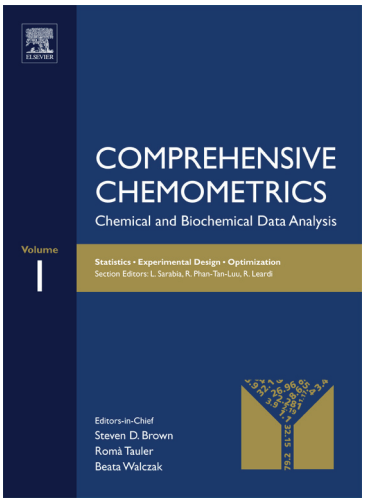

J. M. Bernardo 2009 Bayesian Methodology in Statistics.
In: Brown S, Tauler R, Walczak R (eds.) Comprehensive Chemometrics, volume 1, pp. 213-245 Oxford: Elsevier.

1.08 Bayesian Methodology in Statistics

J. M. Bernardo, Universitat de València, Valencia, Spain

This chapter includes updated sections of the paper *Bayesian Statistics*, prepared by the author for the Encyclopedia of Life Support Systems, a 2003 online UNESCO publication.

Symbols

{*A*, Θ, C}	a decision problem with class of actions *A*, set of relevant events Θ, and set of consequences *C*.
Be(*x*|α, β)	the PDF of a beta distribution for *x*, with parameters α and β.
$d(\tilde{\theta}\|D)$	expected intrinsic discrepancy between $p(D\|\tilde{\theta})$ and the true model, given data D.
Fij(ω)	generic element of the Fisher information matrix of a general model of the form $M \equiv \{p(x\|\omega), x \in X, \omega \in \Omega\}$.
$I\theta\{T, p(\theta)\}$	expected information about the value of θ to be provided by an observation $t \in T$ from $p(t\|\theta)$, when the prior is $p(\theta)$.
$\kappa\{\hat{p}(x)\|p(x)\}$	the logarithmic (Kullback–Leibler) divergence of $\hat{p}x$ from px.
L(*a*, θ)	the loss of action *a* when θ occurs.
N(*x*|μ, σ)	the PDF of a normal distribution for *x*, with mean μ and variance σ2.
p(θ|*D*, *C*)	probability density of the unknown parameter θ given data *D* and conditions *C*.
p(*x*|*D*)	the predictive distribution of a future observation *x*, given data *D*.
p(*x*|θ, *C*)	probability density of the observable random vector *x* given parameter θ and conditions *C*.
Pr(*E*|*D*, *A*, *K*)	the probability of the event *E*, given data *D*, assumptions *A* and any other available knowledge *K*.
$\pi(\theta\|C)$	a possibly improper prior function for θ.
$\pi(\theta\|D, C)$	the posterior distribution of θ after data *D* have been observed, obtained by formal use of Bayes' theorem with the prior function $\pi(\theta\|C)$.
$\pi(\theta\|M, P)$	reference prior for θ given model *M* and class *P* of candidate priors.
S(ω)	inverse of Fisher information matrix *F*(ω).
St(*x*|μ, σ, α)	the PDF of a student distribution for *x*, with location parameter μ scale parameter σ and α degrees of freedom.
$\delta(\tilde{\theta}, \theta)$	intrinsic discrepancy between $p(D\|\theta)$ and $p(D\|\tilde{\theta})$.
$t = t(D)$	a general statistic, function of the data *D*.
U(*a*, θ)	The utility of action *a* when θ occurs.
{*x*, *s2*}	The mean and the variance of a set of observations {x_1,. . ., *xn*}.
$\omega \in \Omega$	a general parameter vector in a model $M \equiv \{p(x\|\omega), x \in X, \omega \in \Omega\}$.

1.08.1 Introduction and Notation

Experimental results usually consist of sets of data of the general form $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where the $\boldsymbol{x}_i$s are somewhat 'homogeneous' (possibly multidimensional) observations. 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 behavior in future instances of the same process. A basic element of any statistical analysis is the specification of a probability model, which is assumed to describe the mechanism that has generated the observed data *D* as a function of a (possibly multidimensional) parameter (vector) $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, sometimes referred to as the state of nature, about which only limited information (if any) is available. All derived statistical conclusions are 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 procedures are tried; see Lindley[1] and Jaynes[2] 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,[3] Jeffreys[4], or de Finetti[5], are actually entitled 'Probability Theory'. The practical consequences of adopting the Bayesian paradigm are far reaching. Indeed, Bayesian methods (1) reduce statistical inference to problems in probability theory, thereby minimizing the need for completely new concepts, and (2) 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 axiomatic 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 frequentist 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 chapter 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 (generally data) and bold italic greek fonts are used for unobservable random vectors (typically parameters); lowercase is used for variables and calligraphic uppercase for their dominion sets. Moreover, the standard mathematical convention of referring to functions, say f and g of $\boldsymbol{x} \in \mathcal{X}$, respectively by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$, will be used throughout. Thus, $p(\boldsymbol{\theta}|D, C)$ and $p(\boldsymbol{x}|\boldsymbol{\theta}, C)$ respectively represent general probability densities of the unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ given data D and conditions C, and of the observable random vector $\boldsymbol{x} \in \mathcal{X}$ conditional on $\boldsymbol{\theta}$ and C. Hence, $p(\boldsymbol{\theta}|D, C) \geq 0$, $\int_{\Theta} p(\boldsymbol{\theta}|D, C)\, \mathrm{d}\boldsymbol{\theta} = 1$ and $p(\boldsymbol{x}|\boldsymbol{\theta}, C) \geq 0$, $\int_{\mathcal{X}} p(\boldsymbol{x}|\boldsymbol{\theta}, C)\, \mathrm{d}\boldsymbol{x} = 1$. Bayesian statistics often make use of improper prior functions for the unknown parameters, that is positive functions whose integral over their dominion is not finite; possibly improper prior functions will be denoted by $\pi(\boldsymbol{\theta}|C)$ and their corresponding posterior densities given data D and conditions C (obtained by formal use of Bayes' theorem) will be denoted by $\pi(\boldsymbol{\theta}|D, C)$. 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 by $\mathrm{N}(x|\mu, \sigma)$.

Bayesian methods make frequent use of the concept of logarithmic divergence, a very general measure of the goodness of the approximation of a probability density $p(\boldsymbol{x})$ by another density $\hat{p}(\boldsymbol{x})$. The Kullback–Leibler or logarithmic divergence of a probability density $\hat{p}(\boldsymbol{x})$ of the random vector $\boldsymbol{x} \in \mathcal{X}$ from its true probability density $p(\boldsymbol{x})$, is defined as

$$\kappa\{\hat{p}(\boldsymbol{x})|p(\boldsymbol{x})\} = \int_{\mathcal{X}} p(\boldsymbol{x}) \log\{p(\boldsymbol{x})/\hat{p}(\boldsymbol{x})\}\, \mathrm{d}\boldsymbol{x} \tag{1}$$

It may be shown that (1) the logarithmic divergence is nonnegative (and it is zero if, and only if, $\hat{p}(\boldsymbol{x}) = p(\boldsymbol{x})$ almost everywhere), and (2) that $\kappa\{\hat{p}(\boldsymbol{x})|p(\boldsymbol{x})\}$ is invariant under one-to-one transformations of $\boldsymbol{x}$.

This chapter contains a brief summary of the foundations of Bayesian statistical methods (Section 1.08.2), an overview of the paradigm (Section 1.08.3), a detailed discussion of objective Bayesian methods (Section 1.08.4), a description of useful objective inference summaries, including estimation and hypothesis testing (Section 1.08.5) and some concluding remarks (Section 1.08.6).

Good pioneering introductions to objective Bayesian statistics include Lindley,[6] Zellner,[7] and Box and Tiao.[8] For more advanced monographs, see Berger[9] and Bernardo and Smith.[10] Bayesian works with specific reference to chemometrics include Rubin[11], Dryden *et al.*[12], and the review by Chen *et al.*[13]

1.08.2 Axiomatic 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 under 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 axiomatic arguments which prove the mathematical inevitability of the use of probability distributions to describe uncertainties; these are summarized later in this section.

1.08.2.1 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 that 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 under 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 under 'similar' conditions. The following examples are intended to illustrate the use of probability as a conditional measure of uncertainty.

1.08.2.1.1 Probabilistic diagnosis

An industrial production is known to contain 0.2% of products, which will eventually fail within their guaranteed period (faulty items). A particular item, randomly selected from that population, is subject to a test which is known to yield positive results, indicating the likely existence of defects, in 98% of faulty items and in 1% of nonfaulty items, so that, if F denotes the event that an item will fail within its guaranteed period and $+$ denotes a positive test result, $\Pr(+|F) = 0.98$ and $\Pr(+|\overline{F}) = 0.01$. Suppose that the result of the test turns out to be positive. Clearly, one is then interested in $\Pr(F|+, A, K)$, the probability that the item will fail, given the positive test result, the assumptions A about the probability mechanism generating the test results, and the available knowledge K of the proportion of faulty items in the population under study (described here by $\Pr(F|K) = 0.002$). An elementary exercise in probability algebra, which involves Bayes' theorem in its simplest form (see Section 1.08.3), yields $\Pr(F|+, A, K) = 0.164$. Notice that all the four probabilities involved in the problem have the same interpretation: they are all conditional measures of uncertainty. Besides, $\Pr(F|+, A, K)$ is both a measure of the uncertainty associated with the event that the particular item which tested positive is actually faulty, and an estimate of the proportion of items in that population (about 16.4%) that would eventually prove to be faulty among those which yielded a positive test.

1.08.2.1.2 Estimation of a proportion

A survey is conducted to estimate the proportion θ of items in a population which 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 behavior 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 screening test for a particular defect where 100 items have been tested, none of which has turned out to have that defect, one may conclude that $\Pr(\theta < 0.01|0, 100, A, K) = 0.844$, that is a probability of 0.844 that the proportion of items with that defect is smaller than 1%.

1.08.2.1.3 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 behavior 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 $\mathrm{m\,s^{-2}}$) 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 in the laboratory may be well 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 was 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.

1.08.2.1.4 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_is, 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 that produces automobile restraint systems may report something like $\Pr(r = 0 | r_1 = \cdots = r_{10} = 0, A, K) = 0.953$, 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 desirable 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.

1.08.2.1.5 Regression

Often, additional information of future observations is provided by related covariates. For instance, after observing the outputs $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ which correspond to a sequence $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of different production conditions, it may be desired to forecast the output $\boldsymbol{y}$ which would correspond to a particular set of production conditions $\boldsymbol{x}$. For example, the viscosity of commercially 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 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, the production engineers will require probabilities of the form $\Pr(a < y < b | \boldsymbol{x}, (y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n), A, K)$. 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 $\boldsymbol{x}$ will actually satisfy the required viscosity specifications.

1.08.2.2 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 $c(a, \boldsymbol{\theta}) \in \mathcal{C}_a$, $\boldsymbol{\theta} \in \Theta_a$, be the consequence of having chosen action a when event $\boldsymbol{\theta}$ 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. All these 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,[14] Savage,[15] DeGroot,[16] Bernardo and Smith,[10] and references therein), but all of them lead basically to the same conclusions, namely

(1) Preferences among consequences should be measured with a real-valued utility function $U(c) = U(a, \boldsymbol{\theta})$, which specifies, on some numerical scale, their desirability.
(2) The uncertainty of relevant events should be measured with a set of probability distributions $\{(p(\boldsymbol{\theta}|C, a), \boldsymbol{\theta} \in \Theta_a), a \in \mathcal{A}\}$ describing their plausibility given the conditions C under which the decision must be taken.
(3) The desirability of each action is measured by its corresponding expected utility,

$$\bar{U}(a|C) = \int_{\Theta_a} U(a, \boldsymbol{\theta}) p(\boldsymbol{\theta}|C, a)\, \mathrm{d}\boldsymbol{\theta}, \quad a \in \mathcal{A} \tag{2}$$

It is often convenient to work in terms of the nonnegative loss function defined by

$$L(a, \boldsymbol{\theta}) = \sup_{a \in \mathcal{A}} \{U(a, \boldsymbol{\theta})\} - U(a, \boldsymbol{\theta}) \tag{3}$$

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

$$\bar{L}(a|C) = \int_{\Theta_a} L(a, \boldsymbol{\theta}) p(\boldsymbol{\theta}|C, a)\, \mathrm{d}\boldsymbol{\theta}, \quad a \in \mathcal{A} \tag{4}$$

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 $\boldsymbol{\theta}$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, (1) 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[14]), and (2) it has been shown (Bernardo[17]) that statistical inference on $\boldsymbol{\theta}$ actually has the mathematical structure of a decision problem, where the class of alternatives is the functional space of the conditional probability distributions $p(\boldsymbol{\theta}|D)$ of $\boldsymbol{\theta}$ given the data, and the utility function is a measure of the amount of information about $\boldsymbol{\theta}$ which the data may be expected to provide.

