, n)}{\int_{g \in G_c} \mathrm{St}(g|m, s/\sqrt{n-1}, n)}, \quad g \in G_c,$$

represented in Figure 3(b), where $G_c = \{g;\, 9.7803 < g < 9.8322\}$. One-dimensional numerical integration may be used to verify that $\Pr(g > 9.792|D, g \in G_c) = 0.95$. Moreover, if inferences about the standard deviation σ of the measurement procedure are also requested, the corresponding posterior distribution is found to be $\pi(\sigma|D) = 2\sigma^{-3}\mathrm{Ga}(\sigma^{-2}|9.5, 0.0183)$. This has a mean $\mathrm{E}[\sigma|D] = 0.0458$ and yields $\Pr(0.0334 < \sigma < 0.0642|D) = 0.95$.

3.2 Predictive Distributions

Let $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathcal{X}$, be a set of exchangeable observations, and consider now a situation where it is desired to predict the value of a future observation $\mathbf{x} \in \mathcal{X}$ generated by the same random mechanism that has generated the data D. It follows from the foundations arguments discussed in Section 2 that the solution to this prediction problem is simply encapsulated by the *predictive* distribution $p(\mathbf{x}|D)$ describing the uncertainty on the value that $\mathbf{x}$ will take, given the information provided by D and any other available knowledge. Suppose that contextual information suggests the assumption that data D may be considered to be a random sample from a distribution in the family $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$, and let $\pi(\omega)$ be a prior distribution describing available information on the value of ω. Since $p(\mathbf{x}|\omega, D) = p(\mathbf{x}|\omega)$, it then follows from standard probability theory that

$$(17)\quad p(\mathbf{x}|D) = \int_\Omega p(\mathbf{x}|\omega)\, \pi(\omega|D)\, d\omega,$$

which is an average of the probability distributions of $\mathbf{x}$ conditional on the (unknown) value of ω, weighted with the posterior distribution of ω given D.

If the assumptions on the probability model are correct, the posterior predictive distribution $p(\mathbf{x}|D)$ will converge, as the sample size increases, to the distribution $p(\mathbf{x}|\omega)$ which has generated the data. Indeed, the best technique to assess the quality of the inferences about ω encapsulated in $\pi(\omega|D)$ is to check against the observed data the predictive distribution $p(\mathbf{x}|D)$ generated by $\pi(\omega|D)$. For a good introduction to Bayesian predictive inference, see Geisser [1993].

EXAMPLE 4 Prediction in a Poisson process. Let $D = \{r_1, \ldots, r_n\}$ be a random sample from a Poisson distribution $\mathrm{Pn}(r|\lambda)$ with parameter λ, so that $p(D|\lambda) \propto \lambda^t e^{-\lambda n}$, where $t = \sum r_i$. It may be shown (see Section 4) that absence of initial information on the value of λ may be formally described by the (improper) prior function $\pi(\lambda) = \lambda^{-1/2}$. Using Bayes' theorem, the corresponding posterior is

$$(18)\quad \pi(\lambda|D) \propto \lambda^t e^{-\lambda n}\, \lambda^{-1/2} \propto \lambda^{t-1/2} e^{-\lambda n},$$

the kernel of a Gamma density $\mathrm{Ga}(\lambda|, t + 1/2, n)$, with mean $(t + 1/2)/n$. The corresponding predictive distribution is the Poisson-Gamma mixture

$$(19)\quad p(r|D) = \int_0^\infty \mathrm{Pn}(r|\lambda)\, \mathrm{Ga}(\lambda|, t + \frac{1}{2}, n)\, d\lambda = \frac{n^{t+1/2}}{\Gamma(t+1/2)}\, \frac{1}{r!}\, \frac{\Gamma(r+t+1/2)}{(1+n)^{r+t+1/2}}\,.$$

Suppose, for example, that in a firm producing automobile restraint systems, the entire production in each of 10 consecutive months has yielded no complaint from their clients. With no additional information on the average number λ of complaints per month, the quality assurance department of the firm may report that the probabilities that r complaints will be received in the next month of production are given by equation (19), with $t = 0$ and $n = 10$. In particular, $p(r = 0|D) = 0.953$, $p(r = 1|D) = 0.043$, and $p(r = 2|D) = 0.003$. Many other situations may be described with the same model. For instance, if metereological conditions remain similar in a given area, $p(r = 0|D) = 0.953$ would describe the chances of no flash flood next year, given 10 years without flash floods in the area.

EXAMPLE 5 Prediction in a Normal process. Consider now prediction of a continuous variable. Let $D = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $N(x|\mu, \sigma)$. As mentioned in Example 3, absence of initial information on the values of both μ and σ is formally described by the *improper* prior function $\pi(\mu, \sigma) = \sigma^{-1}$, and this leads to the joint posterior density (13). The corresponding (posterior) predictive distribution is

$$(20)\quad p(x|D) = \int_0^\infty \int_{-\infty}^\infty \mathbf{N}(x|\mu, \sigma)\, \pi(\mu, \sigma|D)\, d\mu d\sigma = \mathrm{St}(x|\bar{x}, s\sqrt{\frac{n+1}{n-1}}, n-1).$$

If μ is known to be positive, the appropriate prior function will be the restricted function

$$(21)\quad \pi(\mu, \sigma) = \begin{cases} \sigma^{-1} & \text{if } \mu > 0 \\ 0 & \text{otherwise.} \end{cases}$$

However, the result in equation (19) will still hold, provided the likelihood function $p(D|\mu, \sigma)$ is concentrated on positive μ values. Suppose, for example, that in the firm producing automobile restraint systems, the observed breaking strengths of $n = 10$ randomly chosen safety belt webbings have mean $\bar{x} = 28.011$ kN and standard deviation $s = 0.443$ kN, and that the relevant engineering specification requires breaking strengths to be larger than 26 kN. If data may truly be assumed to be a random sample from a normal distribution, the likelihood function is only appreciable for positive μ values, and only the information provided by this small sample is to be used, then the quality engineer may claim that the probability that a safety belt randomly chosen from the same batch as the sample tested would satisfy the required specification is $\Pr(x > 26 \mid D) = 0.9987$. Besides, if production conditions remain constant, 99.87% of the safety belt webbings may be expected to have acceptable breaking strengths.

3.3 Asymptotic Behaviour

The behaviour of posterior distributions when the sample size is large is now considered. This is important for, at least, two different reasons: (i) asymptotic results provide useful first-order approximations when actual samples are relatively large, and (ii) objective Bayesian methods typically depend on the asymptotic properties of the assumed model. Let $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x} \in \mathcal{X}$, be a random sample of size n from $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$. It may be shown that, as $n \to \infty$, the posterior distribution of a *discrete* parameter ω typically converges to a degenerate distribution which gives probability one to the true value of ω, and that the posterior distribution of a *continuous* parameter ω typically converges to a normal distribution centered at its *maximum likelihood estimate* $\hat{\omega}$ (MLE), with a variance matrix which decreases with n as $1/n$.

Consider first the situation where $\Omega = \{\omega_1, \omega_2, \ldots\}$ consists of a *countable* (possibly infinite) set of values, such that the probability model which corresponds to the true parameter value ω_t is *distinguishable* from the others in the sense that the logarithmic divergence $\kappa\{p(\mathbf{x}|\omega_i) | p(\mathbf{x}|\omega_t)\}$ of each of the $p(\mathbf{x}|\omega_i)$ from $p(\mathbf{x}|\omega_t)$ is strictly positive. Taking logarithms in Bayes' theorem, defining $z_j = \log[p(\mathbf{x}_j|\omega_i)/p(\mathbf{x}_j|\omega_t)]$, $j = 1, \ldots, n$, and using the strong law of large numbers on the n conditionally independent and identically distributed random quantities $z_1, \ldots, z_n$, it may be shown that

$$(22)\quad \lim_{n\to\infty} \Pr(\omega_t|\mathbf{x}_1, \ldots, \mathbf{x}_n) = 1, \qquad \lim_{n\to\infty} \Pr(\omega_i|\mathbf{x}_1, \ldots, \mathbf{x}_n) = 0, \quad i \neq t.$$

Thus, under appropriate regularity conditions, the posterior probability of the true parameter value converges to one as the sample size grows.

Consider now the situation where ω is a k-dimensional *continuous* parameter. Expressing Bayes' theorem as $\pi(\omega|\mathbf{x}_1, \ldots, \mathbf{x}_n) \propto \exp\{\log[\pi(\omega)] + \sum_{j=1}^{n} \log[p(\mathbf{x}_j|\omega)]\}$, expanding $\sum_j \log[p(\mathbf{x}_j|\omega)]$ about its maximum (the MLE $\hat{\omega}$), and assuming regularity conditions (to ensure that terms of order higher than quadratic may be ignored and that the sum of the terms from the likelihood will dominate the term from the prior) it is found that the posterior density of ω is the approximate k-variate normal

$$(23)\quad \pi(\omega|\mathbf{x}_1, \ldots, \mathbf{x}_n) \approx \mathbf{N}_k\{\hat{\omega}, \mathbf{S}(D, \hat{\omega})\}, \qquad \mathbf{S}^{-1}(D, \omega) = \left(-\sum_{l=1}^{n} \frac{\partial^2 \log[p(\mathbf{x}_l|\omega)]}{\partial\omega_i \partial\omega_j} \right).$$

A simpler, but somewhat poorer, approximation may be obtained by using the strong law of large numbers on the sums in (22) to establish that $\mathbf{S}^{-1}(D,\hat\omega) \approx n\,\mathbf{F}(\hat\omega)$, where $\mathbf{F}(\omega)$ is *Fisher's information matrix*, with general element

$$(24)\;\; \mathbf{F}_{ij}(\omega) = -\int_X p(\mathbf{x}|\omega)\; \frac{\partial^2 \log[p(\mathbf{x}|\omega)]}{\partial\omega_i \partial\omega_j}\, d\mathbf{x},$$

so that

$$(25)\;\; \pi(\omega|\mathbf{x}_1,\ldots,\mathbf{x}_n) \approx N_k(\omega|\hat\omega, n^{-1}\,\mathbf{F}^{-1}(\hat\omega)).$$

Thus, under appropriate regularity conditions, the posterior probability density of the parameter vector ω approaches, as the sample size grows, a multivarite normal density centered at the MLE $\hat\omega$, with a variance matrix which decreases with n as n^{-1} .

