Table of Contents

Related Articles

Bayesian econometrics
Bayesian methods in macroeconometrics
Bayesian nonparametrics
Bayesian time series analysis
Bayes, Thomas (1702–1761)
de Finetti, Bruno (1906–1985)
Savage, Leonard J. (Jimmie) (1917–1971)
statistical decision theory

Bayesian statistics

José M. Bernardo
From *The New Palgrave Dictionary of Economics*, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Back to top

Statistics is primarily concerned with analysing data, either to assist in appreciating some underlying mechanism or to reach effective decisions. All uncertainties should be described by probabilities, since probability is the only appropriate language for a logic that deals with all degrees of uncertainty, not just absolute truth and falsity. This is the essence of Bayesian statistics. Decision-making is embraced by introducing a utility function and then maximizing expected utility. Bayesian statistics is designed to handle all situations where uncertainty is found. Since some uncertainty is present in most aspects of life, Bayesian statistics arguably should be universally appreciated and used.

Keywords

Back to top

asymptotic behaviour; Bayes, T.; Bayesian reference criterion; Bayesian statistics; exchangeability; expected utility; hypothesis testing; improper prior function; inference; likelihood; nonparametric models; nuisance parameters; point estimation; prediction; probability; reference analysis; region estimation; representation theorems; robustness; statistical decision theory; statistical inference; subjective probability; sufficiency; sure thing principle; uncertainty

Article

Back to top

Bayesian statistics is a comprehensive approach to both statistical inference and decision analysis which derives from the fact that, for rational behaviour, all uncertainties in a problem must necessarily be described by probability distributions.

Unlike most other branches of mathematics, conventional methods of statistical inference do not have 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. 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 for statistical inference, a scientific revolution in Kuhn's sense. Bayesian statistics require only the mathematics of probability theory and the interpretation of probability which most closely corresponds to the standard use of this word in everyday language: a conditional measure of uncertainty. The main consequence of these axiomatic foundations is precisely the requirement to describe with probability distributions all uncertainties present in the problem. Hence, parameters are treated as random variables; this is not a description of their variability (parameters are typically fixed unknown quantities) but a description of the uncertainty about their true values.

The Bayesian paradigm is easily summarized. Thus, if available data D are assumed to have been generated from a probability distribution $p(D|\omega)$ characterized by an unknown parameter vector ω, the uncertainty about the value of ω before the data have been observed must be described by a prior probability distribution $p(\omega)$. After data D have been observed, the uncertainty about the value of ω is described by its posterior distribution $p(\omega|D)$, which is obtained via Bayes's theorem; hence the adjective Bayesian for this form of inference. Point and region estimates for ω may be derived from $p(\omega|D)$ as useful summaries of its contents. Measures of the compatibility of the posterior with a particular set Θ_0 of parameter values may be used to test the hypothesis $H_0 = \{q \in \Theta_0\}$. If data consist of a random sample $D = \{x_1, \ldots, x_n\}$ from a probability distribution $p(x|\omega)$, inferences about the value of a future observation x from the same process are derived from the (posterior) predictive distribution $p(x|D) = \int_\Omega p(x|\omega)\, p(\omega|D)\, d\omega$.

An 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 data. This is addressed by reference analysis which uses information-theoretic concepts to derive the appropriate reference posterior distribution $\pi(\omega|D)$, defined to encapsulate inferential conclusions about the value of ω solely based on the assumed probability model $p(D|\omega)$ and the observed data D.

Pioneering textbooks on Bayesian statistics were Jeffreys (1961), Lindley (1965), Zellner (1971) and Box and Tiao (1973). For modern elementary introductions, see Berry (1996) and Lee (2004). Intermediate to advanced monographs on Bayesian statistics include Berger (1985), Bernardo and Smith (1994), Gelman et al. (2003), O'Hagan (2004) and Robert (2001). This article may be regarded as a very short summary of the material contained in the forthcoming second edition of Bernardo and Smith (1994). For a recent review of objective Bayesian statistics, see Bernardo (2005) and references therein.

The New Palgrave Dictionary of Economics Online

Bayesian statistics

José M. Bernardo
From The New Palgrave Dictionary of Economics, Second Edition, 2008
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

Statistics is primarily concerned with analysing data, either to assist in appreciating some underlying mechanism or to reach effective decisions. All uncertainties should be described by probabilities, since probability is the only appropriate language for a logic that deals with all degrees of uncertainty, not just absolute truth and falsity. This is the essence of Bayesian statistics. Decision-making is embraced by introducing a utility function and then maximizing expected utility. Bayesian statistics is designed to handle all situations where uncertainty is found. Since some uncertainty is present in most aspects of life, Bayesian statistics arguably should be universally appreciated and used.