Another, independent argument for the necessary existence of prior distributions which does not make use of decision theoretical ideas is based on the concept of exchangeability and the related representation theorems. For details, see Bernardo and Smith[10], and references therein.

1.08.3 Bayesian Methodology

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 $\{p(D|\boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ assumed to represent, for some (unknown) value of $\boldsymbol{\omega}$, the probabilistic mechanism which has generated the observed data D. The argument outlined in Section 1.08.2 establishes the logical need to assess a prior probability distribution $p(\boldsymbol{\omega}|K)$ over the parameter space $\boldsymbol{\Omega}$, describing the available knowledge K on the value of $\boldsymbol{\omega}$ 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 $\boldsymbol{\omega}$ after the data D have been observed is contained in the corresponding posterior distribution whose probability density, $p(\boldsymbol{\omega}|D, A, K)$, is immediately obtained from Bayes' theorem,

$$p(\boldsymbol{\omega}|D, A, K) = \frac{p(D|A, \boldsymbol{\omega})p(\boldsymbol{\omega}|K)}{\int_{\Omega} p(D|A, \boldsymbol{\omega})p(\boldsymbol{\omega}|K)\, \mathrm{d}\boldsymbol{\omega}} \tag{5}$$

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 $\boldsymbol{\omega}$ with zero prior density will have zero posterior density. Thus, it is typically assumed (by appropriate restriction, if necessary, of the parameter space $\boldsymbol{\Omega}$) that prior distributions are strictly positive (as Savage[15] 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 $\boldsymbol{\omega}$ 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 assumsed 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

$$\Pr(\theta_i|D) = \frac{p(D|\theta_i)\Pr(\theta_i)}{\sum_{j=1}^{m} p(D|\theta_j)\Pr(\theta_j)}, \quad i = 1, \ldots, m \tag{6}$$

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

An important, frequent application of this simple technique is provided by probabilistic diagnosis. For example, consider again the simple situation where a particular test designed to detect a potentially faulty item is known from laboratory research to give a positive result in 98% of faulty items and in 1% of nonfaulty items. Then, the posterior probability that an item which tested positive is faulty is given by $\Pr(F|+) = (0.98p)/\{0.98p + 0.01(1-p)\}$ as a function of $p = \Pr(F)$, the prior probability of an item being faulty. As one would expect, the posterior probability is 0 only if the prior probability is 0 (so that it is known that the production is free of defects) and it is 1 only if the prior probability is 1 (so that it is known that all the population is faulty). Notice that if the proportion of faulty items is small, then the posterior probability of a randomly chosen item being faulty will be relatively low even if the test is positive. Indeed, say for $\Pr(F) = 0.002$, one finds $\Pr(F|+) = 0.164$, so that in a population where just 0.2% of items are faulty only 16.4% of those testing positive within a random sample will actually prove to be faulty: most positives would actually be false positives.

1.08.3.1 The Learning Process

In this section, the learning process described by Bayes' theorem is described in some detail, discussing its implementation in the presence of nuisance parameters, showing how it can be used to forecast the value of future observations and analyzing its large sample behavior.

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 made easier by noting that Bayes' theorem may be expressed as

$$p(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})p(\boldsymbol{\omega}) \tag{7}$$

(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 $\left[\int_\Omega p(D|\boldsymbol{\omega})p(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}\right]^{-1}$ may always be deduced from the fact that $p(\boldsymbol{\omega}|D)$, a probability density, must integrate to 1. Hence, to identify the form of a posterior distribution it suffices to identify a kernel of the corresponding probability density, that is a function $k(\boldsymbol{\omega})$ such that $p(\boldsymbol{\omega}|D) = c(D)k(\boldsymbol{\omega})$ for some $c(D)$ which does not involve $\boldsymbol{\omega}$. In the examples which follow, this technique will often be used.

An improper prior function is defined as a positive function $\pi(\boldsymbol{\omega})$ such that $\int_\Omega \pi(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}$ is not finite. Equation (7), the formal expression of Bayes' theorem, remains technically valid if $p(\boldsymbol{\omega})$ is actually an improper prior function $\pi(\boldsymbol{\omega})$, provided that $\int_\Omega p(D|\boldsymbol{\omega})\pi(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} < \infty$, thus leading to a well-defined proper posterior density $\pi(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})\pi(\boldsymbol{\omega})$. In particular, as will be justified later (Section 1.08.4), it also remains philosophically valid if $\pi(\boldsymbol{\omega})$ is an appropriately chosen reference (typically improper) prior function. We will use the generic notation $\pi(\boldsymbol{\omega})$ for a possibly improper prior function of $\boldsymbol{\omega}$ and $\pi(\boldsymbol{\omega}|D)$ for the corresponding posterior density (whose propriety should always be verified).

Considered as a function of $\boldsymbol{\omega}$, $l(\boldsymbol{\omega}|D) = p(D|\boldsymbol{\omega})$ 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 (7) that, provided the same (proper or improper) prior function $\pi(\boldsymbol{\omega})$ is used, two different data sets D_1 and D_2, with possibly different probability models $p_1(D_1|\boldsymbol{\omega})$ and $p_2(D_2|\boldsymbol{\omega})$ but yielding proportional likelihood functions, will produce identical posterior distributions for $\boldsymbol{\omega}$. 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 function $\pi(\boldsymbol{\omega})$, 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 $\boldsymbol{\omega}$ 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 $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are sequentially presented, the final result will be the same whether data are globally or sequentially processed. Indeed, $\pi(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1}) \propto p(\boldsymbol{x}_{i+1}|\boldsymbol{\omega})\pi(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$ 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 $\pi(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1})$ will be more concentrated around the true value of $\boldsymbol{\omega}$ than $\pi(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$. However, this is not always the case: occasionally, a 'surprising' observation will increase, rather than decrease, the uncertainty about the value of $\boldsymbol{\omega}$. 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 (i.e., 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 $\boldsymbol{t} = \boldsymbol{t}(D)$ is a sufficient statistic in the sense that, given the model, $\boldsymbol{t}(D)$ contains all information about $\boldsymbol{\omega}$ which is available in D. Formally, $\boldsymbol{t} = \boldsymbol{t}(D)$ is sufficient if (and only if) there exist nonnegative functions f and g such that the likelihood function may be factorized in the general form $p(D|\boldsymbol{\omega}) = f(\boldsymbol{\omega}, \boldsymbol{t})g(D)$. A sufficient statistic always exists, for $\boldsymbol{t}(D) = D$ 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 $\boldsymbol{t}$ is sufficient, the posterior distribution of $\boldsymbol{\omega}$ only depends on the data D through $\boldsymbol{t}(D)$, and may be directly computed in terms of $p(\boldsymbol{t}|\boldsymbol{\omega})$, so that $\pi(\boldsymbol{\omega}|D) = \pi(\boldsymbol{\omega}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\omega})\pi(\boldsymbol{\omega})$.

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 $t(D) = \{r, n\}$ is sufficient. Suppose that prior knowledge about θ is described by a Beta distribution $\text{Be}(\theta|\alpha, \beta)$, so that $p(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$. Using Bayes' theorem, the posterior density of θ is $p(\theta|r, n, \alpha, \beta) \propto \theta^r(1-\theta)^{n-r}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}$ the Beta distribution $\text{Be}(\theta|r+\alpha, n-r+\beta)$.

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 $\text{Be}(\theta|50, 50)$, 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.

A random survey of size 1500 is then conducted, where only 720 citizens declare to be in favor 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 1**; 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 1**) 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 0.933.

1.08.3.1.1 Nuisance parameters

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

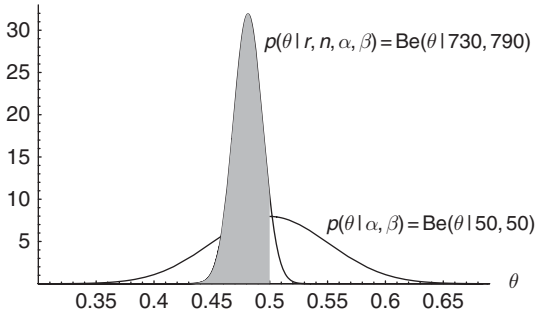

Figure 1 Prior and posterior densities of the proportion θ of citizens who would vote in favor of a referendum.

$\boldsymbol{\theta}(\boldsymbol{\omega})$ is a one-to-one transformation of $\boldsymbol{\omega}$. Using standard change-of-variable probability techniques, the posterior density of $\boldsymbol{\psi}$ is

$$\pi(\boldsymbol{\psi}|D) = \pi(\boldsymbol{\theta}, \boldsymbol{\lambda}|D) = \left[\frac{\pi(\boldsymbol{\omega}|D)}{|\mathcal{J}(\boldsymbol{\omega})|}\right]_{\boldsymbol{\omega}=\boldsymbol{\omega}(\boldsymbol{\psi})} \tag{8}$$

and the required posterior of $\boldsymbol{\theta}$ is the appropriate marginal density, obtained by integration over the nuisance parameter $\boldsymbol{\lambda}$,

$$\pi(\boldsymbol{\theta}|D) = \int_{\Lambda} \pi(\boldsymbol{\theta}, \boldsymbol{\lambda}|D)\, \mathrm{d}\boldsymbol{\lambda} \tag{9}$$

Notice that elimination of unwanted nuisance parameters, a simple integration within the Bayesian paradigm is, however, a difficult (often polemic) problem for frequentist statistics. For further details on the elimination of nuisance parameters see Liseo.[18]

1.08.3.1.2 Domain restrictions

Sometimes the range of possible values of $\boldsymbol{\omega}$ is effectively restricted by contextual considerations. If $\boldsymbol{\omega}$ 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

$$\pi(\boldsymbol{\omega}|D, \boldsymbol{\omega} \in \boldsymbol{\Omega}_c) = \frac{\pi(\boldsymbol{\omega}|D)}{\displaystyle\int_{\Omega_c} \pi(\boldsymbol{\omega}|D)}, \quad \boldsymbol{\omega} \in \Omega_c \tag{10}$$

and obviously vanishes if $\boldsymbol{\omega} \notin \Omega_c$. 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 that satisfy the required condition. Incorporation of known constraints on the parameter values, a simple renormalization within the Bayesian paradigm, is another very difficult problem for conventional statistics.

Example 3. *(Inference on normal parameters)*
Let $D = \{x_1, \ldots, x_n\}$ 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 $n\bar{x} = \Sigma_i x_i$ and $ns^2 = \Sigma_i (x_i - \bar{x})^2$. It may be shown (see Section 1.08.4) 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

$$\pi(\mu, \sigma|D) \propto \sigma^{-(n+1)}\exp[-n\{s^2 + (\bar{x}-\mu)^2\}/(2\sigma^2)] \tag{11}$$

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

$$\pi(\mu|D) \propto \int_0^{\infty} \sigma^{-(n+1)}\exp\left[-\frac{n}{2\sigma^2}[s^2 + (\bar{x}-\mu)^2]\right] \mathrm{d}\sigma \propto [s^2 + (\bar{x}-\mu)^2]^{-n/2} \tag{12}$$

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,

$$\pi(\sigma|D) \propto \int_{-\infty}^{\infty} \sigma^{-(n+1)}\exp\left[-\frac{n}{2\sigma^2}[s^2 + (\bar{x}-\mu)^2]\right] \mathrm{d}\mu \propto \sigma^{-n}\exp\left[-\frac{ns^2}{2\sigma^2}\right] \tag{13}$$

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

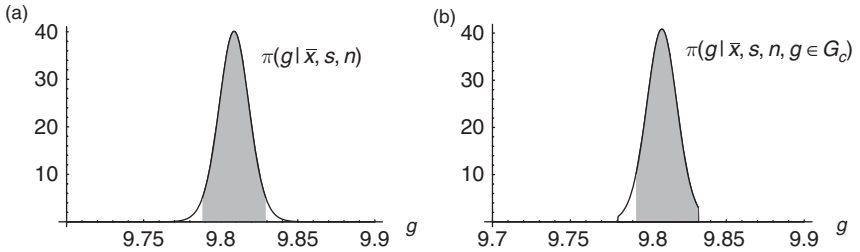

Figure 2 Posterior density $\pi(g|m, s, n)$ of the value g of the gravitational field, given $n = 20$ normal measurements with mean $m = 9.8087$ and standard deviation $s = 0.0428$. (a) With no additional information, and (b) with g restricted to $G_c = \{g;\ 9.7803 < g < 9.8322\}$. Shaded areas represent 95%-credible regions of g.