EXAMPLE 2, continued. Asymptotic approximation with binomial data. Let $D = (x_1,\ldots,x_n)$ consist of n independent Bernoulli trials with parameter θ, so that $p(D|\theta,n) = \theta^r(1-\theta)^{n-r}$. This likelihood function is maximized at $\hat\theta = r/n$, and Fisher's information function is $F(\theta) = \theta^{-1}(1-\theta)^{-1}$. Thus, using the results above, the posterior distribution of θ will be the approximate normal,

$$(26)\;\; \pi(\theta|r,n) \approx \mathbf{N}(\theta|\hat\theta, s(\hat\theta)/\sqrt{n}), \quad s(\theta) = \{\theta(1-\theta)\}^{1/2}$$

with mean $\hat\theta = r/n$ and variance $\hat\theta(1-\hat\theta)/n$. This will provide a reasonable approximation to the exact posterior if (i) the prior $\pi(\theta)$ is relatively "flat" in the region where the likelihood function matters, and (ii) both r and n are moderately large. If, say, $n = 1500$ and $r = 720$, this leads to $\pi(\theta|D) \approx \mathbf{N}(\theta|0.480, 0.013)$, and to $\Pr(\theta > 0.5|D) \approx 0.940$, which may be compared with the exact value $\Pr(\theta > 0.5|D) = 0.933$ obtained from the posterior distribution which corresponds to the prior $\mathrm{Be}(\theta|50,50)$. $\triangleleft$

It follows from the *joint* posterior asymptotic behaviour of ω and from the properties of the multivariate normal distribution that, if the parameter vector is decomposed into $\omega = (\theta,\lambda)$, and Fisher's information matrix is correspondingly partitioned, so that

$$(27)\;\; \mathbf{F}(\omega) = \mathbf{F}(\theta,\lambda) = (\,\mathbf{F}_{\theta\theta}(\theta,\lambda) \quad \mathbf{F}_{\theta\lambda}(\theta,\lambda)\mathbf{F}_{\lambda\theta}(\theta,\lambda) \quad \mathbf{F}_{\lambda\lambda}(\theta,\lambda)\,)$$

and

$$(28)\;\; \mathbf{S}(\theta,\lambda) = \mathbf{F}^{-1}(\theta,\lambda) = (\,\mathbf{S}_{\theta\theta}(\theta,\lambda) \quad \mathbf{S}_{\theta\lambda}(\theta,\lambda)\mathbf{S}_{\lambda\theta}(\theta,\lambda) \quad \mathbf{S}_{\lambda\lambda}(\theta,\lambda)\,)\;,$$

then the *marginal* posterior distribution of θ will be

$$(29)\;\; \pi(\theta|D) \approx \mathbf{N}\{\theta|\hat\theta,\, n^{-1}\,\mathbf{S}_{\theta\theta}(\hat\theta,\hat\lambda)\},$$

while the *conditional* posterior distribution of λ given θ will be

$$(30)\;\; \pi(\lambda|\theta,D) \approx \mathbf{N}\{\lambda|\hat\lambda - \mathbf{F}^{-1}_{\lambda\lambda}(\theta,\hat\lambda)\mathbf{F}_{\lambda\theta}(\theta,\hat\lambda)(\hat\theta-\theta),\, n^{-1}\,\mathbf{F}^{-1}_{\lambda\lambda}(\theta,\hat\lambda)\}.$$

Notice that $\mathbf{F}_{\lambda\lambda}^{-1} = \mathbf{S}_{\lambda\lambda}$ if (and only if) $\mathbf{F}$ is block diagonal, i.e. if (and only if) θ and λ are asymptotically independent.

EXAMPLE 3, continued. Asymptotic approximation with normal data. Let $D = (x_1, \ldots, x_n)$ be a random sample from a normal distribution $\mathbf{N}(x|\mu, \sigma)$. The corresponding likelihood function $p(D|\mu, \sigma)$ is maximized at $(\hat{\mu}, \hat{\sigma}) = (\bar{x}, s)$, and Fisher's information matrix is diagonal, with $F_{\mu\mu} = \sigma^{-2}$. Hence, the posterior distribution of μ is *approximately* $\mathbf{N}(\mu|\bar{x}, s/\sqrt{n})$; this may be compared with the *exact* result $\pi(\mu|D) = \mathrm{St}(\mu|\bar{x}, s/\sqrt{n-1}, n-1)$ obtained previously under the assumption of no prior knowledge. ◁

4 REFERENCE ANALYSIS

Under the Bayesian paradigm, the outcome of any inference problem (the posterior distribution of the quantity of interest) combines the information provided by the data with relevant available prior information. In many situations, however, either the available prior information on the quantity of interest is too vague to warrant the effort required to have it formalized in the form of a probability distribution, or it is too subjective to be useful in scientific communication or public decision making. It is therefore important to be able to identify the mathematical form of a "noninformative" prior, a prior that would have a minimal effect, relative to the data, on the posterior inference. More formally, suppose that the probability mechanism which has generated the available data D is assumed to be $p(D|\omega)$, for some $\omega \in \Omega$, and that the quantity of interest is some real-valued function $\theta = \theta(\omega)$ of the model parameter ω. Without loss of generality, it may be assumed that the probability model is of the form

$$(31)\quad \mathcal{M} = \{p(D|\theta, \lambda), D \in \mathcal{D}, \theta \in \Theta, \lambda \in \Lambda\}$$

$p(D|\theta, \lambda)$, where λ is some appropriately chosen nuisance parameter vector. As described in Section 3, to obtain the required posterior distribution of the quantity of interest $\pi(\theta|D)$ it is necessary to specify a *joint* prior $\pi(\theta, \lambda)$. It is now required to identify the form of that joint prior $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P})$, the *$\theta$-reference prior*, which would have a *minimal effect* on the corresponding posterior distribution of θ,

$$(32)\quad \pi(\theta|D) \propto \int_\Lambda p(D|\theta, \lambda)\, \pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P})\, d\lambda,$$

within the class $\mathcal{P}$ of all the prior disributions compatible with whatever information about (θ, λ) one is prepared to assume, which may just be the class $\mathcal{P}_0$ of *all* strictly positive priors. To simplify the notation, when there is no danger of confusion the reference prior $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P})$ is often simply denoted by $\pi(\theta, \lambda)$, but its dependence on the quantity of interest θ, the assumed model $\mathcal{M}$ and the class $\mathcal{P}$ of priors compatible with assumed knowledge, should always be kept in mind.

To use a conventional expression, the reference prior "would let the data speak for themselves" about the likely value of θ. Properly defined, reference *posterior* distributions have an important role to play in scientific communication, for they provide the answer to a central question in the sciences: conditional on the assumed model $p(D|\theta, \lambda)$, and on any further assumptions of the value of θ on which there might be universal agreement, the reference posterior $\pi(\theta|D)$ should specify what *could* be said about θ *if* the only available information about θ were some well-documented data D and whatever information (if any) one is prepared to assume by restricting the prior to belong to an appropriate class $\mathcal{P}$.

Much work has been done to formulate "reference" priors which would make the idea described above mathematically precise. For historical details, see [Bernardo and Smith, 1994, Sec. 5.6.2; Kass and Wasserman, 1996; Bernardo, 2005a] and references therein. This section concentrates on an approach that is based on information theory to derive reference distributions which may be argued to provide the most advanced general procedure available; this was initiated by Bernardo [1979b; 1981] and further developed by Berger and Bernardo [1989; 1992a; 1982b; 1982c; 1997; 2005a; Bernardo and Ramón, 1998; Berger *et al.*, 2009], and references therein. In the formulation described below, far from ignoring prior knowledge, the reference posterior exploits certain well-defined features of a *possible* prior, namely those describing a situation were relevant knowledge about the quantity of interest (beyond that universally accepted, as specified by the choice of $\mathcal{P}$) may be held to be negligible compared to the information about that quantity which repeated experimentation (from a specific data generating mechanism $\mathcal{M}$) might possibly provide. Reference analysis is appropriate in contexts where the set of inferences which could be drawn in this *possible* situation is considered to be pertinent.

Any statistical analysis contains a fair number of subjective elements; these include (among others) the data selected, the model assumptions, and the choice of the quantities of interest. Reference analysis may be argued to provide an "objective" Bayesian solution to statistical inference problems in just the same sense that conventional statistical methods claim to be "objective": in that the solutions only depend on model assumptions and observed data.

4.1 Reference Distributions

One parameter. Consider the experiment which consists of the observation of data D, generated by a random mechanism $p(D|\theta)$ which only depends on a real-valued parameter $\theta \in \Theta$, and let $\mathbf{t} = \mathbf{t}(D) \in T$ be *any* sufficient statistic (which may well be the complete data set D). In Shannon's general information theory, the *amount of information* $I^\theta\{T, \pi(\theta)\}$ which may be expected to be provided by D, or (equivalently) by $\mathbf{t}(D)$, about the value of θ is defined by

$$(33)\quad I^\theta\{T, \pi(\theta)\} = \kappa\,\{p(\mathbf{t})\pi(\theta)|p(\mathbf{t}|\theta)\pi(\theta)\} = \mathrm{E}_{\mathbf{t}}\left[\int_\Theta \pi(\theta|\mathbf{t}) \log \frac{\pi(\theta|\mathbf{t})}{\pi(\theta)}\, d\theta\right],$$

the expected logarithmic divergence of the prior from the posterior. This is naturally a *functional* of the prior $\pi(\theta)$: the larger the prior information, the smaller the information which the data may be expected to provide. The functional

$I^\theta\{T, \pi(\theta)\}$ is concave, non-negative, and invariant under one-to-one transformations of θ. Consider now the amount of information $I^\theta\{T^k, \pi(\theta)\}$ about θ which may be expected from the experiment which consists of k conditionally independent replications $\{\mathbf{t}_1, \ldots, \mathbf{t}_k\}$ of the original experiment. As $k \to \infty$, such an experiment would provide any *missing information* about θ which could possibly be obtained within this framework; thus, as $k \to \infty$, the functional $I^\theta\{T^k, \pi(\theta)\}$ will approach the missing information about θ associated with the prior $p(\theta)$. Intuitively, a θ-"noninformative" prior is one which *maximizes the missing information* about θ. Formally, if $\pi_k(\theta)$ denotes the prior density which maximizes $I^\theta\{T^k, \pi(\theta)\}$ in the class $\mathcal{P}$ of s prior distributions which are compatible with accepted assumptions on the value of θ (which may well be the class $\mathcal{P}_0$ of *all* strictly positive proper priors) then the θ-reference prior $\pi(\theta \mid \mathcal{M}, \mathcal{P})$ is the limit as $k \to \infty$ (in a sense to be made precise) of the sequence of priors $\{\pi_k(\theta), k = 1, 2, \ldots\}$.

Notice that this limiting procedure is *not* some kind of asymptotic approximation, but an essential element of the *definition* of a reference prior. In particular, this definition implies that reference distributions only depend on the *asymptotic* behaviour of the assumed probability model, a feature which actually simplifies their actual derivation.

EXAMPLE 6 Maximum entropy. If θ may only take a *finite* number of values, so that the parameter space is $\Theta = \{\theta_1, \ldots, \theta_m\}$ and $\pi(\theta) = \{p_1, \ldots, p_m\}$, with $p_i = \Pr(\theta = \theta_i)$, and there is no topology associated to the parameter space Θ, so that the θ_i's are just labels with no quantitative meaning, then the missing information associated to $\{p_1, \ldots, p_m\}$ reduces to

$$\text{(34)} \quad \lim_{k \to \infty} I^\theta\{T^k, \pi(\theta)\} = H(p_1, \ldots, p_m) = -\sum\nolimits_{i=1}^{m} p_i \log(p_i),$$

that is, the *entropy* of the prior distribution $\{p_1, \ldots, p_m\}$.

Thus, in the non-quantitative finite case, the reference prior $\pi(\theta \mid \mathcal{M}, \mathcal{P})$ is that with *maximum entropy* in the class $\mathcal{P}$ of priors compatible with accepted assumptions. Consequently, the reference prior algorithm contains "maximum entropy" priors as the particular case which obtains when the parameter space is a *finite set of labels*, the *only* case where the original concept of entropy as a measure of uncertainty is unambiguous and well-behaved. In particular, if $\mathcal{P}$ is the class $\mathcal{P}_0$ of *all* priors over $\{\theta_1, \ldots, \theta_m\}$, then the reference prior is the uniform prior over the set of possible θ values, $\pi(\theta \mid \mathcal{M}, \mathcal{P}_0) = \{1/m, \ldots, 1/m\}$.

Formally, the *reference prior function* $\pi(\theta \mid \mathcal{M}, \mathcal{P})$ of a univariate parameter θ is defined to be the limit of the sequence of the proper priors $\pi_k(\theta)$ which maximize $I^\theta\{T^k, \pi(\theta)\}$ in the precise sense that, for any value of the sufficient statistic $\mathbf{t} = \mathbf{t}(D)$, the *reference posterior*, the intrinsic[1] limit $\pi(\theta \mid \mathbf{t})$ of the corresponding sequence of posteriors $\{\pi_k(\theta \mid \mathbf{t})\}$, may be obtained from $\pi(\theta \mid \mathcal{M}, \mathcal{P})$ by formal use of Bayes theorem, so that $\pi(\theta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \theta)\, \pi(\theta \mid \mathcal{M}, \mathcal{P})$.

[1] A sequence $\{\pi_k(\theta \mid \mathbf{t})\}$ of posterior distributions converges intrinsically to a limit $\pi(\theta \mid \mathbf{t})$ if the sequence of expected intrinsic discrepancies $\mathrm{E}_{\mathbf{t}}[\delta\{\pi_k(\theta \mid \mathbf{t}), \pi(\theta \mid \mathbf{t})\}]$ converges to 0, where $\delta\{p, q\} = \min\{k(p \mid q), k(q \mid p)\}$, and $k(p \mid q) = \int_\Theta q(\theta) \log[q(\theta)/p(\theta)] d\theta$. For details, see [Berger *et al.*, 2009].