Keywords

asymptotic behaviour; Bayes, T.; Bayesian reference criterion; Bayesian statistics; exchangeability; expected utility; hypothesis testing; improper prior function; inference; likelihood; nonparametric models; nuisance parameters; point estimation; prediction; probability; reference analysis; region estimation; representation theorems; robustness; statistical decision theory; statistical inference; subjective probability; sufficiency; sure thing principle; uncertainty

Article

Bayesian statistics is a comprehensive approach to both statistical inference and decision analysis which derives from the fact that, for rational behaviour, all uncertainties in a problem must necessarily be described by probability distributions.

Unlike most other branches of mathematics, conventional methods of statistical inference do not have 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. 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 for statistical inference, a scientific revolution in Kuhn's sense. Bayesian statistics require only the mathematics of probability theory and the interpretation of probability which most closely corresponds to the standard use of this word in everyday language: a conditional measure of uncertainty. The main consequence of these axiomatic foundations is precisely the requirement to describe with probability distributions all uncertainties present in the problem. Hence, parameters are treated as random variables; this is not a description of their variability (parameters are typically fixed unknown quantities) but a description of the uncertainty about their true values.

The Bayesian paradigm is easily summarized. Thus, if available data D are assumed to have been generated from a probability distribution $p(D|\boldsymbol{\omega})$ characterized by an unknown parameter vector $\boldsymbol{\omega}$, the uncertainty about the value of $\boldsymbol{\omega}$ before the data have been observed must be described by a prior probability distribution $p(\boldsymbol{\omega})$. After data D have been observed, the uncertainty about the value of $\boldsymbol{\omega}$ is described by its posterior distribution $p(\boldsymbol{\omega}|D)$, which is obtained via Bayes's theorem; hence the adjective Bayesian for this form of inference. Point and region estimates for $\boldsymbol{\omega}$ may be derived from $p(\boldsymbol{\omega}|D)$ as useful summaries of its contents. Measures of the compatibility of the posterior with a particular set $\boldsymbol{\Theta}_0$ of parameter values may be used to test the hypothesis $H_0 \equiv \{q \in \boldsymbol{\Theta}_0\}$. If data consist of a random sample $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from a probability distribution $p(\boldsymbol{x}|\boldsymbol{\omega})$, inferences about the value of a future observation $\boldsymbol{x}$ from the same process are derived from the (posterior) predictive distribution $p(\boldsymbol{x}|D) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x}|\boldsymbol{\omega})\ p(\boldsymbol{\omega}|D)\ d\boldsymbol{\omega}$.

An 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 data. This is addressed by reference analysis which uses information-theoretic concepts to derive the appropriate reference posterior distribution $\pi(\boldsymbol{\omega}|D)$, defined to encapsulate inferential conclusions about the value of $\boldsymbol{\omega}$ solely based on the assumed probability model $p(D|\boldsymbol{\omega})$ and the observed data D.

Pioneering textbooks on Bayesian statistics were Jeffreys (1961), Lindley (1965), Zellner (1971) and Box and Tiao (1973). For modern elementary introductions, see Berry (1996) and Lee (2004). Intermediate to advanced monographs on Bayesian statistics include Berger (1985), Bernardo and Smith (1994), Gelman et al. (2003), O'Hagan (2004) and Robert (2001). This article may be regarded as a very short summary of the material contained in the forthcoming second edition of Bernardo and Smith (1994). For a recent review of objective Bayesian statistics, see Bernardo (2005) and references therein.

Foundations

The central element of the Bayesian paradigm is the use of probabilities to describe all relevant uncertainties, interpreting $\Pr(A|H)$, the probability of A given H, as a conditional measure of uncertainty, on a [0,1] scale, about the occurrence of the event A in conditions H. There are two different independent arguments which prove the mathematical inevitability of the use of probabilities to describe uncertainties.