A frequent example of this scenario is provided by laboratory measurements made under 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 s^{-2}) a mean $\bar{x} = 9.8087$ and a standard deviation $s = 0.0428$. Using no other information, the corresponding posterior distribution is $p(g|D) = \mathrm{St}(g|9.8087, 0.0098, 19)$ represented in **Figure 2(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 2(a)**, 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 s^{-2} (average value at the equator) and 9.8322 m s^{-2} (average value at the poles). The updated posterior distribution will then be

$$\pi(g|D, g \in G_c) = \frac{\mathrm{St}(g|m, s/\sqrt{n-1}, n)}{\displaystyle\int_{g \in G_c} \mathrm{St}(g|m, s/\sqrt{n-1}, n)}, \quad g \in G_c \tag{14}$$

and zero if $g \notin G_c$, where $G_c = \{g,\ 9.7803 < g < 9.8322\}$. This is represented in **Figure 2(b)**. One-dimensional numerical integration may be used to obtain a new 0.95-credible interval; indeed $\Pr(g > 9.792|D, g \in G_c) = 0.95$, the shaded region in **Figure 2(b)**. Moreover, if inferences about the standard deviation σ of the measurement procedure are also requested, the corresponding posterior distribution is found to be

$$\pi(\sigma|D) = 2\sigma^{-3}\mathrm{Ga}(\sigma^{-2}|9.5, 0.0183) \tag{15}$$

This has a mean $\mathrm{E}[\sigma|D] = 0.0458$ and yields $\Pr(0.0334 < \sigma < 0.0642|D) = 0.95$.

1.08.3.2 Predictive Distributions

Let data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \chi$, be a random sample from some distribution in the family $\{p(\boldsymbol{x}|\boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, $\pi(\boldsymbol{\omega})$ a (possibly improper) prior function describing available information (if any) on the value of $\boldsymbol{\omega}$, and consider now a situation where it is desired to predict the value of a future observation $\boldsymbol{x} \in \chi$ generated by the same random mechanism that has generated the data D. It follows from the foundations arguments discussed in Section 1.08.2 that the solution to this prediction problem is simply encapsulated by the predictive distribution $p(\boldsymbol{x}|D)$ describing the uncertainty on the value that $\boldsymbol{x}$ will take, given the information provided by D and any other available knowledge. Since $p(\boldsymbol{x}|\boldsymbol{\omega}, D) = p(\boldsymbol{x}|\boldsymbol{\omega})$, it then follows from standard probability theory that

$$p(\boldsymbol{x}|D) = \int_{\Omega} p(\boldsymbol{x}|\boldsymbol{\omega})\, \pi(\boldsymbol{\omega}|D)\mathrm{d}\boldsymbol{\omega} \tag{16}$$

which is an average of the probability distributions of $\boldsymbol{x}$ conditional on the (unknown) value of $\boldsymbol{\omega}$, weighted with the posterior distribution of $\boldsymbol{\omega}$ given D, $\pi(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})\, \pi(\boldsymbol{\omega})$.

If the assumptions on the probability model are correct, the posterior predictive distribution $p(x|D)$ will converge, as the sample size increases, to the distribution $p(x|\omega)$ that has generated the data. Indeed, about 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 $p(x|D)$ generated by $\pi(\omega|D)$. For a good introduction to Bayesian predictive inference, see Geisser.[19]

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 1.08.4) 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

$$\pi(\lambda|D) \propto \lambda^t e^{-\lambda n}\lambda^{-1/2} \propto \lambda^{t-1/2}e^{-\lambda n} \tag{17}$$

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

$$p(r|D)=\int_0^\infty \mathrm{Pn}(r|\lambda)\mathrm{Ga}(\lambda|t+\frac{1}{2},n)\,\mathrm{d}\lambda = \frac{n^{t+1/2}}{\Gamma(t+1/2)}\frac{1}{r!}\frac{\Gamma(r+t+1/2)}{(1+n)^{r+t+1/2}} \tag{18}$$

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 (18), 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 meteorological conditions remain similar in a given area, $p(r=0|D)=0.953$ would describe the chances of there being no flash flood next year, given that there has been no flash floods in the area for 10 years.

Example 5. (*Prediction in a normal process*)
Consider now an example of prediction of the future value of a continuous observable quantity. 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

$$p(x|D)=\int_0^\infty\int_{-\infty}^\infty \mathrm{N}(x|\mu,\sigma)\pi(\mu,\sigma|D)\,\mathrm{d}\mu\,\mathrm{d}\sigma = \mathrm{St}(x|\bar{x},s\sqrt{\frac{n+1}{n-1}},n-1) \tag{19}$$

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

$$\pi(\mu,\sigma)=\begin{cases}\sigma^{-1} & \text{if } \mu>0\\ 0 & \text{otherwise}\end{cases} \tag{20}$$

However, the result in Equation (19) will still basically 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 $\Pr(x>26|D)=0.9987$. Besides, if production conditions remain constant, 99.87% of the safety belt webbings may be expected to have acceptable breaking strengths.

1.08.3.3 Regression

Let data $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, $\boldsymbol{x}_i \in \chi$, $y_i \in \mathcal{Y}$, be a set of n pairs of probabilistically related observations, so that the observation y_i is assumed to be generated from a distribution $p(y_i|\boldsymbol{x}_i, \boldsymbol{\omega})$, which depends on the known observed vector $\boldsymbol{x}_i$ and on an unknown parameter vector $\boldsymbol{\omega} \in \Omega$, and the likelihood function is

$$p(D|\boldsymbol{\omega}) = \prod_{j=1}^{n} p(y_i|x_i, \boldsymbol{\omega}) \tag{21}$$

Let $\pi(\boldsymbol{\omega})$ be a (possibly improper) prior function describing available information (if any) on the value of $\boldsymbol{\omega}$. Consider now a situation where, for some $\boldsymbol{x} \in \chi$, it is desired to predict the value of a future observation $y \in \mathcal{Y}$ generated from $p(y|\boldsymbol{x}, \boldsymbol{\omega})$. It follows again from the foundations arguments discussed in Section 1.08.2 that the solution to this prediction problem is simply encapsulated by the predictive distribution $p(y|\boldsymbol{x}, D)$ describing the uncertainty on the value that y will take, given x and the information provided by D and any other available knowledge. Since $p(y|\boldsymbol{x}, \boldsymbol{\omega}, D) = p(y|\boldsymbol{x}, \boldsymbol{\omega})$, it follows from standard probability theory that

$$p(y|\boldsymbol{x}, D) = \int_{\Omega} p(y|\boldsymbol{x}, \boldsymbol{\omega})\pi(\boldsymbol{\omega}|D)\mathrm{d}\boldsymbol{\omega} \tag{22}$$

which is an average of the probability distributions of y conditional on $\boldsymbol{x}$ and the (unknown) value of $\boldsymbol{\omega}$, weighted with the posterior distribution of $\boldsymbol{\omega}$ given D, $\pi(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})\ \pi(\boldsymbol{\omega})$. If the assumptions on the probability model are correct, the posterior predictive distribution $p(y|\boldsymbol{x}, D)$ will converge, as the sample size n increases, to the distribution $p(y|\boldsymbol{x}, \boldsymbol{\omega})$ that would generate y given $\boldsymbol{x}$.

1.08.3.3.1 The simple linear model

Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a set of n pairs of related real-valued observables, and suppose that the y_is may be assumed to be linearly related to the x_is with normal homoscedastic errors, so that

$$p(D|\alpha, \beta, \sigma) = \prod_{j=1}^{n} \mathrm{N}(y_i|\alpha + \beta x_i, \sigma) \tag{23}$$

As discussed in Section 1.08.4, absence of relevant initial information on the values of α, β, and σ is formally described by the improper prior function $\pi(\alpha, \beta, \sigma) = \sigma^{-1}$. Using Equation (22), this leads to the (posterior) Student t predictive distribution

$$p(y|x, D) = \mathrm{St}(y|\hat{\alpha} + \hat{\beta}x, s\sqrt{\frac{nb(x)}{n-2}}, n-2) \tag{24}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2}, \quad b(x) = 1 + \frac{1}{n}\frac{(x-\bar{x})^2 + s_{xx}^2}{s_{xx}^2} \tag{25}$$

which depends on the set of sufficient statistics

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j, \quad s_{xx}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 \tag{26}$$

$$s_{xy}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - x)(y_j - y), \quad s^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{\alpha} - \hat{\beta}x_j)^2 \tag{27}$$

Notice that the predictive densities are all Student t with $n-2$ degrees of freedom, centered at the regression line $\hat{\alpha} + \hat{\beta}x$, and with a scale parameter which depends on the observed covariate x through the function $b(x)$. As Equation (25) indicates, $b(x)$ attains its minimum value, $(n+1)/n$, when $x = \bar{x}$, where prediction is most precise, and increases as x moves away from $\bar{x}$ so that, as one would expect, prediction is less precise as the covariate x moves away from the data center.

Example 6. *(Calibration)*
In an environmental study, indirect, laser-based automatic Grimm measurements of the contents in the air of micro PM10 particles were to be calibrated with more precise, gravimetric Andersen measurements. A set of $n = 12$ air samples were analyzed yielding the results presented in the following table, and plotted in the left pane of **Figure 3**.

Automatic	48.2	41.4	44.1	50.2	71.2	49.0	14.3	66.2	30.0	36.8	58.1	83.1
Gravimetric	41.2	39.4	38.1	33.4	48.3	35.2	17.5	44.9	24.3	27.8	50.7	68.0

Using Equation (25), the corresponding regression line is $y = 6.089 + 0.668x$. Thus, for small pollution values (below 18 $\mu g\, m^{-3}$), automatic measurements underestimate the correct gravimetric value but, for the important large values, automatic values are noticeably larger than the gravimetric values. For instance, if the observed automatic value is 70 $\mu g\, m^{-3}$ (vertical dashed line in the left pane of the figure), the predictive density of the corresponding gravimetric value (right pane of the figure) is the Student $St(y|52.83, 5.42, 10)$, whose more likely value is 52.83 $\mu g\, m^{-3}$, and has values between 40.74 and 64.92 with probability 0.95 (confidence bands in the left pane at $x = 70$, and shaded region in the right panel of **Figure 3**).

1.08.3.4 Asymptotic Behavior

The behavior of posterior distributions when the sample size is large is now considered. This is important for, at least, two different reasons: (1) asymptotic results provide useful first-order approximations when actual samples are relatively large, and (2) objective Bayesian methods typically depend on the asymptotic properties of the assumed model. Let $D = \{x_1, \ldots, x_n\}$, $x \in \chi$, be a random sample of size n from $\{p(x|\omega), \omega \in \Omega\}$. 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 a probability 1 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 (MLE) $\hat{\omega}$, with a variance matrix which decreases with n as $1/n$.