Reference prior *functions* are often simply called reference priors, even though they are usually *not* probability distributions. They should *not* be considered as expressions of belief, but technical devices to obtain (proper) posterior distributions which are a limiting form of the posteriors which could have been obtained from possible prior beliefs which were relatively uninformative with respect to the quantity of interest when compared with the information which data could provide.

If (i) the sufficient statistic $\mathbf{t} = \mathbf{t}(D)$ is a consistent estimator $\tilde{\theta}$ of a continuous parameter θ, and (ii) the class $\mathcal{P}$ contains *all* strictly positive priors, then the reference prior may be shown to have a simple form in terms of any *asymptotic* approximation to the posterior distribution of θ. Notice that, by construction, an *asymptotic* approximation to the posterior does *not* depend on the prior. Specifically, if the posterior density $\pi(\theta|D)$ has an asymptotic approximation of the form $\pi(\theta|\tilde{\theta}, n)$, the (unrestricted) reference prior is simply

$$(35) \quad \pi(\theta|\mathcal{M}, \mathcal{P}_0) \propto \pi(\theta|\tilde{\theta}, n)\Big|_{\tilde{\theta}=\theta} .$$

One-parameter reference priors are *invariant* under reparametrization; thus, if $\psi = \psi(\theta)$ is a piecewise one-to-one function of θ, then the ψ-reference prior is simply the appropriate probability transformation of the θ-reference prior.

EXAMPLE 7 Jeffreys' prior. If θ is univariate and continuous, and the posterior distribution of θ given $\{x_1 \ldots, x_n\}$ is asymptotically normal with standard deviation $s(\tilde{\theta})/\sqrt{n}$, then, using (34), the reference prior function is $\pi(\theta) \propto s(\theta)^{-1}$. Under regularity conditions (often satisfied in practice, see Section 3.3), the posterior distribution of θ is asymptotically normal with variance $n^{-1}\, F^{-1}(\hat{\theta})$, where $F(\theta)$ is Fisher's information function and $\hat{\theta}$ is the MLE of θ. Hence, the reference prior function in these conditions is $\pi(\theta|\mathcal{M}, \mathcal{P}_0) \propto F(\theta)^{1/2}$, which is known as Jeffreys' prior. It follows that the reference prior algorithm contains Jeffreys' priors as the particular case which obtains when the probability model only depends on a single continuous univariate parameter, there are regularity conditions to guarantee asymptotic normality, and there is no additional information, so that the class of possible priors is the set $\mathcal{P}_0$ of all strictly positive priors over Θ. These are precisely the conditions under which there is general agreement on the use of Jeffreys' prior as a "noninformative" prior.

EXAMPLE 2, continued. Reference prior for a binomial parameter. Let data $D = \{x_1, \ldots, x_n\}$ consist of a sequence of n independent Bernoulli trials, so that $p(x|\theta) = \theta^x(1-\theta)^{1-x}$, $x \in \{0, 1\}$; this is a regular, one-parameter continuous model, whose Fisher's information function is $F(\theta) = \theta^{-1}(1-\theta)^{-1}$. Thus, the reference prior $\pi(\theta)$ is proportional to $\theta^{-1/2}(1-\theta)^{-1/2}$, so that the reference prior is the (proper) Beta distribution $\text{Be}(\theta|1/2, 1/2)$. Since the reference algorithm is invariant under reparametrization, the reference prior of $\phi(\theta) = 2 \arcsin \sqrt{\theta}$ is $\pi(\phi) = \pi(\theta)/|\partial\phi/\partial/\theta| = 1$; thus, the reference prior is *uniform on the variance-stabilizing transformation* $\phi(\theta) = 2\text{arc}\sin\sqrt{\theta}$, a feature generally true under reg-

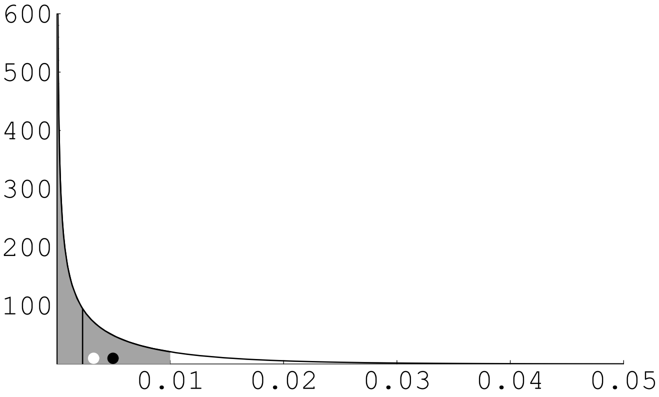

Figure 4. Posterior distribution of the proportion of infected people in the population, given the results of $n = 100$ tests, none of which were positive

ularity conditions. In terms of θ, the reference posterior is $\pi(\theta|D) = \pi(\theta|r, n) = \mathrm{Be}(\theta|r + 1/2, n - r + 1/2)$, where $r = \sum x_j$ is the number of positive trials.

Suppose, for example, that $n = 100$ randomly selected people have been tested for an infection and that all tested negative, so that $r = 0$. The reference posterior distribution of the proportion θ of people infected is then the Beta distribution $\mathrm{Be}(\theta|0.5, 100.5)$, represented in Figure 4. It may well be known that the infection was rare, leading to the assumption that $\theta < \theta_0$, for some upper bound θ_0; the (restricted) reference prior would then be of the form $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$ if $\theta < \theta_0$, and zero otherwise. However, provided the likelihood is concentrated in the region $\theta < \theta_0$, the corresponding posterior would virtually be identical to $\mathrm{Be}(\theta|0.5, 100.5)$. Thus, just on the basis of the observed experimental results, one may claim that the proportion of infected people is surely smaller than 5% (for the reference posterior probability of the event $\theta > 0.05$ is 0.001), that θ is smaller than 0.01 with probability 0.844 (area of the shaded region in Figure 4), that it is equally likely to be over or below 0.23% (for the median, represented by a vertical line, is 0.0023), and that the probability that a person randomly chosen from the population is infected is 0.005 (the posterior mean, represented in the figure by a black circle), since $\Pr(x = 1|r, n) = \mathrm{E}[\theta|r, n] = 0.005$. If a particular point estimate of θ is required (say a number to be quoted in the summary headline) the *intrinsic* estimator suggests itself (see Section 5); this is found to be $\theta^* = 0.0032$ (represented in the figure with a white circle). Notice that the traditional solution to this problem, based on the asymptotic behaviour of the MLE, here $\hat{\theta} = r/n = 0$ for any n, makes absolutely no sense in this scenario. $\triangleleft$

One nuisance parameter. The extension of the reference prior algorithm to the case of two parameters follows the usual mathematical procedure of reducing the problem to a sequential application of the established procedure for the single parameter case. Thus, if the probability model is $p(\mathbf{t}|\theta, \lambda)$, $\theta \in \Theta$, $\lambda \in \Lambda$ and a θ-reference prior $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P})$ is required, the reference algorithm proceeds in two steps:

(i) Conditional on θ, $p(\mathbf{t}|\theta, \lambda)$ only depends on the nuisance parameter λ and, hence, the one-parameter algorithm may be used to obtain the *conditional* reference prior $\pi(\lambda|\theta, \mathcal{M}, \mathcal{P})$.

(ii) If $\pi(\lambda|\theta, \mathcal{M}, \mathcal{P})$ is proper, this may be used to integrate out the nuisance parameter thus obtaining the one-parameter integrated model $p(\mathbf{t}|\theta) = \int_\Lambda p(\mathbf{t}|\theta, \lambda)\, \pi(\lambda|\theta, \mathcal{M}, \mathcal{P})\, d\lambda$, to which the one-parameter algorithm may be applied again to obtain $\pi(\theta|\mathcal{M}, \mathcal{P})$. The θ-reference prior is then $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P}) = \pi(\lambda|\theta, \mathcal{M}, \mathcal{P})\, \pi(\theta|\mathcal{M}, \mathcal{P})$, and the required reference posterior is $\pi(\theta|\mathbf{t}) \propto p(\mathbf{t}|\theta)\, \pi(\theta|\mathcal{M}, \mathcal{P})$.

If the conditional reference prior is *not* proper, then the procedure is performed within an increasing sequence $\{\Lambda_i\}$ of subsets converging to Λ over which $\pi(\lambda|\theta)$ is integrable. This makes it possible to obtain a corresponding sequence of θ-reference posteriors $\{\pi_i(\theta|\mathbf{t}\}$ for the quantity of interest θ, and the required reference posterior is the corresponding intrinsic limit $\pi(\theta|\mathbf{t}) = \lim_i \pi_i(\theta|\mathbf{t})$.

A θ-reference prior is then defined as a positive function $\pi_\theta(\theta, \lambda)$ which may be formally used in Bayes' theorem as a prior to obtain the reference posterior, i.e. such that, for any sufficient $\mathbf{t} \in T$ (which may well be the whole data set D) $\pi(\theta|\mathbf{t}) \propto \int_\Lambda p(\mathbf{t}|\theta, \lambda)\, \pi_\theta(\theta, \lambda)\, d\lambda$. The approximating sequences should be *consistently* chosen within a given model. Thus, given a probability model $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$ an appropriate approximating sequence $\{\Omega_i\}$ should be chosen for the whole parameter space Ω. Thus, if the analysis is done in terms of, say, $\psi = \{\psi_1, \psi_2\} \in \Psi(\Omega)$, the approximating sequence should be chosen such that $\Psi_i = \psi(\Omega_i)$. A natural approximating sequence in location-scale problems is $\{\mu, \log\sigma\} \in [-i, i]^2$.

The θ-reference prior does *not* depend on the choice of the nuisance parameter λ; thus, for any $\psi = \psi(\theta, \lambda)$ such that (θ, ψ) is a one-to-one function of (θ, λ), the θ-reference prior in terms of (θ, ψ) is simply $\pi_\theta(\theta, \psi) = \pi_\theta(\theta, \lambda)/|\partial(\theta, \psi)/\partial(\theta, \lambda)|$, the appropriate probability transformation of the θ-reference prior in terms of (θ, λ). Notice, however, that the reference prior *may* depend on the parameter of interest; thus, the θ-reference prior may differ from the ϕ-reference prior unless either ϕ is a piecewise one-to-one transformation of θ, or ϕ is asymptotically independent of θ. This is an expected consequence of the fact that the conditions under which the missing information about θ is maximized are not generally the same as the conditions which maximize the missing information about an arbitrary function $\phi = \phi(\theta, \lambda)$.

The *non-existence* of a unique "noninformative prior" which would be appropriate for any inference problem within a given model was established by Dawid, Stone and Zidek [1973], when they showed that this is incompatible with *consistent marginalization.* Indeed, if given the model $p(D|\theta, \lambda)$, the reference posterior of the quantity of interest θ, $\pi(\theta|D) = \pi(\theta|\mathbf{t})$, only depends on the data through a statistic $\mathbf{t}$ whose sampling distribution, $p(\mathbf{t}|\theta, \lambda) = p(\mathbf{t}|\theta)$, only depends on θ, one would expect the reference posterior to be of the form $\pi(\theta|\mathbf{t}) \propto \pi(\theta)\, p(\mathbf{t}|\theta)$ for some prior $\pi(\theta)$. However, examples were found where this cannot be the case if a *unique* joint "noninformative" prior were to be used for all possible quantities of interest.