Exchangeability and representation theorems

Available data often consist of a finite set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of 'homogeneous' observations, in the sense that only their values matter, not the order in which they appear. Formally, this is captured by the notion of exchangeability. The set of random vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_j \in \mathrm{X}$, is exchangeable if their joint distribution is invariant under permutations. An infinite sequence of random vectors is exchangeable if all its finite subsequences are exchangeable. Notice that, in particular, any random sample from any model is exchangeable. The general representation theorem implies that, if a set of observations is assumed to be a subset of an exchangeable sequence, then it constitutes a random sample from a probability model $\{p(\boldsymbol{x}|\boldsymbol{\omega}),\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, described in terms of some parameter vector $\boldsymbol{\omega}$; furthermore, this parameter $\boldsymbol{\omega}$ is defined as the limit (as $n \to \infty$) of some function of the observations, and available information about the value of $\boldsymbol{\omega}$ must necessarily be described by some probability distribution $p(\boldsymbol{\omega})$. This formulation includes 'nonparametric' (distribution free) modelling, where $\boldsymbol{\omega}$ may index, for instance, all continuous probability distributions on X. Notice that $p(\boldsymbol{\omega})$ does not model a possible variability of $\boldsymbol{\omega}$ (since $\boldsymbol{\omega}$ will typically be a fixed unknown vector), but models the uncertainty associated with its actual value. Under exchangeability (and therefore under any assumption of random sampling), the general representation theorem provides an existence theorem for a probability distribution $p(\boldsymbol{\omega})$ on the parameter space $\boldsymbol{\Omega}$, and this is an argument which depends only on mathematical probability theory.

Statistical inference and decision theory

Statistical decision theory provides a precise methodology to deal with decision problems under uncertainty, but it also provides a powerful axiomatic basis for the Bayesian approach to statistical inference. A decision problem exists whenever there are two or more possible courses of action. Let A be the class of possible actions, let $\boldsymbol{\Theta}$ be the set of relevant events which may affect the result of choosing an action, and let $c(a, q) \in$ C, be the consequence of having chosen action a when event $\boldsymbol{\theta}$ takes place. The triplet {A, $\boldsymbol{\Theta}$, C} describes the structure of the decision problem. Different sets of principles have been proposed to capture a minimum collection of logical rules that could sensibly be required for rational decision-making. These all consist of axioms with a strong intuitive appeal; examples include the *transitivity* of preferences (if $a_1 \succ a_2$ and $a_2 \succ a_3$, then $a_1 \succ a_3$), and the *sure thing principle* (if $a_1 \succ a_2$ given E, and $a_1 \succ a_2$ given $\bar{E}$, then $a_1 \succ a_2$). 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, but they all lead to the same basic conclusions:

- Preferences among possible consequences should be measured with a *utility* function $u(c) = u(a, q)$ which specifies, on some numerical scale, their desirability.
- The uncertainty about the relevant events should be measured with a probability distribution $p(q|D)$ describing their plausibility given the conditions under which the decision must be taken (assumptions made and available data D).
- The best strategy is to take that action a^* with maximizes the corresponding expected utility, $\int_{\boldsymbol{\Theta}} u(a, q)\ p(q|D)dq$.

Notice that the argument described above establishes (from another perspective) the need to quantify the uncertainty about all relevant unknown quantities (the actual value of the vector $\boldsymbol{\theta}$), and specifies that this must have the mathematical structure of a probability distribution. It has been argued that the development described above (which is not qsted when decisions have to be made) does not apply to problems of statistical inference, where no specific decision making is envisaged. Notice, however, that (*a*) a problem of statistical inference is typically considered worth analysing because it may eventually help make sensible decisions (as Ramsey put it in the 1930s, a lump of arsenic is poisonous because it may kill someone, not because it has actually killed someone), and (*b*) statistical inference on $\boldsymbol{\theta}$ has the mathematical structure of a decision problem, where the class of alternatives is the functional space of all possible conditional probability distributions 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.
In statistical inference it is often convenient to work in terms of the non-negative loss function $\boldsymbol{\ell}(a, q) = \sup_{a \in \mathrm{A}} \{u(a, q)\} - u(a, q)$, which directly measures, as a function of $\boldsymbol{\theta}$, the *penalty* for choosing a wrong action. The undesirability of each possible action $a \in$ A is then measured by its *expected loss*, $l(a|D) = \int_{\boldsymbol{\Theta}} \boldsymbol{\ell}(a, q)\ p(q|D)\ dq$, and the best action a^* is that with the minimum expected loss.