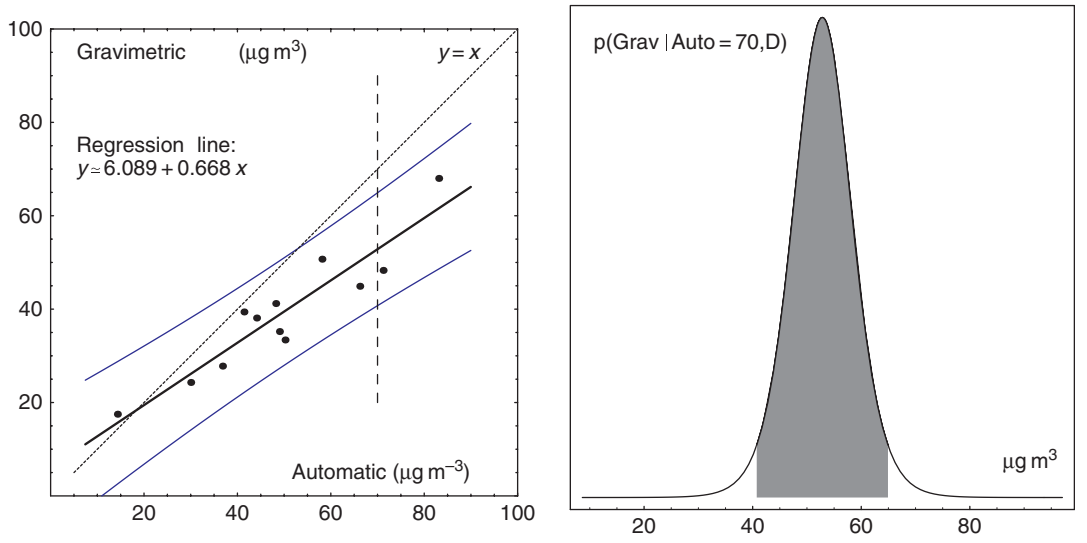

Figure 3 Calibration. Left panel: Data set, regression line and 0.95 credible lines. Right panel: Predictive density of the gravimetric value given an automatic value of 70 $\mu g\, m^{-3}$, with its (shaded) 0.95-credible region.

Consider first the situation where $\Omega = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots\}$ consists of a countable (possibly infinite) set of values, such that the probability model that corresponds to the true parameter value $\boldsymbol{\omega}_t$ is distinguishable from the others, in the sense that the logarithmic divergence $\kappa\{p(\boldsymbol{x}|\boldsymbol{\omega}_i)|p(\boldsymbol{x}|\boldsymbol{\omega}_t)\}$ of each of the $p(\boldsymbol{x}|\boldsymbol{\omega}_i)$ from $p(\boldsymbol{x}|\boldsymbol{\omega}_t)$ is strictly positive. Taking logarithms in Bayes' theorem, defining $z_j = \log[p(\boldsymbol{x}_j|\boldsymbol{\omega}_i)/p(\boldsymbol{x}_j|\boldsymbol{\omega}_t)]$, $j = 1, \ldots, n$, 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

$$\lim_{n\to\infty} \Pr(\boldsymbol{\omega}_t|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 1, \quad \lim_{n\to\infty} \Pr(\boldsymbol{\omega}_i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 0, \quad i \neq t \tag{28}$$

Thus, under appropriate regularity conditions, the posterior probability of the true parameter value converges to 1 as the sample size grows.

Consider now the situation where $\boldsymbol{\omega}$ is a k-dimensional continuous parameter. Expressing Bayes' theorem as $\pi(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \propto \text{explog}[\pi(\boldsymbol{\omega})] + \sum_{j=1}^{n} \log[p(\boldsymbol{x}_j|\boldsymbol{\omega})]$, expanding $\sum_j \log[p(\boldsymbol{x}_j|\boldsymbol{\omega})]$ about its maximum (the MLE $\hat{\boldsymbol{\omega}}$), 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 $\boldsymbol{\omega}$ is the approximate k-variate normal,

$$\pi(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \approx \mathbf{N}_k\{\hat{\boldsymbol{\omega}}, \boldsymbol{S}(D, \hat{\boldsymbol{\omega}})\}, S^{-1}(D, \boldsymbol{\omega}) = \left(-\sum_{l=1}^{n} \frac{\partial^2 \log[p(\boldsymbol{x}_l|\boldsymbol{\omega})]}{\partial \boldsymbol{\omega}_i \partial \boldsymbol{\omega}_j} \right) \tag{29}$$

A simpler, but somewhat poorer, approximation may be obtained by using the strong law of large numbers on the sums in Equation (29) to establish that $\boldsymbol{S}^{-1}(D,\hat{\boldsymbol{\omega}}) \approx n\boldsymbol{F}(\hat{\boldsymbol{\omega}})$, where $\boldsymbol{F}(\boldsymbol{\omega})$ is Fisher's information matrix, with general element

$$\boldsymbol{F}_{ij}(\boldsymbol{\omega}) = -\int_X p(\boldsymbol{x}|\boldsymbol{\omega}) \frac{\partial^2 \log[p(\boldsymbol{x}|\boldsymbol{\omega})]}{\partial \boldsymbol{\omega}_i \partial \boldsymbol{\omega}_j} \, \mathrm{d}\boldsymbol{x} \tag{30}$$

so that

$$\pi(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \approx \boldsymbol{N}_k(\boldsymbol{\omega}|\hat{\boldsymbol{\omega}}, n^{-1}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\omega}})) \tag{31}$$

Thus, under appropriate regularity conditions, the posterior probability density of the parameter vector $\boldsymbol{\omega}$ approaches, as the sample size grows, a multivariate normal density centered at the MLE $\hat{\boldsymbol{\omega}}$, with a variance matrix that 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 above results, the posterior distribution of θ will be the approximate normal,

$$\pi(\theta|r, n) \approx \mathrm{N}(\theta|\hat{\theta}, s(\hat{\theta})/\sqrt{n}), \quad s(\theta) = \{\theta(1-\theta)\}^{1/2} \tag{32}$$

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 (1) the prior $p(\theta)$ is relatively 'flat' in the region where the likelihood function matters, and (2) both r and n are moderately large. If, say, $n = 1500$ and $r = 720$, this leads to $\pi(\theta|D) \approx \mathrm{N}(\theta|0.480, 0.013)$, 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 that corresponds to the prior $\mathrm{Be}(\theta|50, 50)$.

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

$$\boldsymbol{F}(\boldsymbol{\omega}) = \boldsymbol{F}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \begin{pmatrix} \boldsymbol{F}_{\theta\theta}(\boldsymbol{\theta}, \boldsymbol{\lambda}) & \boldsymbol{F}_{\theta\lambda}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ \boldsymbol{F}_{\lambda\theta}(\boldsymbol{\theta}, \boldsymbol{\lambda}) & \boldsymbol{F}_{\lambda\lambda}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \end{pmatrix} \tag{33}$$

and

$$S(\boldsymbol{\theta}, \boldsymbol{\lambda}) = F^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \begin{pmatrix} S_{\theta\theta}(\boldsymbol{\theta}, \boldsymbol{\lambda}) & S_{\theta\lambda}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ S_{\lambda\theta}(\boldsymbol{\theta}, \boldsymbol{\lambda}) & S_{\lambda\lambda}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \end{pmatrix} \tag{34}$$

then the marginal posterior distribution of $\boldsymbol{\theta}$ will be

$$\pi(\boldsymbol{\theta}|D) \approx \mathrm{N}\left\{\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, n^{-1} S_{\theta\theta}\left(\hat{\boldsymbol{\theta}}, \hat{\lambda}\right)\right\} \tag{35}$$

while the conditional posterior distribution of λ given $\boldsymbol{\theta}$ will be

$$\pi(\boldsymbol{\lambda}|\boldsymbol{\theta}, D) \approx \mathrm{N}\left\{\boldsymbol{\lambda}|\hat{\boldsymbol{\lambda}} - \boldsymbol{F}_{\lambda\lambda}^{-1}\left(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}\right) \boldsymbol{F}_{\lambda\theta}\left(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right), n^{-1} \boldsymbol{F}_{\lambda\lambda}^{-1}\left(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}\right)\right\} \tag{36}$$

Notice that $\boldsymbol{F}_{\lambda\lambda}^{-1} = \boldsymbol{S}_{\lambda\lambda}$ if (and only if) $\boldsymbol{F}$ is block diagonal, that is if (and only if) $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ 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 $\mathrm{N}(x|\mu, \sigma)$. 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 $\mathrm{N}(\mu|\bar{x}, s/\sqrt{n})$; this may be compared with the exact result $\pi(\mu|D) = \mathrm{St}(\mu|\bar{x}, s\sqrt{n-1}, n-1)$, previously obtained under the assumption of no prior knowledge.

1.08.4 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 $p(D|\boldsymbol{\omega})$ for some $\boldsymbol{\omega} \in \Omega$, and that the quantity of interest is some real-valued function $\theta = \theta(\boldsymbol{\omega})$ of the model parameter $\boldsymbol{\omega}$. Without loss of generality, it may be assumed that the probability model is of the form

$$\mathcal{M} = \{p(D|\theta, \boldsymbol{\lambda}), D \in \mathcal{D}, \theta \in \Theta, \boldsymbol{\lambda} \in \Lambda\} \tag{37}$$

where $\boldsymbol{\lambda}$ is some appropriately chosen nuisance parameter vector. As described in Section 1.08.3, to obtain the required posterior density of the quantity of interest $p(\theta|D)$, it is necessary to specify a (possibly improper) joint prior function $\pi(\theta, \boldsymbol{\lambda})$. It is now required to identify the form of that joint prior function $\pi_\theta(\theta, \boldsymbol{\lambda}|\mathcal{M}, \mathcal{P})$, the θ-reference prior, which would have a minimal effect on the corresponding posterior distribution of θ,

$$\pi(\theta|D) \propto \int_\Lambda p(D|\theta, \boldsymbol{\lambda}) \pi_\theta(\theta, \boldsymbol{\lambda}|\mathcal{M}, \mathcal{P}) \mathrm{d}\boldsymbol{\lambda} \tag{38}$$

within the class $\mathcal{P}$ of all the prior distributions compatible with whatever information one is prepared to assume about $(\theta, \boldsymbol{\lambda})$, 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 $\pi_\theta(\theta, \boldsymbol{\lambda}|\mathcal{M}, \mathcal{P})$ is often simply denoted by $\pi(\theta, \boldsymbol{\lambda})$, but its dependence on the quantity of interest θ, the assumed model $\mathcal{M}$, and the class $\mathcal{P}$ of priors compatible with the 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 $p(D|\theta,\boldsymbol{\lambda})$, 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[10] (Section 2.17.2), Kass and Wasserman[20] Bernardo and Ramón[21] Bernardo,[22] and references therein. This section focuses 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[23,24] and further developed by Berger and Bernardo,[25–28] Bernardo,[22] Berger *et al.*,[29] and references therein. For a general discussion on 'noninformative' priors, see Bernardo.[30] In the formulation described below, the reference posterior exploits certain well-defined features of a possible prior, namely those describing a situation where 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 that 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 'objective' Bayesian solutions to statistical inference problems in just the same sense that conventional statistical methods claim to be 'objective', that is in the sense that the solutions provided only depend on the assumed model and the observed data.

1.08.4.1 Reference Distributions

1.08.4.1.1 One parameter

Consider an experiment that 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 $\boldsymbol{t} = \boldsymbol{t}(D) \in T$ be any sufficient statistic (which may well be the complete data set D). In Shannon's general information theory, the amount of information $I^{\theta}\{T, p(\theta)\}$ which may be expected to be provided by D, or (equivalently) by $\boldsymbol{t}(D)$, about the value of θ is defined by

$$I^{\theta}\{T, p(\theta)\} = \kappa\{p(\boldsymbol{t})p(\theta)|p(\boldsymbol{t}|\theta)p(\theta)\} = \mathrm{E}_{t}\left[\int_{\Theta} p(\theta|\boldsymbol{t})\log\frac{p(\theta|\boldsymbol{t})}{p(\theta)}\,\mathrm{d}\theta\right] \tag{39}$$

the expected logarithmic divergence of the prior from the posterior. This is naturally a functional of the prior density $p(\theta)$: the larger the prior information, the smaller the information that the data may be expected to provide. The functional $I^{\theta}\{T, p(\theta)\}$ is concave, nonnegative, and invariant under one-to-one transformations of θ. Consider now the amount of information $I^{\theta}\{T^k, p(\theta)\}$ about θ that may be expected from the experiment, which consists of k conditionally independent replications $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k\}$ 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 $I^{\theta}\{T^k, p(\theta)\}$ will approach the missing information about θ associated with the prior $p(\theta)$. Intuitively, a θ-'noninformative' prior is one that maximizes the missing information about θ. Formally, if $p_k(\theta)$ denotes the prior density that maximizes $I^{\theta}\{T^k, p(\theta)\}$ in the class $\mathcal{P}$ of 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 $\pi(\theta|\mathcal{M}, \mathcal{P})$ is the limit as $k \to \infty$ (in a sense to be made precise) of the sequence of priors $\{p_k(\theta),\ k = 1, 2, \ldots\}$.