EXAMPLE 8 Regular two dimensional continuous reference prior functions. If the joint posterior distribution of (θ, λ) is asymptotically normal, then the θ-reference prior may be derived in terms of the corresponding Fisher's information matrix, $\mathbf{F}(\theta, \lambda)$. Indeed, if

$$(36)\quad \mathbf{F}(\theta,\lambda) = \begin{pmatrix} F_{\theta\theta}(\theta,\lambda) & F_{\theta\lambda}(\theta,\lambda) \\ F_{\theta\lambda}(\theta,\lambda) & F_{\lambda\lambda}(\theta,\lambda) \end{pmatrix}, \text{ and } \mathbf{S}(\theta,\lambda) = \mathbf{F}^{-1}(\theta,\lambda),$$

then the unrestricted θ-reference prior is $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P}_0) = \pi(\lambda|\theta)\, \pi(\theta)$, where

$$(37)\quad \pi(\lambda|\theta) \propto F_{\lambda\lambda}^{1/2}(\theta,\lambda), \quad \lambda \in \Lambda.$$

If $\pi(\lambda|\theta)$ is proper,

$$(38)\quad \pi(\theta) \propto \exp\big\{ \int_\Lambda \pi(\lambda|\theta)\, \log[S_{\theta\theta}^{-1/2}(\theta,\lambda)]\, d\lambda \big\}, \quad \theta \in \Theta.$$

If $\pi(\lambda|\theta)$ is not proper, integrations are performed on an approximating sequence $\{\Lambda_i\}$ to obtain a sequence $\{\pi_i(\lambda|\theta)\, \pi_i(\theta)\}$, (where $\pi_i(\lambda|\theta)$ is the proper renormalization of $\pi(\lambda|\theta)$ to Λ_i) and the θ-reference prior $\pi_\theta(\theta, \lambda)$ is defined as its appropriate limit. Moreover, if (i) both $F_{\lambda\lambda}^{1/2}(\theta,\lambda)$ and $S_{\theta\theta}^{-1/2}(\theta,\lambda)$ *factorize*, so that

$$(39)\quad S_{\theta\theta}^{-1/2}(\theta,\lambda) \propto f_\theta(\theta)\, g_\theta(\lambda), \quad F_{\lambda\lambda}^{1/2}(\theta,\lambda) \propto f_\lambda(\theta)\, g_\lambda(\lambda),$$

and (ii) the parameters θ and λ are *variation independent*, so that Λ does not depend on θ, then the θ-reference prior is simply $\pi_\theta(\theta, \lambda) = f_\theta(\theta)\, g_\lambda(\lambda)$, even if the conditional reference prior $\pi(\lambda|\theta) = \pi(\lambda) \propto g_\lambda(\lambda)$ (which will not depend on θ) is actually improper.

EXAMPLE 3, continued. Reference priors for the normal model. The information matrix which corresponds to a normal model $\mathbf{N}(x|\mu, \sigma)$ is

$$(40)\quad \mathbf{F}(\mu,\sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}, \quad \mathbf{S}(\mu,\sigma) = \mathbf{F}^{-1}(\mu,\sigma) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix};$$

hence $F_{\sigma\sigma}^{1/2}(\mu,\sigma) = \sqrt{2}\,\sigma^{-1} = f_\sigma(\mu)\, g_\sigma(\sigma)$, with $g_\sigma(\sigma) = \sigma^{-1}$, and thus $\pi(\sigma|\mu) = \sigma^{-1}$. Similarly, $S_{\mu\mu}^{-1/2}(\mu,\sigma) = \sigma^{-1} = f_\mu(\mu)\, g_\mu(\sigma)$, with $f_\mu(\mu) = 1$, and thus $\pi(\mu) = 1$. Therefore, the μ-reference prior is $\pi_\mu(\mu, \sigma|\mathcal{M}, \mathcal{P}_0) = \pi(\sigma|\mu)\, \pi(\mu) = \sigma^{-1}$, as already anticipated. Moreover, as one would expect from the fact that $\mathbf{F}(\mu, \sigma)$ is diagonal and also anticipated, it is similarly found that the σ-reference prior is $\pi_\sigma(\mu, \sigma|\mathcal{M}, \mathcal{P}_0) = \sigma^{-1}$, the same as before.

Suppose, however, that the quantity of interest is *not* the mean μ or the standard deviation σ, but the *standardized* mean $\phi = \mu/\sigma$. Fisher's information matrix in terms of the parameters ϕ and σ is $\mathbf{F}(\phi,\sigma) = J^t\,\mathbf{F}(\mu,\sigma)\,J$, where $J = (\partial(\mu,\sigma)/\partial(\phi,\sigma))$ is the Jacobian of the inverse transformation; this yields

$$(41)\quad \mathbf{F}(\phi,\sigma) = \begin{pmatrix} 1 & \phi\sigma^{-1} \\ \phi\sigma^{-1} & \sigma^{-2}(2+\phi^2) \end{pmatrix}, \quad \mathbf{S}(\phi,\sigma) = \begin{pmatrix} 1+\frac{1}{2}\phi^2 & -\frac{1}{2}\phi\sigma \\ -\frac{1}{2}\phi\sigma & \frac{1}{2}\sigma^2 \end{pmatrix}.$$

Thus, $S_{\phi\phi}^{-1/2}(\phi,\sigma) \propto (1+\frac{1}{2}\phi^2)^{-1/2}$ and $F_{\sigma\sigma}^{1/2}(\phi,\sigma) \propto \sigma^{-1}(2+\phi^2)^{1/2}$. Hence, using again the results in Example 8, $\pi_\phi(\phi,\sigma|\mathcal{M},\mathcal{P}_0) = (1+\frac{1}{2}\phi^2)^{-1/2}\sigma^{-1}$. In the original parametrization, this is $\pi_\phi(\mu,\sigma|\mathcal{M},\mathcal{P}_0) = (1+\frac{1}{2}(\mu/\sigma)^2)^{-1/2}\sigma^{-2}$, which is *very* different from $\pi_\mu(\mu,\sigma|\mathcal{M},\mathcal{P}_0) = \pi_\sigma(\mu,\sigma|\mathcal{M},\mathcal{P}_0) = \sigma^{-1}$. The corresponding reference posterior of ϕ is $\pi(\phi|x_1,\ldots,x_n) \propto (1+\frac{1}{2}\phi^2)^{-1/2}\,p(t|\phi)$ where $t = (\sum x_j)/(\sum x_j^2)^{1/2}$, a one-dimensional (marginally sufficient) statistic whose sampling distribution, $p(t|\mu,\sigma) = p(t|\phi)$, only depends on ϕ. Thus, the reference prior algorithm is seen to be consistent under marginalization. ◁

Many parameters. The reference algorithm is easily generalized to an arbitrary number of parameters. If the model is $p(\mathbf{t}|\omega_1,\ldots,\omega_m)$, a joint reference prior

$$(42)\quad \pi(\theta_m|\theta_{m-1},\ldots,\theta_1)\times\ldots\times\pi(\theta_2|\theta_1)\times\pi(\theta_1)$$

may sequentially be obtained for each *ordered* parametrization $\{\theta_1(\omega),\ldots,\theta_m(\omega)\}$ of interest, and these are invariant under reparametrization of any of the $\theta_i(\omega)$'s. The choice of the ordered parametrization $\{\theta_1,\ldots,\theta_m\}$ precisely describes the particular prior required, namely that which *sequentially* maximizes the missing information about each of the θ_i's, conditional on $\{\theta_1,\ldots,\theta_{i-1}\}$, for $i = m$, $m-1,\ldots,1$.

EXAMPLE 9 Stein's paradox. Let D be a random sample from a m-variate normal distribution with mean $\mu = \{\mu_1,\ldots,\mu_m\}$ and unitary variance matrix. The reference prior which corresponds to any permutation of the μ_i's is uniform, and this prior leads indeed to appropriate reference posterior distributions for any of the μ_i's, namely $\pi(\mu_i|D) = N(\mu_i|\bar{x}_i, 1/\sqrt{n})$. Suppose, however, that the quantity of interest is $\theta = \sum_i \mu_i^2$, the distance of μ to the origin. As showed by Stein [1959], the posterior distribution of θ based on that uniform prior (or in any "flat" *proper* approximation) has very undesirable properties; this is due to the fact that a uniform (or nearly uniform) prior, although "noninformative" with respect to each of the individual μ_i's, is actually highly informative on the sum of their squares, introducing a severe positive bias (Stein's paradox). However, the reference prior which corresponds to a parametrization of the form $\{\theta,\lambda_1,\ldots,\lambda_{m-1}\}$ produces, for any choice of the nuisance parameters $\lambda_i = \lambda_i(\mu)$, the reference posterior $\pi(\theta|D) = \pi(\theta|t) \propto \theta^{-1/2}\chi^2(nt|m,n\theta)$, where $t = \sum_i \bar{x}_i^2$, and this posterior is shown to have the appropriate consistency properties.

Far from being specific to Stein's example, the inappropriate behaviour in problems with many parameters of specific marginal posterior distributions derived from multivariate "flat" priors (proper or improper) is indeed very frequent. Hence, sloppy, uncontrolled use of "flat" priors (rather than the relevant reference priors), is very strongly discouraged.

Limited information Although often used in contexts where no universally agreed prior knowledge about the quantity of interest is available, the reference algorithm may be used to specify a prior which incorporates any acceptable prior knowledge; it suffices to maximize the missing information within the class $\mathcal{P}$ of priors which is compatible with such accepted knowledge. Indeed, by progressive incorporation of further restrictions into $\mathcal{P}$, the reference prior algorithm becomes a method of (prior) *probability assessment*. As described below, the problem has a fairly simple analytical solution when those restrictions take the form of known expected values. The incorporation of other type of restrictions usually involves numerical computations.

EXAMPLE 10 Univariate restricted reference priors. If the probability mechanism which is assumed to have generated the available data only depends on a univarite continuous parameter $\theta \in \Theta \subset \Re$, and the class $\mathcal{P}$ of acceptable priors is a class of proper priors which satisfies some expected value restrictions, so that

$$(43) \quad \mathcal{P} = \left\{ \pi(\theta); \quad \pi(\theta) > 0, \int_\Theta \pi(\theta)\, d\theta = 1, \quad \int_\Theta g_i(\theta)\, \pi(\theta)\, d\theta = \beta_i,\, i = 1, \ldots, m \right\}$$

then the (restricted) reference prior is

$$(44) \quad \pi(\theta \mid \mathcal{M}, \mathcal{P}) \propto \pi(\theta \mid \mathcal{M}, \mathcal{P}_0)\, \exp\left[\sum\nolimits_{j=1}^{m} \gamma_i\, g_i(\theta) \right]$$

where $\pi(\theta \mid \mathcal{M}, \mathcal{P}_0)$ is the unrestricted reference prior and the γ_i's are constants (the corresponding Lagrange multipliers), to be determined by the restrictions which define $\mathcal{P}$. Suppose, for instance, that data are considered to be a random sample from a location model centered at θ, and that it is further assumed that $\mathrm{E}[\theta] = \mu_0$ and that $\mathrm{Var}[\theta] = \sigma_0^2$. The unrestricted reference prior for any regular location problem may be shown to be uniform, so that here $\pi(\theta \mid \mathcal{M}, \mathcal{P}_0) = 1$. Thus, the restricted reference prior must be of the form $\pi(\theta \mid \mathcal{M}, \mathcal{P}) \propto \exp\{\gamma_1 \theta + \gamma_2 (\theta - \mu_0)^2\}$, with $\int_\Theta \theta\, \pi(\theta \mid \mathcal{M}, \mathcal{P})\, d\theta = \mu_0$ and $\int_\Theta (\theta - \mu_0)^2\, \pi(\theta \mid \mathcal{M}, \mathcal{P})\, d\theta = \sigma_0^2$. Hence, $\pi(\theta \mid \mathcal{M}, \mathcal{P})$ is the *normal* distribution with the specified mean and variance, $\mathbf{N}(\theta \mid \mu_0, \sigma_0)$.

4.2 Frequentist Properties

Bayesian methods provide a *direct* solution to the problems typically posed in statistical inference; indeed, posterior distributions precisely state what can be said about unknown quantities of interest *given* available data and prior knowledge. In particular, unrestricted reference posterior distributions state what could be said if no prior knowledge about the quantities of interest were available.

A frequentist analysis of the behaviour of Bayesian procedures under repeated sampling may, however, be illuminating, for this provides some interesting connections between frequentist and Bayesian inference. It is found that the frequentist properties of Bayesian reference procedures are typically excellent, and may be used to provide a form of calibration for reference posterior probabilities.