The Bayesian paradigm

The statistical analysis of some observed data set $D \in$ D typically begins with some informal descriptive evaluation, which is used to suggest a tentative, formal probability model $\{p(D|\boldsymbol{\omega}, H),\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ which, given some assumptions H, is supposed to represent, for some (unknown) value of $\boldsymbol{\omega}$, the probabilistic mechanism which has generated the observed data D. The arguments outlined above establish the logical need to assess a prior probability distribution $p(\boldsymbol{\omega}|H)$ over the parameter space $\boldsymbol{\Omega}$, describing the available knowledge about the value of $\boldsymbol{\omega}$ under the accepted assumptions H, prior to the data being observed. It then follows from Bayes's theorem 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*,

$$p(\boldsymbol{\omega}|D, H) = \frac{p(D|\boldsymbol{\omega}, H)\, p(\boldsymbol{\omega}|H)}{\int_{\boldsymbol{\Omega}} p(D|\boldsymbol{\omega}, H)\, p(\boldsymbol{\omega}|H) d\boldsymbol{\omega}},\ \boldsymbol{\omega} \in \boldsymbol{\Omega} \tag{1}$$

It is this systematic use of Bayes's 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's 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. To simplify the presentation, the assumptions H are often omitted from the notation, but the fact that all statements about $\boldsymbol{\omega}$ given D are also conditional to H should always be kept in mind.
Computation of posterior densities is often facilitated by noting that Bayes's theorem may be simply expressed as $p(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})\, p(\boldsymbol{\omega})$ (where $\propto$ stands for 'proportional to' and where, for simplicity, the assumptions H have been omitted from the notation), since the missing proportionality constant $[\int_{\boldsymbol{\Omega}} p(D|\boldsymbol{\omega})\, p(\boldsymbol{\omega})\, d\boldsymbol{\omega}]^{-1}$ may always be deduced from the fact that $p(\boldsymbol{\omega}|D)$, a probability density, must integrate to 1.

Improper priors

An improper prior function is defined as non-negative function $\pi(\boldsymbol{\omega})$ such that $\int_{\boldsymbol{\Omega}} \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ is not finite. The formal expression of Bayes's theorem remains, however technically valid if $p(\boldsymbol{\omega})$ is replaced by an improper prior function $\pi(\boldsymbol{\omega})$, provided the proportionality constant exists, thus leading to a well-defined *proper* posterior density $\pi(\boldsymbol{\omega}|D) \propto p(D|\boldsymbol{\omega})\pi(\boldsymbol{\omega})$, which does integrate to 1.

Likelihood principle

Considered as a function of $\boldsymbol{\omega}$ for fixed data D, $p(D|\boldsymbol{\omega})$ is often referred to as the likelihood function. Thus, Bayes's theorem is simply expressed in words by the statement that the posterior is proportional to the likelihood times the prior. It follows from (1) that, provided the same prior $p(\boldsymbol{\omega})$ is used, two different data sets D_1 and D_2, with possibly different probability models $p_1(D_1|\boldsymbol{\omega})$ and $p_2(D_2|\boldsymbol{\omega})$ which yield proportional likelihood functions, will produce identical posterior distributions for $\boldsymbol{\omega}$. This immediate consequence of Bayes's 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 $p(\boldsymbol{\omega})$, the posterior distribution does not depend on the set D of possible data values (the *outcome space*). Notice, however, that the likelihood principle applies only to inferences about the parameter vector $\boldsymbol{\omega}$ once the data have been obtained. Consideration of the outcome 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.

Sequential learning

Naturally, the terms 'prior' and 'posterior' are only relative to a particular set of data. As one would expect, if exchangeable 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, $p(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1}) \propto p(\boldsymbol{x}_{i+1}|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$, for $i = 1, \ldots, n-1$, so that the 'posterior' at a given stage becomes the 'prior' at the next.

Sufficiency

For a given probability model, one may find that some particular function of the data $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathrm{T}$ 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, $\boldsymbol{t}$ is sufficient if (and only if) there exist non-negative functions f and g such that the likelihood function may be factorized in the 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, then the posterior distribution of $\boldsymbol{\omega}$ depends only on the data D through $\boldsymbol{t}(D)$, and $p(\boldsymbol{\omega}|D) = p(\boldsymbol{\omega}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\omega})\ p(\boldsymbol{\omega})$.

Robustness

As one would expect, for fixed data and model assumptions, different priors generally 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 analyse the robustness of the posterior to changes in the prior. Objective posterior distributions based on reference priors (see below) play a central role in this 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.