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 behavior of the assumed probability model, a feature which actually simplifies their actual derivation.

Example 7. *(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 $p(\theta) = \{p_1, \ldots, p_m\}$, with $p_i = \Pr(\theta = \theta_i)$, and there is no topology associated to the parameter space Θ, so that the θ_is are just labels with no quantitative meaning, then the missing information associated to $\{p_1, \ldots, p_m\}$ reduces to

$$\lim_{k\to\infty} I^{\theta}\{T^k, p(\theta)\} = H(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log(p_i) \tag{40}$$

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

Thus, in the non structured finite case, the reference prior $\pi(\theta|\mathcal{M}, \mathcal{P})$ (which in this case is obviously always proper) 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 results when the parameter space is a finite set of nonquantitative 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, $\pi(\theta|\mathcal{M}, \mathcal{P}_0) = \{1/m, \ldots, 1/m\}$.

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 $p_k(\theta)$, which maximize $I^{\theta}\{T^k, p(\theta)\}$ in the precise sense that, for any value of the sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(D)$, the reference posterior, the intrinsic limit $\pi(\theta|\boldsymbol{t})$ of the corresponding sequence of posteriors $\{p_k(\theta|\boldsymbol{t})\}$, may be obtained from $\pi(\theta|\mathcal{M}, \mathcal{P})$ by formal use of Baye's theorem, so that $\pi(\theta|\boldsymbol{t}) \propto p(\boldsymbol{t}|\theta)\pi(\theta|\mathcal{M}, \mathcal{P})$.A sequence $\{p_k(\theta \mid \boldsymbol{t})\}$ of posterior distributions converges intrinsically to a limit $\pi(\theta \mid \boldsymbol{t})$ if the sequence of expected intrinsic discrepancies $\mathrm{E}_{\boldsymbol{t}}[\delta\{p_k(\theta \mid \boldsymbol{t}), \pi(\theta \mid \boldsymbol{t})\}]$ converges to 0, where $\delta\{p, q\} = \min\{k(p \mid q), k(q \mid p)\}$ and $k(p \mid q) = \int_{\Theta} q(\theta) \log[q(\theta)/p(\theta)]\mathrm{d}\theta$. For details, see Berger *et al.*[29]

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 that 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 the data could provide.

If (1) the sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(D)$ is a consistent estimator θ of a continuous parameter θ, and (2) 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 $p(\theta \mid D)$ has an asymptotic approximation of the form $p(\theta|\tilde{\theta}, n)$, the (unrestricted) reference prior is simply

$$\pi(\theta|\mathcal{M}, \mathcal{P}_0) \propto p(\theta|\tilde{\theta}, n)|_{\tilde{\theta}=\theta} \tag{41}$$

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 8. *(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 Equation (41), the reference prior function is $\pi(\theta) \propto s(\theta)^{-1}$. Under regularity conditions (often satisfied in practice, see Section 1.08.3.3), 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 $\hat{\theta}$ the MLE of θ. Hence, the reference prior function under 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, and its posterior distribution is asymptotically normal.

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 $p(x \mid \theta) = \theta^x(1-\theta)^{1-x}$, with $x \in \{0, 1\}$; 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 is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, so that it is actually the (proper) Beta

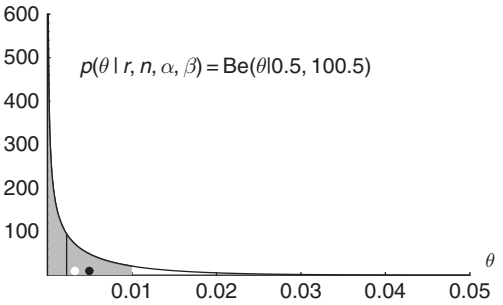

Figure 4 Posterior distribution of the proportion of defective items in a production batch, given the analysis of $n = 100$ items, none of which were defective.

distribution $\text{Be}(\theta|1/2, 1/2)$. Since the reference algorithm is invariant under reparametrization, the reference prior of $\phi(\theta) = 2\text{arcsin}\sqrt{\theta}$ is $\pi(\phi) = \pi(\theta)/|\partial\phi/\partial/\theta| = 1$; thus, the reference prior is uniform on the variance-stabilizing transformation $\phi(\theta) = 2\text{arcsin}\sqrt{\theta}$, a general feature under regularity conditions. In terms of θ, the reference posterior is $\pi(\theta|D) = \pi(\theta|\, r, n) = \text{Be}(\theta\,|r + 1/2, n - r + 1/2)$, where $r = \sum x_j$ is the number of positive trials.

Suppose, for example, that $n = 100$ randomly selected items from a production batch have been tested for a particular defect and that all tested negative so that $r = 0$. The reference posterior distribution of the proportion θ of items with the defect is then the Beta distribution $\text{Be}(\theta|0.5, 100.5)$ represented in **Figure 4**.

Thus, on the basis of the assumed model and the observed experimental results, one may claim that the proportion of defective items is surely smaller than 5% (for the reference posterior probability of the event $\theta > 0.05$ is 0.001), θ is smaller than 0.01 with probability 0.844 (area of the shaded region in **Figure 4**), it is equally likely to be over or below 0.23% (for the median, represented by a vertical line, is 0.0023), and 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 $\Pr(x = 1|r, n) = \text{E}[\theta|r, n] = 0.005$. 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 1.08.5); 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 behavior of the MLE, here $\hat{\theta} = r/n = 0$ for any n, makes absolutely no sense in this scenario.

1.08.4.1.2 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 $p(\boldsymbol{t}|\theta, \lambda), \theta \in \Theta, \lambda \in \Lambda$, and a θ-reference prior $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P})$ is required, the reference algorithm proceeds in two steps:

(1) Conditional on θ, $p(\boldsymbol{t}|\theta, \lambda)$ only depends on the nuisance parameter λ and, hence, the one-parameter algorithm may be used to obtain the conditional reference prior $\pi(\lambda|\theta, \mathcal{M}, \mathcal{P})$.
(2) If $\pi(\lambda|\,\theta, \mathcal{M}, \mathcal{P})$ is proper, this may be used to integrate the nuisance parameter, thus obtaining the one-parameter integrated model $p(\boldsymbol{t}|\theta) = \int_\Lambda p(\boldsymbol{t}|\theta, \lambda)\pi(\lambda|\theta, \mathcal{M}, \mathcal{P})\text{d}\lambda$, 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 $\pi(\theta|\boldsymbol{t}) \propto p(\boldsymbol{t}|\theta)\pi(\theta|\mathcal{M}, \mathcal{P})$.

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 $\{\pi_i(\theta|\boldsymbol{t}\}$ for the quantity of interest θ, and the required reference posterior is the corresponding intrinsic limit $\pi(\theta|\boldsymbol{t}) = \lim_i \pi_i(\theta|\boldsymbol{t})$.

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, that is for any sufficient $\boldsymbol{t} \in \text{T}$ (which may well be the whole data set D) $\pi(\theta|\boldsymbol{t}) \propto \int_\Lambda p(\boldsymbol{t}|\,\theta, \lambda)\pi_\theta(\theta, \lambda)\text{d}\lambda$. The approximating sequences should be consistently chosen within a given model. Thus, given a probability model $\{p(\boldsymbol{x}|\boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, an appropriate approximating sequence $\{\Omega_i\}$ should be chosen for the whole parameter space Ω. If the analysis is done in terms of, say,

$\boldsymbol{\psi} = \{\psi_1, \psi_2\} \in \Psi(\Omega)$, 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 nonexistence of a unique 'noninformative prior' that would be appropriate for any inference problem within a given model was established by Dawid *et al.*[31] 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, $\pi(\theta|D) = \pi(\theta|\boldsymbol{t})$, only depends on the data through a statistic $\boldsymbol{t}$ whose sampling distribution, $p(\boldsymbol{t}|\theta, \lambda) = p(\boldsymbol{t}|\theta)$, only depends on θ, then one would expect the reference posterior to be of the form $\pi(\theta|\boldsymbol{t}) \propto \pi(\theta)p(\boldsymbol{t}|\theta)$ for some prior $\pi(\theta)$. However, examples were found where this cannot be the case if a unique joint 'noninformative' prior was to be used for all possible quantities of interest.

Example 9. *(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, $\boldsymbol{F}(\theta, \lambda)$. Indeed, if

$$\boldsymbol{F}(\theta, \lambda) = \begin{pmatrix} F_{\theta\theta}(\theta, \lambda) & F_{\boldsymbol{\theta}\lambda}(\theta, \lambda) \\ F_{\theta\lambda}(\theta, \lambda) & F_{\lambda\lambda}(\theta, \lambda) \end{pmatrix} \quad \text{and} \quad \boldsymbol{S}(\theta, \lambda) = \boldsymbol{F}^{-1}(\theta, \lambda) \tag{42}$$

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

$$\pi(\lambda|\theta) \propto F_{\lambda\lambda}^{1/2}(\theta, \lambda), \quad \lambda \in \Lambda \tag{43}$$

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

$$\pi(\theta) \propto \exp\left\{\int_\Lambda \pi(\lambda|\theta) \log[\boldsymbol{S}_{\theta\theta}^{-1/2}(\theta, \lambda)]\, \mathrm{d}\lambda\right\}, \quad \theta \in \Theta \tag{44}$$

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 (1) both $F_{\lambda\lambda}^{1/2}(\theta, \lambda)$ and $S_{\theta\theta}^{-1/2}(\theta, \lambda)$ factorize, so that

$$S_{\theta\theta}^{-1/2}(\theta, \lambda) \propto f_\theta(\theta)g_\theta(\lambda), \quad F_{\lambda\lambda}^{1/2}(\theta, \lambda) \propto f_\lambda(\theta)\, g_\lambda(\lambda) \tag{45}$$

and (2) 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 that corresponds to a normal model $\mathrm{N}(x|\mu, \sigma)$ is

$$\boldsymbol{F}(\mu, \sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}, \quad \boldsymbol{S}(\mu, \sigma) = \boldsymbol{F}^{-1}(\mu, \sigma) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix} \tag{46}$$

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 $\boldsymbol{F}(\mu,\sigma)$ 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 $\boldsymbol{F}(\phi,\sigma)=\mathcal{J}^t\boldsymbol{F}(\mu,\sigma)\mathcal{J}$, where $\mathcal{J}=(\partial(\mu,\sigma)/\partial(\phi,\sigma))$ is the Jacobian of the inverse transformation; this yields

$$\boldsymbol{F}(\phi,\sigma)=\begin{pmatrix}1 & \phi\sigma^{-1}\\ \phi\sigma^{-1} & \sigma^{-2}(2+\phi^2)\end{pmatrix},\quad \boldsymbol{S}(\phi,\sigma)=\begin{pmatrix}1+\frac{1}{2}\phi^2 & -\frac{1}{2}\phi\sigma\\ -\frac{1}{2}\phi\sigma & \frac{1}{2}\sigma^2\end{pmatrix} \tag{47}$$

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 9, $\pi_\phi(\phi,\sigma|\mathcal{M},\mathcal{P}_0)=(1+\frac{1}{2}\phi^2)^{-1/2}\sigma^{-1}$. 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 $\pi_\mu(\mu,\sigma\mid\mathcal{M},\mathcal{P}_0)=\pi_\sigma(\mu,\sigma\mid\mathcal{M},\mathcal{P}_0)=\sigma^{-1}$. 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 $t=\left(\sum x_j\right)/\left(\sum x_j^2\right)^{1/2}$, 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.