Point Estimation It is generally accepted that, as the sample size increases, a "good" estimator $\tilde{\theta}$ of θ ought to get the correct value of θ eventually, that is to be *consistent*. Under appropriate regularity conditions, any Bayes estimator ϕ^* of any function $\phi(\theta)$ converges in probability to $\phi(\theta)$, so that sequences of Bayes estimators are typically *consistent*. Indeed, it is known that if there is a consistent sequence of estimators, then Bayes estimators are consistent. The rate of convergence is often best for reference Bayes estimators.

It is also generally accepted that a "good" estimator should be *admissible*, that is, *not dominated* by any other estimator in the sense that its expected loss under sampling (conditional to θ) cannot be larger for all θ values than that corresponding to another estimator. Any *proper* Bayes estimator is admissible; moreover, as established by Wald [1950], a procedure *must* be Bayesian (proper or improper) to be admissible. Most published admissibility results refer to quadratic loss functions, but they often extend to more general loss funtions. Reference Bayes estimators are typically admissible with respect to appropriate loss functions.

Notice, however, that many other apparently intuitive frequentist ideas on estimation have been proved to be potentially misleading. For example, given a sequence of n Bernoulli observations with parameter θ resulting in r positive trials, the *best unbiased* estimate of θ^2 is found to be $r(r-1)/\{n(n-1)\}$, which yields $\tilde{\theta}^2 = 0$ when $r = 1$; but to estimate the probability of two positive trials as zero, when one positive trial has been observed, is less than sensible. In marked contrast, any Bayes reference estimator provides a reasonable answer. For example, the intrinsic estimator of θ^2 is simply $(\theta^*)^2$, where θ^* is the intrinsic estimator of θ described in Section 5.1. In particular, if $r = 1$ and $n = 2$ the intrinsic estimator of θ^2 is (as one would naturally expect) $(\theta^*)^2 = 1/4$.

Interval Estimation As the sample size increases, the frequentist coverage probability of a posterior q-credible region typically converges to q so that, for *large samples*, Bayesian credible intervals may (under regularity conditions) be interpreted as *approximate* frequentist confidence regions: under repeated sampling, a Bayesian q-credible region of θ based on a large sample will cover the true value of θ approximately $100q\%$ of times. Detailed results are readily available for univariate problems. For instance, consider the probability model $\{p(D|\omega), \omega \in \Omega\}$, let $\theta = \theta(\omega)$ be any univariate quantity of interest, and let $\mathbf{t} = \mathbf{t}(D) \in T$ be any sufficient statistic. If $\theta_q(\mathbf{t})$ denotes the $100q\%$ quantile of the posterior distribution of θ which corresponds to some unspecified prior, so that

$$(45)\quad \Pr[\theta \le \theta_q(\mathbf{t})|\mathbf{t}] = \int_{\theta \le \theta_q(\mathbf{t})} \pi(\theta|\mathbf{t})\, d\theta = q,$$

then the coverage probability of the q-credible interval $\{\theta; \theta \leq \theta_q(\mathbf{t})\}$,

$$(46)\quad \Pr[\theta_q(\mathbf{t}) \geq \theta | \omega] = \int_{\theta_q(\mathbf{t}) \geq \theta} p(\mathbf{t}|\omega)\, d\mathbf{t},$$

is such that

$$(47)\quad \Pr[\theta_q(\mathbf{t}) \geq \theta | \omega] = \Pr[\theta \leq \theta_q(\mathbf{t}) | \mathbf{t}] + O(n^{-1/2}).$$

This *asymptotic* approximation is true for *all* (sufficiently regular) positive priors. However, the approximation is better, actually $O(n^{-1})$, for a particular class of priors known as (first-order) *probability matching* priors. For details on probablity matching priors see Datta and Sweeting [2005] and references therein. Reference priors are typically found to be probability matching priors, so that they provide this improved asymptotic agreement. As a matter of fact, the agreement (in regular problems) is typically quite good even for relatively small samples.

EXAMPLE 11 Product of normal means. Consider the case where independent random samples $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ have respectively been taken from the normal densities $N(x|\omega_1, 1)$ and $N(y|\omega_2, 1)$, and suppose that the quantity of interest is the product of their means, $\phi = \omega_1 \omega_2$ (for instance, one may be interested in inferences about the area ϕ of a rectangular piece of land, given measurements $\{x_i\}$ and $\{y_j\}$ of its sides). Notice that this is a simplified version of a problem that it is often encountered in the sciences, where one is interested in the product of several magnitudes, all of which have been measured with error. Using the procedure described in Example 8, with the natural approximating sequence induced by $(\omega_1, \omega_2) \in [-i, i]^2$, the ϕ-reference prior is found to be

$$(48)\quad \pi_\phi(\omega_1, \omega_2 | \mathcal{M}, \mathcal{P}_0) \propto (n\,\omega_1^2 + m\,\omega_2^2)^{-1/2},$$

very different from the uniform prior $\pi_{\omega_1}(\omega_1, \omega_2 | \mathcal{M}, \mathcal{P}_0) = \pi_{\omega_2}(\omega_1, \omega_2 | \mathcal{M}, \mathcal{P}_0) = 1$ which should be used to make objective inferences about either ω_1 or ω_2. The prior $\pi_\phi(\omega_1, \omega_2)$ may be shown to provide approximate agreement between Bayesian credible regions and frequentist confidence intervals for ϕ; indeed, this prior (with $m = n$) was originally suggested by Stein in the 1980's to obtain such approximate agreement. The same example was later used by Efron [1986] to stress the fact that, even within a fixed probability model $\{p(D|\omega), \omega \in \Omega\}$, the prior required to make objective inferences about some function of the parameters $\phi = \phi(\omega)$ must generally depend on the function ϕ. For further details on the reference analysis of this problem, see [Berger and Bernardo, 1989].

The numerical agreement between reference Bayesian credible regions and frequentist confidence intervals is actually perfect in special circumstances. Indeed, as Lindley [1958] pointed out, this is the case in those problems of inference which may be transformed to location-scale problems.

EXAMPLE 3, continued. Inference on normal parameters. Let $D = \{x_1, \ldots x_n\}$ be a random sample from a normal distribution $N(x|\mu, \sigma)$. As mentioned before, the reference posterior of the quantity of interest μ is the Student distribution $\mathrm{St}(\mu|\bar{x}, s/\sqrt{n-1}, n-1)$. Thus, normalizing μ, the *posterior* distribution of $t(\mu) = \sqrt{n-1}(\bar{x}-\mu)/s$, as a function of μ given D, is the standard Student $\mathrm{St}(t|0,1,n-1)$ with $n-1$ degrees of freedom. On the other hand, this function t is recognized to be precisely the conventional t statistic, whose *sampling distribution* is well known to *also* be standard Student with $n-1$ degrees of freedom. It follows that, *for all sample sizes*, posterior *reference* credible intervals for μ given the data will be *numerically identical* to frequentist confidence intervals based on the sampling distribution of t.

A similar result is obtained in inferences about the variance. Thus, the reference *posterior* distribution of $\lambda = \sigma^{-2}$ is the Gamma distribution $\mathrm{Ga}(\lambda|(n-1)/2, ns^2/2)$ and, hence, the *posterior* distribution of $r = ns^2/\sigma^2$, as a function of σ^2 given D, is a (central) χ^2 with $n-1$ degrees of freedom. But the function r is recognized to be a conventional statistic for this problem, whose *sampling distribution* is well known to *also* be χ^2 with $n-1$ degrees of freedom. It follows that, *for all sample sizes*, posterior *reference* credible intervals for σ^2 (or any one-to-one function of σ^2) given the data will be *numerically identical* to frequentist confidence intervals based on the sampling distribution of r. $\triangleleft$

5 INFERENCE SUMMARIES

From a Bayesian viewpoint, the final outcome of a problem of inference about *any* unknown quantity is nothing but the corresponding posterior distribution. Thus, given some data D and conditions C, *all* that can be said about any function ω of the parameters which govern the model is contained in the posterior distribution $\pi(\omega|D, C)$, and *all* that can be said about some function $\mathbf{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\mathbf{y}|D, C)$. Indeed, Bayesian inference may technically be described as a decision problem where the space of available actions is the class of those posterior probability distributions of the quantity of interest which are compatible with accepted assumptions.

However, to make it easier for the user to assimilate the appropriate conclusions, it is often convenient to *summarize* the information contained in the posterior distribution by (i) providing values of the quantity of interest which, in the light of the data, are likely to be "close" to its true value and by (ii) measuring the compatibility of the results with hypothetical values of the quantity of interest which might have been suggested in the context of the investigation. In this section, those Bayesian counterparts of traditional *estimation* and *hypothesis testing* problems are briefly considered.

5.1 Estimation

In one or two dimensions, a graph of the posterior probability density of the quantity of interest (or the probability mass function in the discrete case) immediately conveys an intuitive, "impressionist" summary of the main conclusions which may possibly be drawn on its value. Indeed, this is greatly appreciated by users, and may be quoted as an important asset of Bayesian methods. From a plot of its posterior density, the region where (given the data) a univariate quantity of interest is likely to lie is easily distinguished. For instance, all important conclusions about the value of the gravitational field in Example 3 are qualitatively available from Figure 3. However, this does not easily extend to more than two dimensions and, besides, *quantitative* conclusions (in a simpler form than that provided by the mathematical expression of the posterior distribution) are often required.

Point Estimation Let D be the available data, which are assumed to have been generated by a probability model $\{p(D|\omega), \omega \in \Omega\}$, and let $\theta = \theta(\omega) \in \Theta$ be the quantity of interest. A *point estimator* of θ is some function of the data $\tilde{\theta} = \tilde{\theta}(D)$ which could be regarded as an appropriate proxy for the actual, unknown value of θ. Formally, to choose a point estimate for θ is a *decision problem*, where the action space is the class Θ of possible θ values. From a decision-theoretic perspective, to choose a point estimate $\tilde{\theta}$ of some quantity θ is a *decision* to act as though $\tilde{\theta}$ were θ, not to assert something about the value of θ (although desire to assert something simple may well be the reason to obtain an estimate). As prescribed by the foundations of decision theory (Section 2), to solve this decision problem it is necessary to specify a *loss function* $L(\tilde{\theta}, \theta)$ measuring the consequences of acting *as if* the true value of the quantity of interest were $\tilde{\theta}$, when it is actually θ. The expected posterior loss if $\tilde{\theta}$ were used is

$$(49)\quad \overline{L}[\tilde{\theta}|D] = \int_{\Theta} L(\tilde{\theta}, \theta)\, \pi(\theta|D)\, d\theta,$$

and the corresponding *Bayes estimator* θ^* is that function of the data, $\theta^* = \theta^*(D)$, which minimizes this expectation.

EXAMPLE 12 Conventional Bayes estimators. For any given model and data, the Bayes estimator obviously depends on the chosen loss function. The loss function is context specific, and should be chosen in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for those situations where no particular uses are envisaged. These loss functions produce estimates which may be regarded as simple descriptions of the *location* of the posterior distribution. For example, if the loss function is quadratic, so that $L(\tilde{\theta}, \theta) = (\tilde{\theta} - \theta)^t(\tilde{\theta} - \theta)$, then the Bayes estimator is the *posterior mean* $\theta^* = E[\theta|D]$, assuming that the mean exists. Similarly, if the loss function is a zero-one function, so that $L(\tilde{\theta}, \theta) = 0$ if $\tilde{\theta}$ belongs to a ball or radius ϵ centered in θ and $L(\tilde{\theta}, \theta) = 1$ otherwise, then the Bayes estimator θ^* tends to the *posterior mode* as the ball radius ϵ tends to zero, assuming that a unique mode exists. If θ is univariate and the loss function is linear, so that $L(\tilde{\theta}, \theta) = c_1(\tilde{\theta} - \theta)$ if $\tilde{\theta} \geq \theta$, and $L(\tilde{\theta}, \theta) = c_2(\theta - \tilde{\theta})$ otherwise, then the Bayes estimator is the *posterior quantile* of order $c_2/(c_1 + c_2)$, so that $\Pr[\theta < \theta^*] = c_2/(c_1 + c_2)$. In particular, if $c_1 = c_2$, the Bayes estimator is the *posterior median*. The results derived for linear loss funtions clearly illustrate the fact that *any* possible parameter value may turn out be the Bayes estimator: it all depends on the loss function describing the consequences of the anticipated uses of the estimate.