Nuisance parameters

Typically, the quantity of interest is not the whole parameter vector $\boldsymbol{\omega}$, but some function $q = q(\boldsymbol{\omega})$ of possibly lower dimension than $\boldsymbol{\omega}$. Any valid conclusion on the value of $\boldsymbol{\theta}$ will be contained in its posterior probability distribution $p(q|D)$, which may be derived from $p(\boldsymbol{\omega}|D)$ by standard use of probability calculus. Indeed, if $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\omega}) \in \Lambda$ is some other function of $\boldsymbol{\omega}$ such that $\boldsymbol{\psi} = \{\boldsymbol{\theta}, \boldsymbol{\lambda}\}$ is a one-to-one transformation of $\boldsymbol{\omega}$, and $J(\boldsymbol{\omega}) = (\partial\boldsymbol{\psi}/\partial\boldsymbol{\omega})$ is the corresponding Jacobian matrix, one may change variables to obtain $p(\boldsymbol{\psi}|D) = p(q, \boldsymbol{\lambda}|D) = p(\boldsymbol{\omega}|D)/|J(\boldsymbol{\omega})|$, and the required posterior of $\boldsymbol{\theta}$ is $p(q|D) = \int_{\boldsymbol{\Lambda}} p(q, \boldsymbol{\lambda}|D)\, d\boldsymbol{\lambda}$, the marginal density obtained by integrating out the nuisance parameter $\boldsymbol{\lambda}$. Naturally, introduction of $\boldsymbol{\lambda}$ is not necessary if $\boldsymbol{\theta}(\boldsymbol{\omega})$ is a one-to-one transformation of $\boldsymbol{\omega}$. Notice that elimination of unwanted nuisance parameters, a simple integration within the Bayesian paradigm, is a difficult (often polemic) problem for conventional statistics.

Restricted parameter space

Sometimes, the range of possible values of $\boldsymbol{\omega}$ is effectively restricted by contextual considerations. If $\boldsymbol{\omega}$ is known to belong to $\boldsymbol{\Omega}_c \subset \boldsymbol{\Omega}$, the prior distribution is positive only in $\boldsymbol{\Omega}_c$ and, if one uses Bayes's theorem, it is immediately found that the restricted posterior is

$$p(\boldsymbol{\omega}|D,\ \boldsymbol{\omega} \in \boldsymbol{\Omega}_c) = p(\boldsymbol{\omega}|D) \Big/ \int_{\boldsymbol{\Omega}_c} p(\boldsymbol{\omega}|D) d\boldsymbol{\omega},$$

for $\boldsymbol{\omega} \in \boldsymbol{\Omega}_c$ (and obviously vanishes if $\boldsymbol{\omega} \notin \boldsymbol{\Omega}_c$). Thus, to incorporate a restriction on the possible values of the parameters, it suffices to renormalize the unrestricted posterior distribution to the set $\boldsymbol{\Omega}_c \subset \boldsymbol{\Omega}$ of parameter values which 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.

Asymptotic behaviour

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

$$F_{ij}(\boldsymbol{\omega}) = -E_{\boldsymbol{x}|\boldsymbol{\omega}}[-\partial^2 \log[p(\boldsymbol{x}|\boldsymbol{\omega})]/(\partial\boldsymbol{\omega}_i \partial\boldsymbol{\omega}_j)].$$

Prediction

When data consist of a set $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of homogeneous observations, one is often interested in predicting the value of a future observation $\boldsymbol{x}$ generated by the same random mechanism that has generated the observations in D. It follows from the foundations arguments discussed above that the solution to this prediction problem must be a probability distribution $p(\boldsymbol{x}|D)$ which describes the uncertainty about the value that $\boldsymbol{x}$ will take, given the information provided by D, and any other available knowledge. In particular, if contextual information suggests that data D may be considered to be a random sample from a distribution in the family $\{p(\boldsymbol{x}|\boldsymbol{\omega}),\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and $p(\boldsymbol{\omega})$ is a probability distribution which encapsulates all available prior information on the value of $\boldsymbol{\omega}$, the corresponding posterior will be (by Bayes's theorem) $p(\boldsymbol{\omega}|D) \propto \prod_{j=1}^{n} p(\boldsymbol{x}_j|\boldsymbol{\omega})\ p(\boldsymbol{\omega})$. Since $p(\boldsymbol{x}|\boldsymbol{\omega}, D) = p(\boldsymbol{x}|\boldsymbol{\omega})$, the total probability theorem may then be used to obtain the desired posterior *predictive* distribution