1.08.4.1.3 Many parameters

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

$$\pi(\theta_m|\theta_{m-1},\ldots,\theta_1)\times\cdots\times\pi(\theta_2|\theta_1)\times\pi(\theta_1) \tag{48}$$

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

Example 10. *(Stein's paradox)*
Let D be a random sample from an m-variate normal distribution with mean $\boldsymbol{\mu}=\{\mu_1,\ldots,\mu_m\}$ and unitary variance matrix. The reference prior that corresponds to any permutation of the μ_is is uniform, and this prior leads indeed to appropriate reference posterior distributions for any of the μ_is, 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 squared distance of $\boldsymbol{\mu}$ to the origin. As shown by Stein,[32] 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_is, is actually highly informative on the sum of their squares, introducing a severe positive bias (Stein's paradox). However, the reference prior that corresponds to a parametrization of the form $\{\theta,\lambda_1,\ldots,\lambda_{m-1}\}$ produces, for any choice of the nuisance parameters $\lambda_i=\lambda_i(\boldsymbol{\mu})$, 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 behavior 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.

1.08.4.1.4 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 11. *(Univariate restricted reference priors)*
If the probability mechanism which is assumed to have generated the available data only depends on a univariate 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

$$\mathcal{P} = \left\{ p(\theta); \quad p(\theta) > 0, \int_{\Theta} p(\theta)\, \mathrm{d}\theta = 1, \quad \int_{\Theta} g_i(\theta) p(\theta)\, \mathrm{d}\theta = \beta_i, \ i = 1, \ldots, m \right\} \tag{49}$$

then the (restricted) reference prior is

$$\pi(\theta|\mathcal{M}, \mathcal{P}) \propto \pi(\theta|\mathcal{M}, \mathcal{P}_0) \exp\left[\sum_{j=1}^{m} \gamma_i g_i(\theta)\right] \tag{50}$$

where $\pi(\theta|\mathcal{M}, \mathcal{P}_0)$ is the unrestricted reference prior and the γ_is 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 it is further assumed that $\mathrm{E}[\theta] = \mu_0$ and $\mathrm{Var}[\theta] = \sigma_0^2$. The unrestricted reference prior for any regular location problem may be shown to be uniform, so that here $\pi(\theta|\mathcal{M}, \mathcal{P}_0) = 1$. Thus, the restricted reference prior must be of the form $\pi(\theta|\mathcal{M}, \mathcal{P}) \propto \exp\{\gamma_1 \theta + \gamma_2 (\theta - \mu_0)^2\}$, with $\int_{\Theta} \theta \pi(\theta|\mathcal{M}, \mathcal{P})\, \mathrm{d}\theta = \mu_0$ and $\int_{\Theta} (\theta - \mu_0)^2 \pi(\theta|\mathcal{M}, \mathcal{P})\, \mathrm{d}\theta = \sigma_0^2$. Hence, this is the normal distribution with the specified mean and variance, $\pi(\theta|\mathcal{M}, \mathcal{P}) = \mathrm{N}(\theta|\mu_0, \sigma_0)$.

1.08.4.2 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 behavior 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.

1.08.4.2.1 Point estimation

It is generally accepted that, as the sample size increases, a 'good' estimator $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ ought to get the correct value of $\boldsymbol{\theta}$ eventually, that is to be consistent. Under appropriate regularity conditions, any Bayes estimator $\boldsymbol{\phi}^*$ of any function $\boldsymbol{\phi}(\boldsymbol{\theta})$ converges in probability to $\boldsymbol{\phi}(\boldsymbol{\theta})$, 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 $\boldsymbol{\theta}$) cannot be larger for all $\boldsymbol{\theta}$ values than that corresponding to another estimator. Any proper Bayes estimator is admissible; moreover, as established by Wald,[33] 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 functions. Reference Bayes estimators are typically admissible with respect to reasonable 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 1.08.5.1. In particular, if $r=1$ and $n=2$ the intrinsic estimator of θ^2 is (as one would naturally expect) $(\theta^*)^2=1/4$.

1.08.4.2.2 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 $\boldsymbol{\theta}$ based on a large sample will cover the true value of $\boldsymbol{\theta}$ approximately $100q\%$ of times. Detailed results are readily available for univariate problems. For instance, consider the probability model $\{p(D|\boldsymbol{\omega}), \boldsymbol{\omega}\in\Omega\}$; let $\theta=\theta(\boldsymbol{\omega})$ be any univariate quantity of interest, and let $\boldsymbol{t}=\boldsymbol{t}(D)\in T$ be any sufficient statistic. If $\theta_q(\boldsymbol{t})$ denotes the $100q\%$ quantile of the posterior distribution of θ, which corresponds to some unspecified prior, so that

$$\Pr[\theta\le\theta_q(\boldsymbol{t})|\boldsymbol{t}]=\int_{\theta\le\theta_q(\boldsymbol{t})}\pi(\theta|\boldsymbol{t})\,\mathrm{d}\theta=q \tag{51}$$

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

$$\Pr[\theta_q(\boldsymbol{t})\ge\theta|\boldsymbol{\omega}]=\int_{\theta_q(\boldsymbol{t})\ge\theta}p(\boldsymbol{t}|\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{t} \tag{52}$$

is such that

$$\Pr[\theta_q(\boldsymbol{t})\ge\theta|\boldsymbol{\omega}]=\Pr[\theta\le\theta_q(\boldsymbol{t})|\boldsymbol{t}]+O(n^{-1/2}) \tag{53}$$

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 probability matching priors see Datta and Sweeting[34] and references therein. Reference priors are typically found to be probability matching priors, hence 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 12. *(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, $\phi=\omega_1\omega_2$ (e.g., 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 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 9, with the natural approximating sequence induced by $(\omega_1,\omega_2)\in[-i,i]^2$, the ϕ-reference prior is found to be

$$\pi_\phi(\omega_1,\omega_2|\mathcal{M},\mathcal{P}_0)\propto(n\omega_1^2+m\omega_2^2)^{-1/2} \tag{54}$$

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 an unpublished report to obtain such an approximate agreement. The same example was later used by Efron[35] to stress the fact that, even within a fixed probability model $\{p(D|\boldsymbol{\omega}),\boldsymbol{\omega}\in\Omega\}$, the prior required to make objective inferences about some function of the parameters $\phi=\phi(\boldsymbol{\omega})$ must generally depend on the function ϕ. For further details on the reference analysis of this problem, see Berger and Bernardo.[25]

The numerical agreement between reference Bayesian credible regions and frequentist confidence intervals is actually perfect under special circumstances. Indeed, as Lindley[36] 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. In contrast, this function t is recognized to be precisely the conventional t statistic, whose sampling distribution is well known to also be a 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 varaince. 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.

1.08.5 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 $\boldsymbol{\omega}$ of the parameters which govern the model is contained in the posterior distribution $p(\boldsymbol{\omega}|D, C)$, 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 (1) 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 (2) 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.

1.08.5.1 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 2. 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.

1.08.5.1.1 Point Estimation

Let D be the available data, which are assumed to have been generated by a probability model $\{p(D|\boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, and let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \Theta$ be the quantity of interest. A point estimator of $\boldsymbol{\theta}$ is some function of the data $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(D)$ which could be regarded as an appropriate proxy for the actual, unknown value of $\boldsymbol{\theta}$. Formally, to choose a point estimate for $\boldsymbol{\theta}$ is a decision problem, where the action space is the class Θ of possible

$\boldsymbol{\theta}$ values. From a decision-theoretic perspective, to choose a point estimate $\tilde{\boldsymbol{\theta}}$ of some quantity $\boldsymbol{\theta}$ is a decision to act as though $\tilde{\boldsymbol{\theta}}$ were $\boldsymbol{\theta}$, not to assert something about the value of $\boldsymbol{\theta}$ (although desire to assert something simple may well be the reason to obtain an estimate). As prescribed by the foundations of decision theory (Section 1.08.2), to solve this decision problem it is necessary to specify a loss function $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ measuring the consequences of acting as if the true value of the quantity of interest were $\tilde{\boldsymbol{\theta}}$, when it is actually $\boldsymbol{\theta}$. The expected posterior loss if $\tilde{\boldsymbol{\theta}}$ were used is

$$\bar{L}[\tilde{\boldsymbol{\theta}}\,|D] = \int_{\Theta} L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D)\,\mathrm{d}\boldsymbol{\theta} \tag{55}$$

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

Example 13. *(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 $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$, then the Bayes estimator is the posterior mean $\boldsymbol{\theta}^* = E[\boldsymbol{\theta}|D]$, assuming that the mean exists. Similarly, if the loss function is a zero-one function, so that $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$ if $\tilde{\boldsymbol{\theta}}$ belongs to a ball or radius ϵ centered in $\boldsymbol{\theta}$ and $L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 1$ otherwise, then the Bayes estimator $\boldsymbol{\theta}^*$ 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 functions 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 14. *(Intrinsic estimation)*
Conventional loss functions are typically noninvariant under reparametrization. It follows that the Bayes estimator $\boldsymbol{\phi}^*$ of a one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\boldsymbol{\phi}(\boldsymbol{\theta}^*)$ (the univariate posterior median, which is invariant, is an interesting exception). Moreover, conventional loss functions focus on the 'distance' between the estimate $\tilde{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}$, rather than on the 'distance' between the probability models they label. Inference-oriented loss functions directly focus on how different the probability model $p(D|\boldsymbol{\theta}, \boldsymbol{\lambda})$ is from its closest approximation within the family $\{p(D|\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}_i), \boldsymbol{\lambda}_i \in \Lambda\}$ and typically produce invariant solutions. An attractive example is the intrinsic discrepancy, $\delta(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$, defined as the minimum logarithmic divergence between a probability model labeled by $\boldsymbol{\theta}$ and a probability model labeled by $\tilde{\boldsymbol{\theta}}$. When there are no nuisance parameters, this is given by

$$\delta(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \min\left\{\kappa(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}), \kappa(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})\right\}, \quad \kappa(\boldsymbol{\theta}_i|\boldsymbol{\theta}) = \int_T p(\boldsymbol{t}|\boldsymbol{\theta}) \log \frac{p(\boldsymbol{t}|\boldsymbol{\theta})}{p(\boldsymbol{t}|\boldsymbol{\theta}_i)}\,\mathrm{d}t \tag{56}$$

where $\boldsymbol{t} = \boldsymbol{t}(D) \in T$ 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,

$$\delta(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda}_i \in \Lambda} \delta(\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}_i, \boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{57}$$

measures the logarithmic divergence from $p(t|\boldsymbol{\theta}, \boldsymbol{\lambda})$ of its closest approximation with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, and the loss function now depends on the complete parameter vector $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. 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 $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from some model $p(\boldsymbol{x}|\boldsymbol{\theta})$, then $\kappa(\boldsymbol{\theta}_i|\boldsymbol{\theta}) = n \int_X p(\boldsymbol{x}|\boldsymbol{\theta}) \log[p(\boldsymbol{x}|\boldsymbol{\theta})/p(\boldsymbol{x}|\boldsymbol{\theta}_i)]\mathrm{d}\boldsymbol{x}$

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, nonnegative 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 $p(D|\boldsymbol{\theta}, \boldsymbol{\lambda})$ and its closest approximation within the class $\{p(D|\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}_i), \boldsymbol{\lambda}_i \in \Lambda\}$. Moreover, it is invariant under one-to-one reparametrization of the parameter of interest $\boldsymbol{\theta}$, and does not depend on the choice of the nuisance parameter $\boldsymbol{\lambda}$. The intrinsic estimator is naturally obtained by minimizing the reference posterior expected intrinsic discrepancy

$$\mathrm{d}(\tilde{\boldsymbol{\theta}}|D) = \int_{\Lambda} \int_{\Theta} \delta(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \pi(\boldsymbol{\theta}, \boldsymbol{\lambda}|D)\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\lambda} \tag{58}$$

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[37] and Bernardo.[38]

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 $\phi(E[\theta \mid D])$. The intrinsic loss function in this problem is

$$\delta(\tilde{\theta}, \theta) = n \min\{\kappa(\tilde{\theta}|\theta), \kappa(\theta|\tilde{\theta})\},\ \kappa(\theta_i|\theta) = \theta \log\frac{\theta}{\theta_i} + (1-\theta) \log\frac{1-\theta}{1-\theta_i} \tag{59}$$

and the corresponding intrinsic estimator θ^* is obtained by minimizing the expected posterior loss $\mathrm{d}(\tilde{\theta}|D) = \int \delta(\tilde{\theta}, \theta)\pi(\theta|D)\, \mathrm{d}\theta$. 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 $\phi(E[\theta|D])$.