EXAMPLE 13 Intrinsic estimation. Conventional loss functions are typically non-invariant under reparametrization. It follows that the Bayes estimator ϕ^* of a one-to-one transformation $\phi = \phi(\theta)$ of the original parameter θ is not necessarily $\phi(\theta^*)$ (the *univariate* posterior median, which *is* invariant, is an interesting exception). Moreover, conventional loss functions focus on the "distance" between the estimate $\tilde{\theta}$ and the true value θ, rather then on the "distance" between the probability models they label. Inference-oriented loss functions directly focus on how different the probability *model* $p(D|\theta, \lambda)$ is from its closest approximation within the family $\{p(D|\tilde{\theta}, \lambda_i), \lambda_i \in \Lambda\}$, and typically produce invariant solutions. An attractive example is the *intrinsic discrepancy*, $\delta(\tilde{\theta}, \theta)$ defined as the minimum logarithmic divergence between a probability model labeled by θ and a probability model labeled by $\tilde{\theta}$. When there are no nuisance parameters, this is given by

$$(50)\quad \delta(\tilde{\theta}, \theta) = \min\{\kappa(\tilde{\theta}|\theta), \kappa(\theta|\tilde{\theta})\}, \quad \kappa(\theta_i|\theta) = \int_T p(\mathbf{t}|\theta)\, \log \frac{p(\mathbf{t}|\theta)}{p(\mathbf{t}|\theta_i)}\, d\mathbf{t},$$

where $\mathbf{t} = \mathbf{t}(D) \in T$ is *any* sufficient statistic (which may well be the whole data set D). The definition is easily extended to problems with nuisance parameters; in this case,

$$(51)\quad \delta(\tilde{\theta}, \theta, \lambda) = \min_{\lambda_i \in \Lambda} \delta(\tilde{\theta}, \lambda_i, \theta, \lambda)$$

measures the logarithmic divergence from $p(\mathbf{t}|\theta, \lambda)$ of its closest approximation with $\theta = \tilde{\theta}$, and the loss function now depends on the complete parameter vector (θ, λ). Although not explicitly shown in the notation, the intrinsic discrepancy function typically depends on the sample size n; indeed, when the data consist of a random sample $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from some model $p(\mathbf{x}|\theta)$ then $\kappa(\theta_i|\theta) = n \int_X p(\mathbf{x}|\theta) \log[p(\mathbf{x}|\theta)/p(\mathbf{x}|\theta_i)]\, d\mathbf{x}$ so that the discrepancy associated with the full model is simply n times the discrepancy which corresponds to a single observation. The intrinsic discrepancy is a symmetric, non-negative loss function with a direct interpretation in information-theoretic terms as the minimum amount of information which is expected to be necessary to distinguish between the model $p(D|\theta, \lambda)$ and its closest approximation within the class $\{p(D|\tilde{\theta}, \lambda_i), \lambda_i \in \Lambda\}$. Moreover, it is invariant under one-to-one reparametrization of the parameter of interest θ, and does not depend on the choice of the nuisance parameter λ. The *intrinsic estimator* is naturally obtained by minimizing the reference posterior expected intrinsic discrepancy

$$(52)\quad d(\tilde{\theta}|D) = \int_\Lambda \int_\Theta \delta(\tilde{\theta}, \theta, \lambda)\, \pi(\theta, \lambda|D)\, d\theta d\lambda.$$

Since the intrinsic discrepancy is invariant under reparametrization, minimizing its posterior expectation produces invariant estimators. For further details on intrinsic point estimation see [Bernardo and Juárez, 2003; Bernardo, 2006].

EXAMPLE 2, continued. Intrinsic estimation of a binomial parameter. In the estimation of a binomial proportion θ, given data $D = (n, r)$, the Bayes reference estimator associated with the quadratic loss (the corresponding posterior mean) is $E[\theta|D] = (r + \frac{1}{2})/(n+1)$, while the quadratic loss based estimator of, say, the log-odds $\phi(\theta) = \log[\theta/(1-\theta)]$, is found to be $E[\phi|D] = \psi(r + \frac{1}{2}) - \psi(n - r + \frac{1}{2})$ (where $\psi(x) = d\log[\Gamma(x)]/dx$ is the *digamma* function), which is *not* equal to $\phi(E[\theta|D])$. The intrinsic loss function in this problem is

$$(53)\quad \delta(\tilde{\theta}, \theta) = n\, \min\{\kappa(\tilde{\theta}|\theta), \kappa(\theta|\tilde{\theta})\}, \quad \kappa(\theta_i|\theta) = \theta \log\frac{\theta}{\theta_i} + (1-\theta)\log\frac{1-\theta}{1-\theta_i}\,,$$

and the corresponding intrinsic estimator θ^* is obtained by minimizing the expected posterior loss $d(\tilde{\theta}|D) = \int \delta(\tilde{\theta}, \theta)\, \pi(\theta|D)\, d\theta$. The exact value of θ^* may be obtained by numerical minimization, but a very good approximation is given by $\theta^* \approx (r + \frac{1}{3})/(n + \frac{2}{3})$.

Since intrinsic estimation is an invariant procedure, the intrinsic estimator of the log-odds will simply be the log-odds of the intrinsic estimator of θ. As one would expect, when r and $n - r$ are both large, all Bayes estimators of any well-behaved function $\phi(\theta)$ will cluster around $\phi(E[\theta|D])$. ◁

Interval Estimation To describe the inferential content of the posterior distribution of the quantity of interest $\pi(\theta|D)$ it is often convenient to quote regions $R \subset \Theta$ of given probability under $\pi(\theta|D)$. For example, the identification of regions containing 50%, 90%, 95%, or 99% of the probability under the posterior may be sufficient to convey the general quantitative messages implicit in $\pi(\theta|D)$; indeed, this is the intuitive basis of graphical representations of univariate distributions like those provided by boxplots.

Any region $R \subset \Theta$ such that $\int_R \pi(\theta|D)d\theta = q$ (so that, given data D, the true value of θ belongs to R with probability q), is said to be a posterior *q-credible region* of θ. Notice that this provides immediately a direct intuitive statement about the unknown quantity of interest θ in probability terms, in marked contrast to the circumlocutory statements provided by frequentist confidence intervals. Clearly, for any given q there are generally infinitely many credible regions. A credible region is invariant under reparametrization; thus, for any q-credible region R of θ, $\phi(R)$ is a q-credible region of $\phi = \phi(\theta)$. Sometimes, credible regions are selected to have minimum size (length, area, volume), resulting in *highest probability density* (HPD) regions, where all points in the region have larger probability density than all points outside. However, HPD regions are *not* invariant under reparametrization: the image $\phi(R)$ of an HPD region R will be a credible region for ϕ, but will not generally be HPD; indeed, there is no compelling reason to restrict attention to HPD credible regions. In one-dimensional problems, posterior quantiles are often used to derive credible regions. Thus, if $\theta_q = \theta_q(D)$ is the $100q\%$ posterior quantile of θ, then $R = \{\theta;\, \theta \leq \theta_q\}$ is a one-sided, typically unique q-credible region, and it is invariant under reparametrization. Indeed, *probability centered* q-credible regions of the form $R = \{\theta;\, \theta_{(1-q)/2} \leq \theta \leq \theta_{(1+q)/2}\}$ are easier to compute, and are often quoted in preference to HPD regions.

EXAMPLE 3. Inference on normal parameters, continued. In the numerical example about the value of the gravitational field described in Figure 3a, the interval $[9.788, 9.829]$ in the unrestricted posterior density of g is a HPD, 95%-credible region for g. Similarly, the interval $[9.7803, 9.8322]$ in Figure 3b is also a 95%-credible region for g, but it is not HPD. $\triangleleft$

Decision theory may also be used to select credible regions. Thus, *lowest posterior loss* (LPL) regions, are defined as those where all points in the region have smaller posterior expected loss than all points outside. Using the intrinsic discrepancy as a loss function yields *intrinsic credible regions* which, as one would expect from an invariant loss function, are coherent under one-to-one transformations. For details, see [Bernardo, 2005b; 2007].

The concept of a credible region for a function $\theta = \theta(\omega)$ of the parameter vector is trivially extended to prediction problems. Thus, a posterior q-credible region for $\mathbf{x} \in \mathcal{X}$ is a subset R of the sample space $\mathcal{X}$ with posterior predictive probability q, so that $\int_R p(\mathbf{x}|D)d\mathbf{x} = q$.

5.2 Hypothesis Testing

The reference posterior distribution $\pi(\theta|D)$ of the quantity of interest θ conveys immediate intuitive information on those values of θ which, given the assumed model, may be taken to be *compatible* with the observed data D, namely, those with a relatively high probability density. Sometimes, a *restriction* $\theta \in \Theta_0 \subset \Theta$ of the possible values of the quantity of interest (where Θ_0 may possibly consists of a single value $\theta_{\mathbf{0}}$) is suggested in the course of the investigation as deserving special consideration, either because restricting θ to Θ_0 would greatly simplify the model, or because there are additional, context specific arguments suggesting that $\theta \in \Theta_0$. Intuitively, the *hypothesis* $H_0 \equiv \{\theta \in \Theta_0\}$ should be judged to be *compatible* with the observed data D if there are elements in Θ_0 with a relatively high posterior density. However, a more precise conclusion is often required and, once again, this is made possible by adopting a decision-oriented approach. Formally, testing the hypothesis $H_0 \equiv \{\theta \in \Theta_0\}$ is a *decision problem* where the action space has only two elements, namely to accept (a_0) or to reject (a_1) the proposed restriction. To solve this decision problem, it is necessary to specify an appropriate loss function, $L(a_i, \theta)$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value θ of the vector of interest. Notice that this requires the statement of an *alternative* a_1 to accepting H_0; this is only to be expected, for an action is taken not because it is good, but because it is better than anything else that has been imagined.

Given data D, the optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting, $\int_\Theta L(a_0, \theta)\, \pi(\theta|D)\, d\theta$, is larger than the expected posterior loss of rejecting, $\int_\Theta L(a_1, \theta)\, \pi(\theta|D)\, d\theta$, that is, if (and only if)

$$(54)\quad \int_{\Theta}[L(a_0,\theta)-L(a_1,\theta)]\,\pi(\theta|D)\,d\theta=\int_{\Theta}\Delta L(\theta)\,\pi(\theta|D)\,d\theta>0.$$

Therefore, only the loss difference $\Delta L(\theta)=L(a_0,\theta)-L(a_1,\theta)$, which measures the *advantage* of rejecting H_0 as a function of θ, has to be specified. Thus, as common sense dictates, the hypothesis H_0 should be rejected whenever the expected advantage of rejecting H_0 is positive.

A crucial element in the specification of the loss function is a description of what is actually meant by rejecting H_0. By assumption a_0 means to act *as if* H_0 were true, i.e. as if $\theta\in\Theta_0$, but there are at least two options for the alternative action a_1. This may either mean (i) the *negation* of H_0, that is to act as if $\theta\notin\Theta_0$ or, alternatively, it may rather mean (ii) to reject the simplification implied by H_0 and to keep the unrestricted model, $\theta\in\Theta$, which is true by assumption. Both options have been analyzed in the literature, although it may be argued that the problems of scientific data analysis where hypothesis testing procedures are typically used are better described by the second alternative. Indeed, an established model, identified by $H_0\equiv\{\theta\in\Theta_0\}$, is often embedded into a more general model, $\{\theta\in\Theta,\Theta_0\subset\Theta\}$, constructed to include possibly promising departures from H_0, and it is required to verify whether presently available data D are still compatible with $\theta\in\Theta_0$, or whether the extension to $\theta\in\Theta$ is really required.