$$p(\boldsymbol{x}|D) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x}|\boldsymbol{\omega})\, p(\boldsymbol{\omega}|D) d\boldsymbol{\omega} \tag{2}$$

which has the form of a *weighted average*: the average of all possible probability distributions of $\boldsymbol{x}$, weighted with their corresponding posterior densities. Notice that the conventional practice of plugging in some point estimate $\tilde{\boldsymbol{\omega}} = \tilde{\boldsymbol{\omega}}(D)$ and using $p(\boldsymbol{x}|\tilde{\boldsymbol{\omega}})$ to predict $\boldsymbol{x}$ may be seriously misleading, for this totally ignores the uncertainty about the true value of $\boldsymbol{\omega}$. If the assumptions on the probability model are correct, the posterior predictive distribution $p(\boldsymbol{x}|D)$ will converge, as the sample size increases, to the distribution $p(\boldsymbol{x}|\boldsymbol{\omega})$ which has generated the data. Indeed, a good technique to assess the quality of the inferences about $\boldsymbol{\omega}$ encapsulated in $p(\boldsymbol{\omega}|D)$ is to check against the observed data the predictive distribution $p(\boldsymbol{x}|D)$ generated from $p(\boldsymbol{\omega}|D)$. The argument used to derive $p(\boldsymbol{x}|D)$ may be extended to obtain the predictive distribution of any function $\boldsymbol{y}$ of future observations generated by the same process, namely, $p(\boldsymbol{y}|D) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{y}|\boldsymbol{\omega})\ p(\boldsymbol{\omega}|D)$.

Reference analysis

The posterior distribution combines the information provided by the data with relevant available prior information. In many situations, however, either the available prior information 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 identify the mathematical form of a reference prior, a prior that would have a minimal effect, relative to the data, on the posterior inference. Much work has been done to formulate priors which would make this idea mathematically precise. This section summarizes an approach, based on information theory, which may be argued to provide the most advanced general procedure available. In this formulation, the reference prior is that which maximizes the missing information about the quantity of interest.

Reference distributions

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

$$I\left\{\mathrm{T},\ p(\theta)\right\} = \int_{\mathrm{T}}\int_{\Theta} p(\boldsymbol{t},\ \theta)\, \log\, \frac{p(\boldsymbol{t},\ \theta)}{p(\boldsymbol{t})\, p(\theta)}\, d\theta d\boldsymbol{t} = E_t\left[\int_{\Theta} p(\theta|\boldsymbol{t})\, \log\, \frac{p(\theta|\boldsymbol{t})}{p(\theta)}\, d\theta\right] \tag{3}$$

the expected logarithmic divergence of the prior from the posterior. This is a *functional* of the prior distribution $p(\theta)$: the larger the prior information, the smaller the information which the data may be expected to provide. The functional $I\{\mathrm{T},\ p(\theta)\}$ is concave, non-negative, and invariant under one-to-one transformations of θ. Consider now the amount of information $I\{\mathrm{T}^k,\ p(\theta)\}$ about θ which 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 $I\{\mathrm{T}^k,\ p(\theta)\}$ will approach the *missing information* about θ associated with the prior $p(\theta)$. Intuitively, the reference prior for θ is that which maximizes the missing information about θ. If $\pi_k(\theta|\mathrm{P})$ denotes the prior density which maximizes $I\{\mathrm{T}^k,\ p(\theta)\}$ in the class P of strictly positive prior distributions which are compatible with accepted assumptions on the value of θ (which may well be the class of all strictly positive proper priors), then the θ-reference prior $\pi(\theta|\mathrm{P})$ is the limit of the sequence of priors $\{\pi_k(\theta|\mathrm{P})\}_{k=1}^{\infty}$. The limit is taken in the precise sense that, for any value of the sufficient statistic $\boldsymbol{t}$, the reference posterior, the pointwise limit $\pi(\theta|\boldsymbol{t},\ \mathrm{P})$ of the corresponding sequence of posteriors $\{\pi_k(\theta|\boldsymbol{t},\ \mathrm{P})\}_{k=1}^{\infty}$, where $\pi_k(\theta|\boldsymbol{t},\ \mathrm{P}) \propto p(\boldsymbol{t}|\theta)\ \pi_k(\theta|\mathrm{P})$, may be obtained from $\pi(\theta|\mathrm{P})$ by formal use of Bayes' theorem, so that $\pi(\theta|\boldsymbol{t},\ \mathrm{P}) \propto p(\boldsymbol{t}|\theta)\ \pi(\theta|\mathrm{P})$.
The limiting procedure in the definition of a reference prior is not some kind of asymptotic approximation, but an essential element of the definition, required to capture the basic concept of missing information. Notice that, by definition, reference distributions depend only on the asymptotic behaviour of the assumed probability model, a feature which greatly simplifies their actual derivation.
Reference prior *functions* are often simply called reference priors, even though they are usually improper. 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 would have been obtained from prior beliefs which, when compared with the information which data could provide, are relatively uninformative with respect to the quantity of interest.
If θ may take only a *finite* number m of different values, the missing information about θ associated to the prior $p(\theta)$ is its entropy, $H\left\{p(\theta)\right\} = -\sum_{j=1}^{m} p(\theta_j)\, \log\, p(\theta_j)$. Hence the reference prior $\pi(\theta|\mathrm{P})$ is in this case is the prior with *maximum entropy* within P. In particular, if P contains all priors over $\{\theta_1, \ldots, \theta_m\}$, then the reference prior when θ is the quantity of interest is the uniform prior $\pi(\theta) = \{1/m, \ldots, 1/m\}$.
If the sufficient statistic $\boldsymbol{t}$ is a consistent, asymptotically sufficient estimator $\tilde{\theta}$ of a continuous parameter θ, and the class of priors is the set P_0 of all strictly positive priors, then the reference prior is simply