1.08.5.1.2 Interval estimation

To describe the inferential content of the posterior distribution of the quantity of interest $p(\boldsymbol{\theta}|D)$, it is often convenient to quote regions $R \subset \Theta$ of given probability under $p(\boldsymbol{\theta}|D)$. 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 $p(\boldsymbol{\theta}|D)$; indeed, this is the intuitive basis of graphical representations of univariate distributions like those provided by boxplots.

Any region $R \subset \Theta$ such that $\int_R p(\boldsymbol{\theta}|D)\, \mathrm{d}\boldsymbol{\theta} = q$ (so that, given data D, the true value of $\boldsymbol{\theta}$ belongs to R with probability q) is said to be a posterior q-credible region of $\boldsymbol{\theta}$. Notice that this immediately provides a direct intuitive statement about the unknown quantity of interest $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$, $\phi(R)$ is a q-credible region of $\phi = \phi(\boldsymbol{\theta})$. 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 (continued). (Inference on normal parameters)
In the numerical example about the value of the gravitational field described in Figure 2(a), 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 2(b) is also a 95%-credible region for g, but it is not a HPD.

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.[39,40]

The concept of a credible region for a function $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ of the parameter vector is trivially extended to prediction problems. Thus, a posterior q-credible region for $\boldsymbol{x} \in \chi$ is a subset R of the sample space χ with posterior predictive probability q, so that $\int_R p(\boldsymbol{x}|D)d\boldsymbol{x} = q$ (see e.g., Example 6).

1.08.5.2 Hypothesis Testing

The reference posterior distribution $p(\boldsymbol{\theta}|D)$ of the quantity of interest $\boldsymbol{\theta}$ conveys immediate intuitive information on those values of $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta} \in \Theta_0 \subset \Theta$ of the possible values of the quantity of interest (where Θ_0 may possibly consist of a single value $\boldsymbol{\theta}_0$) is suggested in the course of the investigation as deserving special consideration, either because restricting $\boldsymbol{\theta}$ to Θ_0 would greatly simplify the model, or because there are additional, context-specific arguments suggesting that $\boldsymbol{\theta} \in \Theta_0$. Intuitively, the hypothesis $H_0 \equiv \{\boldsymbol{\theta} \in \Theta_0\}$ 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 $H_0 \equiv \{\boldsymbol{\theta} \in \Theta_0\}$ 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, $L(a_i, \boldsymbol{\theta})$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value $\boldsymbol{\theta}$ 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, $\int_\Theta L(a_0, \boldsymbol{\theta})p(\boldsymbol{\theta}|D)\,\mathrm{d}\boldsymbol{\theta}$, is larger than the expected posterior loss of rejecting, $\int_\Theta L(a_1, \boldsymbol{\theta})p(\boldsymbol{\theta}|D)\,\mathrm{d}\boldsymbol{\theta}$, that is if (and only if)

$$\int_\Theta [L(a_0, \boldsymbol{\theta}) - L(a_1, \boldsymbol{\theta})]p(\boldsymbol{\theta}|D)\,\mathrm{d}\boldsymbol{\theta} = \int_\Theta \Delta L(\boldsymbol{\theta})p(\boldsymbol{\theta}|D)\,\mathrm{d}\boldsymbol{\theta} > 0 \tag{60}$$

Therefore, only the loss difference $\Delta L(\boldsymbol{\theta}) = L(a_0, \boldsymbol{\theta}) - L(a_1, \boldsymbol{\theta})$, which measures the advantage of rejecting H_0 as a function of $\boldsymbol{\theta}$, 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, that is as if $\boldsymbol{\theta} \in \Theta_0$, but there are at least two options for the alternative action a_1. This may either mean (1) the negation of H_0, that is to act as if $\boldsymbol{\theta} \notin \Theta_0$ or, alternatively, it may rather mean (2) to reject the simplification implied by H_0 and to keep the unrestricted model, $\boldsymbol{\theta} \in \Theta$, 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 $H_0 \equiv \{\boldsymbol{\theta} \in \Theta_0\}$, is often embedded into a more general model, $\{\boldsymbol{\theta} \in \Theta, \Theta_0 \subset \Theta\}$, constructed to include possibly promising departures from H_0, and it is required to verify whether presently available data D are still compatible with $\boldsymbol{\theta} \in \Theta_0$, or whether the extension to $\boldsymbol{\theta} \in \Theta$ is really required.

Example 15. (Conventional hypothesis testing)
Let $p(\boldsymbol{\theta}|D), \boldsymbol{\theta} \in \Theta$, be the posterior distribution of the quantity of interest, a_0 be the decision to work under the restriction $\boldsymbol{\theta} \in \Theta_0$, and a_1 be the decision to work under the complementary restriction $\boldsymbol{\theta} \notin \Theta_0$. Suppose, moreover, that the loss structure has the simple, zero-one form given by $\{L(a_0, \boldsymbol{\theta}) = 0, L(a_1, \boldsymbol{\theta}) = 1\}$ if $\boldsymbol{\theta} \in \Theta_0$

and, similarly, $\{L(a_0,\boldsymbol{\theta})=1, L(a_1,\boldsymbol{\theta})=0\}$ if $\boldsymbol{\theta}\notin\Theta_0$, so that the advantage $\Delta L(\boldsymbol{\theta})$ of rejecting H_0 is 1 if $\boldsymbol{\theta}\notin\Theta_0$ 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) $\Pr(\boldsymbol{\theta}\notin\Theta_0|D) > \Pr(\boldsymbol{\theta}\in\Theta_0|D)$. Notice that this formulation requires that $\Pr(\boldsymbol{\theta}\in\Theta_0)>0$, that is the hypothesis H_0 has a strictly positive prior probability. If $\boldsymbol{\theta}$ is a continuous parameter and Θ_0 has zero measure (e.g., if H_0 consists of a single point $\boldsymbol{\theta}_0$), this requires the use of a nonregular 'sharp' prior concentrating a positive probability mass on Θ_0. For details see Kass and Raftery[41] and references therein.

Example 16. *(Intrinsic hypothesis testing)*
Again, let $p(\boldsymbol{\theta}|D), \boldsymbol{\theta}\in\Theta$, be the posterior distribution of the quantity of interest, a_0 be the decision to work under the restriction $\boldsymbol{\theta}\in\Theta_0$, and a_1 be the decision to keep the general, unrestricted model $\boldsymbol{\theta}\in\Theta$. In this case, the advantage $\Delta L(\boldsymbol{\theta})$ of rejecting H_0 as a function of $\boldsymbol{\theta}$ may safely be assumed to have the form $\Delta L(\boldsymbol{\theta}) = \delta(\Theta_0,\boldsymbol{\theta}) - \delta^*$, for some $\delta^*>0$, where (1) $\delta(\Theta_0,\boldsymbol{\theta})$ is some measure of the discrepancy between the assumed model $p(D|\boldsymbol{\theta})$ and its closest approximation within the class $\{p(D|\boldsymbol{\theta}_0), \boldsymbol{\theta}_0\in\Theta_0\}$, such that $\delta(\Theta_0,\boldsymbol{\theta})=0$ whenever $\boldsymbol{\theta}\in\Theta_0$, and (2) 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 $\delta(\Theta_0,\boldsymbol{\theta})$ 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 $\delta(\Theta_0,\boldsymbol{\theta})$ is an appropriate extension of the intrinsic discrepancy; when there are no nuisance parameters, this is given by

$$\delta(\Theta_0,\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}_0\in\Theta_0} \min\{\kappa(\boldsymbol{\theta}_0|\boldsymbol{\theta}), \kappa(\boldsymbol{\theta}|\boldsymbol{\theta}_0)\} \tag{61}$$

where $\kappa(\boldsymbol{\theta}_0|\boldsymbol{\theta}) = \int_T p(\boldsymbol{t}|\boldsymbol{\theta}) \log\{p(\boldsymbol{t}|\boldsymbol{\theta})/p(\boldsymbol{t}|\boldsymbol{\theta}_0)\}\,\mathrm{d}\boldsymbol{t}$, and $\boldsymbol{t}=\boldsymbol{t}(D)\in T$ is any sufficient statistic, which may well be the whole dataset D. As before, if the data $D=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ consist of a random sample from $p(\boldsymbol{x}|\boldsymbol{\theta})$, then

$$\kappa(\boldsymbol{\theta}_0|\boldsymbol{\theta}) = n\int_X p(\boldsymbol{x}|\boldsymbol{\theta}) \log\frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0)}\,\mathrm{d}\boldsymbol{x} \tag{62}$$

Naturally, the loss function $\delta(\Theta_0,\boldsymbol{\theta})$ reduces to the intrinsic discrepancy $\delta(\boldsymbol{\theta}_0,\boldsymbol{\theta})$ of Example 14 when Θ_0 contains a single element $\boldsymbol{\theta}_0$. Besides, as in the case of estimation, the definition is easily extended to problems with nuisance parameters, with

$$\delta(\Theta_0,\boldsymbol{\theta},\boldsymbol{\lambda}) = \inf_{\boldsymbol{\theta}_0\in\Theta_0,\,\boldsymbol{\lambda}_0\in\Lambda} \delta(\boldsymbol{\theta}_0,\boldsymbol{\lambda}_o,\boldsymbol{\theta},\boldsymbol{\lambda}) \tag{63}$$

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

$$\mathrm{d}(\Theta_0|D) = \int_\Lambda\int_\Theta \delta(\Theta_0,\boldsymbol{\theta},\boldsymbol{\lambda})\pi(\boldsymbol{\theta},\boldsymbol{\lambda}|D)\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\lambda} > \delta^* \tag{64}$$

for some $\delta^*>0$. As an expectation of a nonnegative quantity, $\mathrm{d}(\Theta_0, D)$ is obviously nonnegative. Moreover, if $\phi=\phi(\boldsymbol{\theta})$ is a one-to-one transformation of $\boldsymbol{\theta}$, then $\mathrm{d}(\phi(\Theta_0), D) = \mathrm{d}(\Theta_0, D)$ 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 $\mathrm{d}(\Theta_0, D)$ under sampling tends to one when H_0 is true, and tends to infinity otherwise; thus $\mathrm{d}(\Theta_0, D)$ 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, $\mathrm{d}(\Theta_0, D)$ is a test statistic for H_0 and the hypothesis should be rejected if the value of $\mathrm{d}(\Theta_0, D)$ 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 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 $\boldsymbol{\theta}$ is continuous and Θ_0 contains a single point $\boldsymbol{\theta}_0$).

Naturally, to implement the test, the utility constant δ^* which defines the rejection region must be chosen. Values of $\mathrm{d}(\Theta_0, D)$ 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. It 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[42] and Bernardo and Pérez.[43]

Example 17. *(Testing the value of a normal mean)*
Let the data $D = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $N(x \mid \mu, \sigma)$, where σ is assumed to be known, and consider the problem of testing whether these data are compatible or not 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 nonregular 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,

$$\Pr[\mu_0|D, \lambda] = \frac{B_{01}(D, \lambda)p_0}{(1-p_0) + p_0 B_{01}(D, \lambda)} \tag{65}$$

$$B_{01}(D, \lambda) = \left(1 + \frac{n}{\lambda}\right)^{1/2} \exp\left[-\frac{1}{2}\frac{n}{n+\lambda}z^2\right] \tag{66}$$

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 favor 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, see Barlett,[44] 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 behavior of posterior probabilities based on sharp, nonregular priors generally known as Lindley's paradox[45] 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 predicted by the general theory, the expected value of $d(\mu_0, D)$ under repeated sampling is 1 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

$$\delta(\mu_0, \mu, \sigma) = \frac{n}{2}\log\left[1 + \left(\frac{\mu - \mu_0}{\sigma}\right)^2\right] \tag{67}$$

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

$$\mathrm{d}(\mu_0, D) \approx \frac{1}{2} + \frac{n}{2}\log\left(1 + \frac{t^2}{n}\right) \tag{68}$$

where t is the conventional 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.

1.08.6 Discussion

This chapter 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 may use the list of references for further information. This final section briefly reviews the main arguments for a Bayesian approach.

1.08.6.1 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 axiomatic foundations prove the need of such a distribution for rational, coherent analysis. 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.