EXAMPLE 14 Conventional hypothesis testing. Let $\pi(\theta|D)$, $\theta\in\Theta$, be the posterior distribution of the quantity of interest, let a_0 be the decision to work under the restriction $\theta\in\Theta_0$ and let a_1 be the decision to work under the complementary restriction $\theta\notin\Theta_0$. Suppose, moreover, that the loss structure has the simple, zero-one form given by $\{L(a_0,\theta)=0,L(a_1,\theta)=1\}$ if $\theta\in\Theta_0$ and, similarly, $\{L(a_0,\theta)=1,L(a_1,\theta)=0\}$ if $\theta\notin\Theta_0$, so that the *advantage* $\Delta L(\theta)$ of rejecting H_0 is 1 if $\theta\notin\Theta_0$ and it is -1 otherwise. With this loss function it is immediately found that the optimal action is to reject H_0 if (and only if) $\Pr(\theta\notin\Theta_0|D)>\Pr(\theta\in\Theta_0|D)$. Notice that this formulation requires that $\Pr(\theta\in\Theta_0)>0$, that is, that the hypothesis H_0 has a strictly positive prior probability. If θ is a continuous parameter and Θ_0 has zero measure (for instance if H_0 consists of a single point $\theta_{\mathbf{0}}$), this requires the use of a non-regular "sharp" prior concentrating a positive probability mass on Θ_0. For details see [Kaas and Rafetery, 1995] and references therein.

EXAMPLE 15 Intrinsic hypothesis testing. Again, let $\pi(\theta|D)$, $\theta\in\Theta$, be the posterior distribution of the quantity of interest, and let a_0 be the decision to work under the restriction $\theta\in\Theta_0$, but let a_1 now be the decision to keep the general, unrestricted model $\theta\in\Theta$. In this case, the advantage $\Delta L(\theta)$ of rejecting H_0 as a function of θ may safely be assumed to have the form $\Delta L(\theta)=\delta(\Theta_0,\theta)-\delta^*$, for some $\delta^*>0$, where (i) $\delta(\Theta_0,\theta)$ is some measure of the discrepancy between the assumed model $p(D|\theta)$ and its closest approximation within the class $\{p(D|\theta_0),\theta_0\in\Theta_0\}$, such that $\delta(\Theta_0,\theta)=0$ whenever $\theta\in\Theta_0$, and (ii) δ^* is a context dependent *utility constant* which measures the (necessarily positive) advantage of being able to work with the simpler model when it is true. Choices for both $\delta(\Theta_0,\theta)$ and δ^* which may be appropriate for general use will now be described.

For reasons similar to those supporting its use in point estimation, an attractive choice for the function $d(\Theta_0, \theta)$ is an appropriate extension of the intrinsic discrepancy; when there are no nuisance parameters, this is given by

$$(55)\quad \delta(\Theta_0, \theta) = \inf_{\theta_0 \in \boldsymbol{\Theta}_0} \min\{\kappa(\theta_0|\theta),\, \kappa(\theta|\theta_0)\}$$

where $\kappa(\theta_0|\theta) = \int_T p(\mathbf{t}|\theta) \log\{p(\mathbf{t}|\theta)/p(\mathbf{t}|\theta_0)\} d\mathbf{t}$, and $\mathbf{t} = \mathbf{t}(D) \in T$ is *any* sufficient statistic, which may well be the whole dataset D. As before, if the data $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ consist of a random sample from $p(\mathbf{x}|\theta)$, then

$$(56)\quad \kappa(\theta_0|\theta) = n \int_X p(\mathbf{x}|\theta) \log \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\theta_0)}\, d\mathbf{x}.$$

Naturally, the loss function $\delta(\Theta_0, \theta)$ reduces to the intrinsic discrepancy $\delta(\theta_0, \theta)$ of Example 13 when Θ_0 contains a single element θ_0. Besides, as in the case of estimation, the definition is easily extended to problems with nuisance parameters, with

$$(57)\quad \delta(\boldsymbol{\Theta}_0, \theta, \lambda) = \inf_{\theta_0 \in \boldsymbol{\Theta}_0, \boldsymbol{\lambda_0} \in \Lambda} \delta(\theta_0, \lambda_o, \theta, \lambda).$$

The hypothesis H_0 should be rejected if the posterior expected advantage of rejecting is

$$(58)\quad d(\boldsymbol{\Theta}_0|D) = \int_\Lambda \int_\Theta \delta(\boldsymbol{\Theta}_0, \theta, \lambda)\, \pi(\theta, \lambda|D)\, d\theta d\lambda > \delta^*,$$

for some $\delta^* > 0$. As an expectation of a non-negative quantity, $d(\Theta_0, D)$ is obviously nonnegative. Morovever, if $\phi = \phi(\theta)$ is a one-to-one transformation of θ, then $d(\phi(\Theta_0), D) = d(\Theta_0, D)$ so that, as one should clearly require, the expected intrinsic loss of rejecting H_0 is invariant under reparametrization.

It may be shown that, as the sample size increases, the expected value of $d(\Theta_0, D)$ under sampling tends to one when H_0 is true, and tends to infinity otherwise; thus $d(\Theta_0, D)$ may be regarded as a continuous, positive measure of how inappropriate (in loss of information units) it would be to simplify the model by accepting H_0. In traditional language, $d(\Theta_0, D)$ is a *test statistic* for H_0 and the hypothesis should be rejected if the value of $d(\Theta_0, D)$ exceeds some *critical value* δ^*. In sharp contrast to conventional hypothesis testing, this critical value δ^* is found to be a context specific, positive utility constant δ^*, which may precisely be described as the number of *information units* which the decision maker is prepared to lose in order to be able to work with the simpler model H_0, and does not depend on the sampling properties of the probability model. The procedure may be used with standard, continuous regular priors even in *sharp* hypothesis testing, when Θ_0 is a zero-measure set (as would be the case if θ is continuous and Θ_0 contains a single point θ_0).

Naturally, to implement the test, the utility constant δ^* which defines the rejection region must be chosen. Values of $d(\Theta_0, D)$ of about 1 should be regarded as an indication of no evidence against H_0, since this is precisely the expected value of the test statistic $d(\Theta_0, D)$ under repeated sampling from the null. If follows from its definition that $d(\Theta_0, D)$ is the reference posterior expectation of the log-likelihood ratio against the null. Hence, values of $d(\Theta_0, D)$ of about $\log[12] \approx 2.5$, and $\log[150] \approx 5$ should be respectively regarded as an indication of mild evidence against H_0, and significant evidence against H_0. In the canonical problem of testing a value $\mu = \mu_0$ for the mean of a normal distribution with known variance (see below), these values correspond to the observed sample mean $\bar{x}$ respectively lying 2 or 3 posterior standard deviations from the null value μ_0. Notice that, in sharp contrast to frequentist hypothesis testing, where it is hazily recommended to adjust the significance level for dimensionality and sample size, this provides an absolute scale (in information units) which remains valid for any sample size and any dimensionality.

For further details on intrinsic hypothesis testing see [Bernardo and Rueda, 2003; Bernardo and Pérez, 2007].

EXAMPLE 16 Testing the value of a normal mean. Let the data $D = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $\mathbf{N}(x|\mu, \sigma)$, where σ is assumed to be known, and consider the problem of testing whether these data are or are not compatible with some specific sharp hypothesis $H_0 \equiv \{\mu = \mu_0\}$ on the value of the mean.

The conventional approach to this problem requires a non-regular prior which places a probability mass, say p_0, on the value μ_0 to be tested, with the remaining $1 - p_0$ probability continuously distributed over $\Re$. If this prior is chosen to be $\pi(\mu|\mu \neq \mu_0) = N(\mu|\mu_0, \sigma_0)$, Bayes theorem may be used to obtain the corresponding posterior probability,

$$(59)\quad \Pr[\mu_0|D, \lambda] = \frac{B_{01}(D, \lambda)\, p_0}{(1 - p_0) + p_0\, B_{01}(D, \lambda)}\,,$$

$$(60)\quad B_{01}(D, \lambda) = \Big(1 + \frac{n}{\lambda}\Big)^{1/2} \exp\left[-\frac{1}{2}\frac{n}{n + \lambda}\, z^2\right],$$

where $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ measures, in standard deviations, the distance between $\bar{x}$ and μ_0 and $\lambda = \sigma^2/\sigma_0^2$ is the ratio of model to prior variance. The function $B_{01}(D, \lambda)$, a ratio of (integrated) likelihood functions, is called the *Bayes factor* in favour of H_0. With a conventional zero-one loss function, H_0 should be rejected if $\Pr[\mu_0|D, \lambda] < 1/2$. The choices $p_0 = 1/2$ and $\lambda = 1$ or $\lambda = 1/2$, describing particular forms of *sharp* prior knowledge, have been suggested in the literature for routine use. The conventional approach to sharp hypothesis testing deals with situations of *concentrated* prior probability; it *assumes* important prior knowledge about the value of μ and, hence, should *not* be used unless this is an appropriate assumption. Moreover [Bartlett, 1957], the resulting posterior probability is extremely sensitive to the specific prior specification. In most applications, H_0 is really a hazily defined small region rather than a point. For moderate sample sizes, the posterior probability $\Pr[\mu_0|D, \lambda]$ is an *approximation* to the posterior probability $\Pr[\mu_0 - \epsilon < \mu < \mu_0 - \epsilon | D, \lambda]$ for some small interval around μ_0 which would have been obtained from a regular, continuous prior heavily concentrated around μ_0; however, this approximation *always* breaks down for sufficiently large sample sizes. One consequence (which is immediately apparent from the last two equations) is that for any *fixed* value of the pertinent statistic z, the posterior probability of the null, $\Pr[\mu_0 | D, \lambda]$, tends to one as $n \to \infty$. Far from being specific to this example, this unappealing behaviour of posterior probabilities based on sharp, non-regular priors generally known as *Lindley's paradox* [Lindley, 1957] is *always* present in the conventional Bayesian approach to *sharp* hypothesis testing.

The intrinsic approach may be used without assuming any sharp prior knowledge. The intrinsic discrepancy is $\delta(\mu_0, \mu) = n(\mu - \mu_0)^2/(2\sigma^2)$, a simple transformation of the standardized distance between μ and μ_0. The reference prior is uniform and the corresponding (proper) posterior distribution is $\pi(\mu|D) = N(\mu|\bar{x}, \sigma/\sqrt{n})$. The expected value of $\delta(\mu_0, \mu)$ with respect to this posterior is $d(\mu_0, D) = (1 + z^2)/2$, where $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ is the standardized distance between $\bar{x}$ and μ_0. As foretold by the general theory, the expected value of $d(\mu_0, D)$ under repeated sampling is one if $\mu = \mu_0$, and increases linearly with n if $\mu = \mu_0$. Moreover, in this canonical example, to reject H_0 whenever $|z| > 2$ or $|z| > 3$, that is whenever μ_0 is 2 or 3 posterior standard deviations away from $\bar{x}$, respectively corresponds to rejecting H_0 whenever $d(\mu_0, D)$ is larger than 2.5, or larger than 5.

If σ is unknown, the reference prior is $\pi(\mu, \sigma) = \sigma^{-1}$, and the intrinsic discrepancy becomes

$$(61) \quad \delta(\mu_0, \mu, \sigma) = \frac{n}{2} \log \left[1 + \left(\frac{\mu - \mu_0}{\sigma} \right)^2 \right].$$

The intrinsic test statistic $d(\mu_0, D)$ is found as the expected value of $\delta(\mu_0, \mu, \sigma)$ under the corresponding joint referenceposterior distribution; this may be exactly expressed in terms of hypergeometric functions, and is well approximated by

$$(62) \quad d(\mu_0, D) \approx \frac{1}{2} + \frac{n}{2} \log \left(1 + \frac{t^2}{n} \right),$$

where t is the traditional statistic $t = \sqrt{n-1}(\bar{x} - \mu_0)/s$, $ns^2 = \sum_j (x_j - \bar{x})^2$. For instance, for samples sizes 5, 30 and 1000, and using the utility constant $\delta^* = 5$, the hypothesis H_0 would be rejected whenever $|t|$ is respectively larger than 5.025, 3.240, and 3.007.