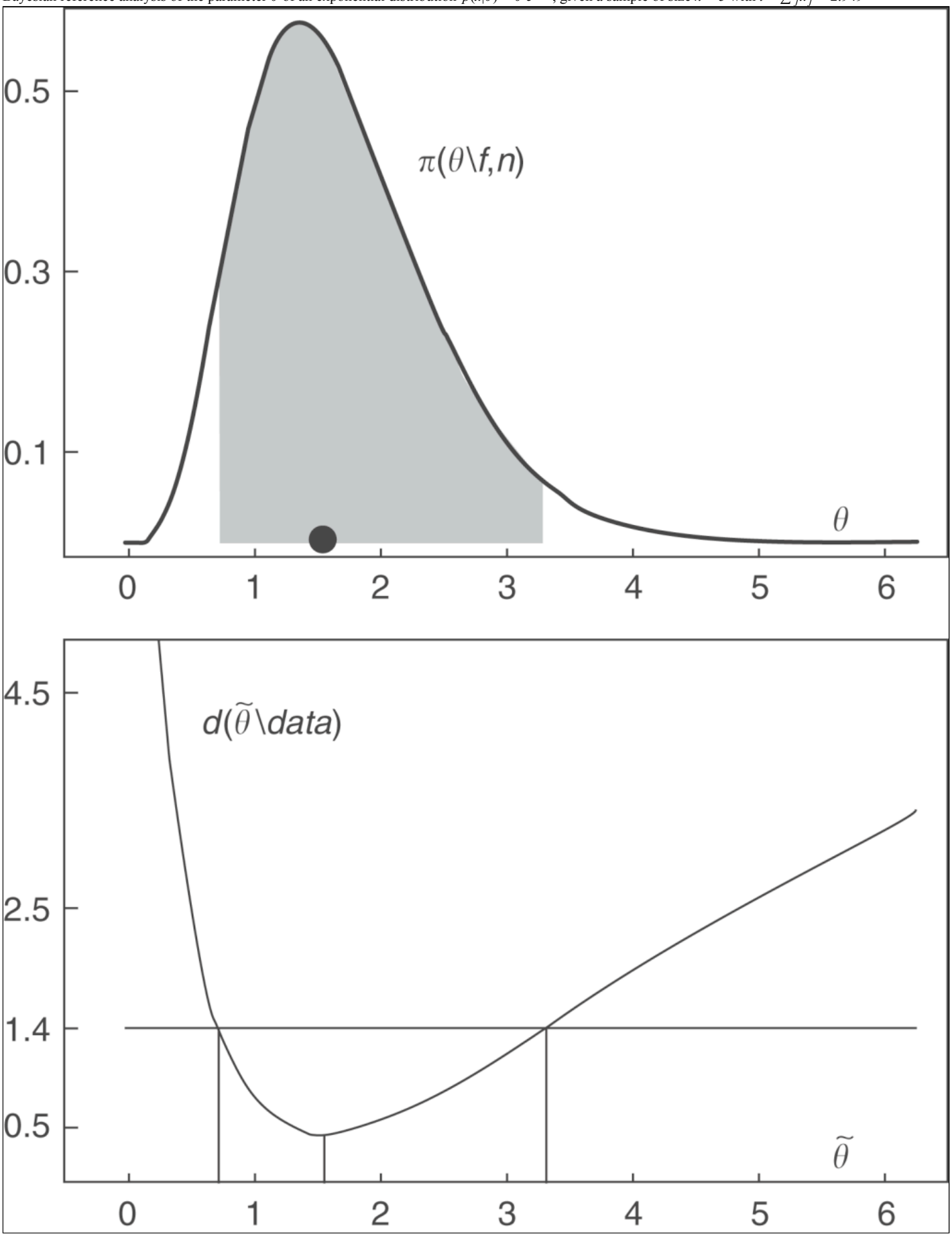
$$\pi(\theta|\mathrm{P}_0) \propto p(\theta|\tilde{\theta})|_{\tilde{\theta}=\theta} \propto p(\theta|\tilde{\theta})|_{\tilde{\theta}=\theta}, \tag{4}$$