1.08.6.2 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 functions, which (1) are mathematically derived from the accepted probability model (and, hence, they are 'objective' insofar as the choice of that model might be objective) and, (2) by construction, they produce posterior probability distributions which, given the accepted probability model, only contain the information provided by the data and, optionally, and further contextual information over which there might be universal agreement.

1.08.6.3 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 behavior under repeated sampling provides a suggestive form of calibration. Indeed, $\Pr[\boldsymbol{\theta} \in R \mid D] = \int_R \pi(\boldsymbol{\theta}|D)\, \mathrm{d}\boldsymbol{\theta}$, the reference posterior probability that $\boldsymbol{\theta} \in R$, 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 $\boldsymbol{\theta}$ belongs to $R \subset \Theta$, and the limiting proportion of the regions which would cover $\boldsymbol{\theta}$ under repeated sampling conditional on data 'sufficiently similar' to D. Under broad conditions (to guarantee regular asymptotic behavior), all large data sets from the same model are sufficiently similar among themselves in this sense, and hence, reference posterior credible regions are typically approximate frequentist confidence regions.

The conditions for this approximate equivalence to hold exclude, however, important special cases, like those involving 'extreme' or 'relevant' observations, where conventional statistical methods often produce untenable answers. On the other hand, in very special situations (essentially when probability models may be transformed to location-scale models or when there exist pivotal quantities), there is an exact equivalence; in those cases reference posterior credible intervals are, for any sample size, exact frequentist confidence intervals.

1.08.6.4 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 approximations, 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 techniques 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. For an introduction to MCMC methods in Bayesian inference, see Gilks *et al.*,[46] Mira,[47] and references therein.

Acknowledgments

This work has been partially funded with grant MTM 2006-07801 of the MEC, Spain.

References

1. Lindley, D. V. *Bayesian Statistics, A Review*; SIAM: Philadelphia, PA, 1972.
2. Jaynes, E. T. Confidence Intervals vs. Bayesian Intervals. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*; Harper, W. L., Hooker, C. A., Eds.; Reidel: Dordrecht, 1976; Vol. 2, pp 175–257 (with discussion).
3. Laplace, P. S. *Théorie Analytique des Probabilités*. Courcier: Paris. Reprinted as *Oeuvres Complètes de Laplace*; Gauthier-Villars: Paris, 1812; Vol. 7 (1878–1912).

4. Jeffreys, H. *Theory of Probability*; Oxford University Press: Oxford, 1939 (3rd ed. in 1961).
5. de Finetti, B. *Teoria delle Probabilità*, Turin: Einaudi. English translation as *Theory of Probability*; Wiley: Chichester, 1970 (1975)
6. Lindley, D. V. *Introduction to Probability and Statistics from a Bayesian Viewpoint*; Cambridge University Press: Cambridge, 1965.
7. Zellner, A. *An Introduction to Bayesian Inference in Econometrics*; Wiley: New York, 1971. Reprinted in Krieger: Melbourne, FL, 1987.
8. Box, G. E. P.; Tiao, G. C. *Bayesian Inference in Statistical Analysis*; Addison-Wesley: Reading, MA, 1973.
9. Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*; Springer: Berlin, 1985.
10. Bernardo, J. M.; Smith, A. F. M. *Bayesian Theory*; Wiley: Chichester, 1994; Chapters 2, 5 (2nd ed. forthcoming).
11. Rubin, D. B. The Broad Role of Multiple Imputation in Statistical Science. *Compstat14*; Bethlehem, J. G., van der Heijden, P. G., Eds.; Physica-Verlag: Heidelberg, 2000; pp 3–14.
12. Dryden, I. L.; Hirst, J. D.; Melville, J. L. Statistical Analysis of Nnlabeled Pointsets: Comparing Molecules in Chemoinformatics. *Biometrics* **2007**, *63*, 237–251.
13. Chen, H.; Bakshia, B. R.; Goel, P. K. Toward Bayesian Chemometrics—A Tutorial on Some Recent Advances. *Anal. Chim. Acta* **2007**, *602*, 1–16.
14. Ramsey, F. P. Truth and Probability. *The Foundations of Mathematics and Other Logical Essays*; Braithwaite, R. B., Ed.; Kegan Paul: London, 1926, pp. 156–198 (1931). Reprinted in *Studies in Subjective Probability*; Kyburg H. E., Smokler, H. E., Eds.; Dover: New York, 1980; pp 61–92.
15. Savage, L. J. *The Foundations of Statistics*; Dover: New York, 1954 (2nd ed. in 1972).
16. DeGroot, M. H. *Optimal Statistical Decisions*; McGraw-Hill: New York, 1970; Chapter 6.
17. Bernardo, J. M. Expected Information as Expected Utility. *Ann. Stat.* **1979**, *7*, 686–690.
18. Liseo, B. The Elimination of Nuisance Parameters. *Handbook of Statistics*; Dey, D. K., Rao, C. R., Eds.; Elsevier: Amsterdam, 2005; Vol. 25, pp 193–219.
19. Geisser, S. *Predictive Inference: An Introduction*; Chapman and Hall: London, 1993.
20. Kass, R. E.; Wasserman, L. The Selection of Prior Distributions by Formal Rules. *J. Am. Stat. Assoc.* **1996**, *91*, 1343–1370.
21. Bernardo, J. M.; Ramón, J. M. An Introduction to Bayesian Reference Analysis: Inference on the Ratio of Multinomial Parameters. *Statistician* **1998**, *47*, 1–35.
22. Bernardo, J. M. Reference Analysis. In *Handbook of Statistics*; Dey, D. K., Rao, C. R., Eds.; Elsevier: Amsterdam, 2005; Vol. 25, pp 17–90.
23. Bernardo, J. M. Reference Posterior Distributions for Bayesian Inference. *J. R. Stat. Soc. B* **1979**, *41*, 113–147; (with discussion). Reprinted in *Bayesian Inference* **1**; Tiao, G. C., Polson, N. G., Eds.; Edward Elgar: Oxford, 1995; pp 229–263.
24. Bernardo, J. M. Reference Decisions. *Symp. Mathematica* **1981**, *25*, 85–94.
25. Berger, J. O.; Bernardo, J. M. Estimating a product of Means: Bayesian Analysis with Reference Priors. *J. Am. Stat. Assoc.* **1989**, *84*, 200–207.
26. Berger, J. O.; Bernardo, J. M. Ordered Group Reference Priors with Applications to a Multinomial Problem. *Biometrika* **1992**, *79*, 25–37.
27. Berger, J. O.; Bernardo, J. M. Reference Priors in a Variance Components Problem. In *Bayesian Analysis in Statistics and Econometrics*; Goel, P. K., Iyengar, N. S., Eds.; Springer: Berlin, 1992; pp 323–340.
28. Berger, J. O.; Bernardo, J. M. On the Development of Reference Priors. In *Bayesian Statistics 4*; Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M., Eds.; University PressOxford, 1992; pp 35–60, (with discussion).
29. Berger, J. O.; Bernardo, J. M.; Sun, D. The Formal Definition of Reference Priors. *Ann. Stat*. (to appear).
30. Bernardo, J. M. Noninformative Priors Do Not Exist. *J. Stat. Plan. Inference* **1997**, *65*, 159–189, (with discussion).
31. Dawid, A. P.; Stone, M.; Zidek, J. V. Marginalization Paradoxes in Bayesian and Structural Inference. *J. R. Stat. Soc. B* **1973**, *35*, 189–233, (with discussion).
32. Stein, C. An Example of Wide Discrepancy Between Fiducial and Confidence Intervals. *Ann. Math. Stat.* **1959**, *30*, 877–880.
33. Wald, A. *Statistical Decision Functions*; Wiley: New York, 1950.
34. Datta, G. S.; Sweeting, T. J. Probability Matching Priors. In *Handbook of Statistics*; Dey, D. K., Rao, C. R., Eds.; Elsevier: Amsterdam, 2005; pp 91–114.
35. Efron, B. Why Isn't Everyone a Bayesian? *Am. Stat.* **1986**, *40*, 1–11, (with discussion).
36. Lindley, D. V. Fiducial Distribution and Bayes' Theorem. *J. R. Stat. Soc. B* **1958**, *20*, 102–107.
37. Bernardo, J. M.; Juárez, M. A. Intrinsic Estimation. In *Bayesian Statistics 7*; Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M., Eds.; University Press: Oxford, 2003; pp 465–476.
38. Bernardo, J. M. Intrinsic Point Estimation of the Normal Variance. In *Bayesian Statistics and Its Applications*; Upadhyay, S. K., Singh, U., Dey, D. K., Eds.; Anamaya Publications: New Delhi, 2006; pp 110–121.
39. Bernardo, J. M. Intrinsic Credible Regions: An Objective Bayesian Approach to Interval Estimation. *Test* **2005**, *14*, 317–384, (with discussion).
40. Bernardo, J. M. Objective Bayesian Point and Region Estimation in Location-Scale Models. *Sort* **2007**, *14*, 3–44.
41. Kass, R. E.; Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795.
42. Bernardo, J. M.; Rueda, R. Bayesian Hypothesis Testing: A Reference Approach. *Inter. Stat. Rev.* **2002**, *70*, 351–372.
43. Bernardo, J. M.; Pérez, S. Comparing Normal Means: New Methods for An Old Problem. *Bayesian Anal.* **2007**, *2*, 45–58.
44. Bartlett, M. A Comment on D. V. Lindley's Statistical Paradox. *Biometrika* **1957**, *44*, 533–534.
45. Lindley, D. V. A Statistical Paradox. *Biometrika* **1957**, *44*, 187–192.
46. Gilks, W. R., Richardson, S. Y., Spiegelhalter, D. J., Eds.; *Markov Chain Monte Carlo in Practice*; Chapman and Hall: London, 1996.
47. Mira, A. MCMC Methods to Estimate Bayesian Parametric Models. In *Handbook of Statistics*; Dey, D. K., Rao, C. R., Eds.; Elsevier: Amsterdam, 2005; Vol. 25, pp 415–436.

Biographical Sketch

José M. Bernardo is currently Professor of Statistics at the University of Valencia, Spain. He received Ph.D. in Mathematics from the University of Valencia in 1974 and Ph.D. in *Statistics from the* University College London in 1976. He worked as a lecturer (1972–77) and as Professor of Biostatistics (1978–82) in the University of Valencia prior to becoming Professor of Statistics. His research interests include the following: the nature and scope of Bayesian Statistics; the relationship between information theory and statistics; the foundations and the derivation of nonsubjective, *reference* priors; the foundations of decision theory; Bayesian hypothesis testing and model criticism; point and region intrinsic estimation; probabilistic classification; applications of Bayesian Statistics and decision theory to medical problems (automatic diagnosis), political analysis (election forecasting, sample surveys), and industrial problems (quality assurance). He has won many awards and fellowships (Spanish National Graduation Award (Mathematics); Spanish National Doctorate Award (Mathematics); British Council Research Fellowship; Yale University Postdoctoral Fellowship; Fellow of the *American Statistical Association*; Académico Correspondiente de la *Real Academia de Ciencias de Madrid*; Fellow of the *Royal Statistical Society*; Fellow of the *Institute of Statisticians*; elected member of the *International Statistical Institute*; Founder co-President of the *International Society for Bayesian Analysis 1992–94*; listed in *Who is Who in the World (1984–)*). He has been Advisor on Statistics and Decision Theory to the President of the State of Valencia (1989–95); General Director of Decision Analysis, Government of the State of Valencia (1991–93); Founding Editor of *Test* (1992–97); Associate Editor of *Estadística Española* (1986–), the *Journal of the Royal Statistical Society* (Series B) (1989–93), the *Journal of the Iranian Statistical Society* (2002–), *The Statistician* (1987–97), and *Questio* (1983–2002); Contributing Editor of the *Current Index of Statistics* (1984–) and *Statistics Theory and Methods Abstracts* (1996–2003); consultant for the *Spanish Scientific and Technological Commission* (Madrid), the *National Science Foundation* (USA), the *National Security Agency* (USA), the *Research Council of Canada*, the *Fondo Nacional de Investigaciòn Cientìfica* (Chile), the *Foundation for Research and Development* (South Africa), the *Czech Academy of Sciences*, and the *Academy of Finland*. He has supervised 17 completed Ph.D. theses. He has conducted courses all over the world.