6 DISCUSSION

This article focuses on the basic concepts of the Bayesian paradigm, with special emphasis on the derivation of "objective" methods, where the results only depend on the data obtained and the model assumed. Many technical aspects have been spared; the interested reader is referred to the bibliography for further information. This final section briefly reviews the main arguments for an objective Bayesian approach.

6.1 Coherence

By using probability distributions to characterize *all* uncertainties in the problem, the Bayesian paradigm reduces statistical inference to applied probability, thereby ensuring the coherence of the proposed solutions. There is no need to investigate, on a case by case basis, whether or not the solution to a particular problem is logically correct: a Bayesian result is only a *mathematical consequence of explicitly stated assumptions* and hence, unless a logical mistake has been committed in its derivation, it cannot be formally wrong. In marked contrast, conventional statistical methods are plagued with counterexamples. These include, among many others, negative estimators of positive quantities, q-confidence regions ($q < 1$) which consist of the whole parameter space, empty sets of "appropriate" solutions, and incompatible answers from alternative methodologies simultaneously supported by the theory.

The Bayesian approach does require, however, the specification of a (prior) probability distribution over the parameter space. The sentence "a prior distribution does not exist for this problem" is often stated to justify the use of non-Bayesian methods. However, the general representation theorem *proves the existence* of such a distribution whenever the observations are assumed to be exchangeable (and, if they are assumed to be a random sample then, *a fortiori*, they are assumed to be exchangeable). To ignore this fact, and to proceed as if a prior distribution did not exist, just because it is not easy to specify, is mathematically untenable.

6.2 Objectivity

It is generally accepted that any statistical analysis is subjective, in the sense that it is always conditional on accepted assumptions (on the structure of the data, on the probability model, and on the outcome space) and those assumptions, although possibly well founded, are definitely *subjective* choices. It is, therefore, mandatory to make all assumptions very explicit.

Users of conventional statistical methods rarely dispute the mathematical foundations of the Bayesian approach, but claim to be able to produce "objective" answers in contrast to the possibly subjective elements involved in the choice of the prior distribution.

Bayesian methods do indeed require the choice of a prior distribution, and critics of the Bayesian approach systematically point out that in many important situations, including scientific reporting and public decision making, the results must exclusively depend on documented data which might be subject to independent scrutiny. This is of course true, but those critics choose to ignore the fact that this particular case is covered within the Bayesian approach by the use of *reference* prior distributions which (i) are mathematically derived from the accepted probability model (and, hence, they are "objective" insofar as the choice of that model might be objective) and, (ii) by construction, they produce posterior probability distributions which, given the accepted probability model, *only* contain the information about their values which data may provide and, *optionally*, any further contextual information over which there might be universal agreement.

6.3 Operational meaning

An issue related to objectivity is that of the operational meaning of reference posterior probabilities; it is found that the analysis of their behaviour under repeated sampling provides a suggestive form of calibration. Indeed, $\Pr[\theta \in R|D] = \int_R \pi(\theta|D)\, d\theta$, the reference posterior probability that $\theta \in R$, is *both* a measure of the conditional uncertainty (given the assumed model and the observed data D) about the event that the unknown value of θ belongs to $R \subset \Theta$, and the limiting proportion of the regions which would cover θ under repeated sampling conditional on data "sufficiently similar" to D. Under broad conditions (to guarantee regular asymptotic behaviour), all large data sets from the same model are "sufficiently similar" among themselves in this sense and hence, given those conditions, reference posterior credible regions are *approximate* frequentist confidence regions.

The conditions for this approximate equivalence to hold exclude, however, important special cases, like those involving "extreme" or "relevant" observations. In very special situations, when probability models may be transformed to location-scale models, there is an exact equivalence; in those cases reference posterior credible intervals are, for any sample size, exact frequentist confidence intervals.

6.4 Generality

In sharp contrast to most conventional statistical methods, which may only be exactly applied to a handful of relatively simple stylized situations, Bayesian methods are defined to be totally general. Indeed, for a given probability model and prior distribution over its parameters, the derivation of posterior distributions is a well-defined mathematical exercise. In particular, Bayesian methods do not require any particular regularity conditions on the probability model, do not depend on the existence of sufficient statistics of finite dimension, do not rely on asymptotic relations, and do not require the derivation of any sampling distribution, nor (a fortiori) the existence of a "pivotal" statistic whose sampling distribution is independent of the parameters.

However, when used in complex models with many parameters, Bayesian methods often require the computation of multidimensional definite integrals and, for a long time in the past, this requirement effectively placed practical limits on the complexity of the problems which could be handled. This has dramatically changed in recent years with the general availability of large computing power, and the parallel development of simulation-based numerical integration techniues like *importance sampling* or *Markov chain Monte Carlo* (MCMC). These methods provide a structure within which many complex models may be analyzed using generic software. MCMC is numerical integration using Markov chains. Monte Carlo integration proceeds by drawing samples from the required distributions, and computing sample averages to approximate expectations. MCMC methods draw the required samples by running appropriately defined Markov chains for a long time; specific methods to construct those chains include the Gibbs sampler and the Metropolis algorithm, originated in the 1950's in the literature of statistical physics. The development of improved algorithms and appropriate diagnostic tools to establish their convergence, remains a very active research area.

For an introduction to MCMC methods in Bayesian inference, see [Gilks *et al.*, 1996; Mira, 2005], and references therein.

BIBLIOGRAPHY

[Bartlett, 1957] M. Bartlett. A comment on D. V. Lindley's statistical paradox. 44, 533–534, 1957.

[Berger and Bernardo, 1989] J. O. Berger and J. M. Bernardo. Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.,*84, 200–207, 1989.

[Berger and Bernardo, 1992a] J. O. Berger and J. M. Bernardo. Ordered group reference priors with applications to a multinomial problem. 79, 25–37, 1992.

[Berger and Bernardo, 1992b] J. O. Berger and J. M. Bernardo. Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics,*, 323–340, 1992.

[Berger and Bernardo, 1992c] J. O. Berger and J. M. Bernardo. On the development of reference priors. 4, 35–60, 1992 (with discussion).

[Berger *et al.*, 2009] J. O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *Ann Statist.*, 37, 905–938, 2009.

[Berger, 1985] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Berlin: Springer., 1985.

[Bernardo, 1979a] J. M. Bernardo. Expected information as expected utility. *Ann Statist.,*7, 686–690, 1979.

[Bernardo, 1979b] J. M. Bernardo. Reference posterior distributions for Bayesian inference. *J. R. Statist. Soc. B* **41**, 113-147 (with discussion). Reprinted in *Bayesian Inference* **1** (G. C. Tiao and N. G. Polson, eds). Oxford: Edward Elgar, 229-263, 1979.

[Bernardo, 1981] J. M. Bernardo. Reference decisions. *Symposia Mathematica* **25**, 85–94, 1981.

[Bernardo, 1997] J. M. Bernardo. Noninformative priors do not exist. *J. Statist. Planning and Inference,*65, 159–189, 1997 (with discussion).

[Bernardo, 2005a] J. M. Bernardo. Reference analysis. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.) Amsterdam: Elsevier, 17–90, 2005.

[Bernardo, 2005b] J. M. Bernardo. Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test,*14, 317–384, 2005 (with discussion).

[Bernardo, 2006] J. M. Bernardo. Intrinsic point estimation of the normal variance. *Bayesian Statistics and its Applications.* (S. K. Upadhyay, U. Singh and D. K. Dey, eds.) New Delhi: Anamaya Pub, 110-121, 2006.

[Bernardo, 2007] J. M. Bernardo. Objective Bayesian point and region estimation in location-scale models. *Sort* **14**, 3–44, 2007.

[Bernardo and Juárez, 2003] J. M. Bernardo and M. A. Juärez. Intrinsic Estimation. 7, 465–476, 2003.

[Bernardo and Pérez, 2007] J. M. Bernardo and S. Péez. Comparing normal means: New methods for an old problem. *Bayesian Analysis* **2**, 45–58, 2007.

[Bernardo and Rámon, 1998] J. M. Bernardo and J. M. Ramón. An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician,*47, 1–35, 1998.

[Bernardo and Rueda, 2002] J. M. Bernardo and R. Rueda. Bayesian hypothesis testing: A reference approach. *International Statistical Review* **70** , 351-372, 2002.

[Bernardo and Smith, 1994] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*, Chichester: Wiley, 1994. Second edition forthcoming.

[Box and Tiao, 1973] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley, 1973.

[Datta and Sweeting, 2005] G. S. Datta and T. J. Sweeting. Probability matching priors. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.) Amsterdam: Elsevier, 91–114, 2005.

[Dawid *et al.*, 1973] A. P. Dawid, M. Stone, and J. V. Zidek. Marginalization paradoxes in Bayesian and structural inference. *J. R. Statist. Soc. B* **35**, 189-233, 1973 (with discussion).

[de Finetti, 1937] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68, 1937. Reprinted in 1980 as 'Foresight; its logical laws, its subjective sources' in *Studies in Subjective Probability,*, 93–158.

[de Finetti, 1970] B. de Finetti *Teoria delle Probabilité* Turin: Einaudi, 1970. English translation as *Theory of Probability* Chichester: Wiley,, 1975.

[DeGroot, 1970] M. H. DeGroot. *Optimal Statistical Decisions*, New York: McGraw Hill,, 1970.

[Efron, 1986] B. Efron. Why isn't everyone a Bayesian? *Amer. Statist.*,40, 1–11, 198 (with discussion).

[Geisser, 1993] S. Geisser. *Predictive Inference: an Introduction.* London: Chapman and Hall, 1993.

[Gilks *et al.*, 1996] W. R. Gilks, S. Y. Richardson, and D. J. Spiegelhalter, eds. *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall, 1996.

[Jaynes, 1976] E. T. Jaynes. Confidence intervals vs. Bayesian intervals. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* **2** (W. L. Harper and and C. A. Hooker, eds). Dordrecht: Reidel, 175–257, 1976 (with discussion).

[Jeffreys, 1939] H. Jeffreys. *Theory of Probability.* Oxford: Oxford University Press. Third edition in 1961, Oxford: Oxford University Press, 1939.

[Kass and Raftery, 1995] R. E. Kass and A. E. Raftery. Bayes factors. *J. Amer. Statist. Assoc.,*90, 773–795, 1995.

[Kass and Wasserman, 1996] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.,*91, 1343–1370, 1996.

[Laplace, 1812] P. S. Laplace. *Théorie Analytique des Probabilités.* Paris: Courcier, 1812. Reprinted as *Oeuvres Complétes de Laplace* **7**, 1878–1912. Paris: Gauthier-Villars.

[Lindley, 1957] D. V. Lindley. A statistical paradox. 44, 187–192, 1957.

[Lindley, 1958] D. V. Lindley. Fiducial distribution and Bayes' Theorem. *J. Roy. Statis. Soc.*,20, 102–107, 1958.

[Lindley, 1965] D. V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint.* Cambridge: Cambridge University Press, 1965.

[Lindley, 1972] D. V. Lindley. *Bayesian Statistics, a Review.* Philadelphia, PA: SIAM, 1972.

[Liseo, 2005] B. Liseo. The elimination of nuisance parameters. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.) Amsterdam: Elsevier, 193–219, 2005.

[Mira, 2005] A. Mira. MCMC methods to estimate Bayesian parametric models. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.) Amsterdam: Elsevier, 415–436, 2005.

[Ramsey, 1926] F. P. Ramsey. Truth and probability. *The Foundations of Mathematics and Other Logical Essays* (R. B. Braithwaite, ed.). London: Kegan Paul 1926 (1931), 156–198. Reprinted in 1980 in *Studies in Subjective Probability*,, 61–92.

[Savage, 1954] L. J. Savage. *The Foundations of Statistics.* New York: Wiley, 1954. Second edition in 1972, New York: Dover.

[Stein, 1959] C. Stein. An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.*, 30, 877–880, 1959.

[Zellner, 1971] A. Zellner. *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley, 1971. Reprinted in 1987, Melbourne, FL: Kreiger.