where $p(\tilde{\theta}|\theta)$ is any asymptotic approximation to the posterior distribution of θ, and $p(\tilde{\theta}|\theta)$ is the sampling distribution of $\tilde{\theta}$. Under conditions which guarantee asymptotic posterior normality, this reduces to *Jeffreys prior*, $\pi(\theta_0|\mathrm{P}) \propto F(\theta)^{1/2}$, where $F(\theta)$ is Fisher information function. One-parameter reference priors are consistent under re-parametrization; 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 1.: **Exponential data**. If $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a random sample from $\theta e^{-\theta x}$, the reference prior is Jeffreys prior $\pi(\theta) = \theta^{-1}$, and the reference posterior is a gamma distribution $\pi(\theta|x) = Ga(\theta|n,\ t)$, where $t = \sum_{j=1}^{n} x_j$. With a random sample of size $n = 5$ (simulated from an exponential distribution with $\theta = 2$), which yielded a sufficient statistic $t = \sum_j x_j = 2.949$, the result is represented in the upper panel of Figure 1. Inferences about the value of a future observation

from the same process may are described by the reference predictive posterior

$$\pi(x|\boldsymbol{t}) = \int_0^\infty \theta e^{-\theta x} \mathrm{Ga}(\theta|n,\, t) d\theta = n\, t^n (x+t)^{-(n+1)}.$$

Figure 1
Bayesian reference analysis of the parameter θ of an exponential distribution $p(x|\theta) = \theta\, e^{-x\theta}$, given a sample of size $n = 5$ with $t = \sum_j x_j = 2.949$



Nuisance parameters

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 one drops explicit mention to the class P of priors compatible with accepted assumptions to simplify notation, if the probability model is $\{p(\boldsymbol{t}|\theta, \lambda),\ \theta \in \Theta,\ \lambda \in \Lambda\}$ and a θ-reference prior $\pi_\theta(\theta, \lambda)$ is required, the reference algorithm proceeds in two steps:

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

$$p(\boldsymbol{t}|\theta) = \int_\Lambda p(\boldsymbol{t}|\theta, \lambda)\ \pi(\lambda|\theta)\ d\lambda$$

to which the one-parameter algorithm may be applied again to obtain $\pi(\theta)$. The θ-reference prior is then $\pi_\theta(\theta, \lambda) = \pi(\lambda|\theta)\ \pi(\theta)$, and the required reference posterior is $\pi(\theta|\boldsymbol{t}) \propto p(\boldsymbol{t}|\theta)\pi(\theta)$.

If the conditional reference prior $\pi(\lambda|\theta)$ 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 pointwise limit $\pi(\theta|\boldsymbol{t}) = \lim_i \pi_i(\theta|\boldsymbol{t})$.

The θ-reference prior does not depend on the choice of the nuisance parameter 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 may be different from the conditions under which the missing information about some function $\phi = \phi(\theta, \lambda)$ is maximized.

The preceding algorithm may be generalized to any number of parameters. Thus, if the model is $p(\boldsymbol{t}|\omega_1, \ldots, \omega_m)$, a reference prior $\pi(\theta_m|\theta_{m-1}, \ldots, \theta_1)\times\cdot\times\pi(\theta_2|\theta_1)\times\pi(\theta_1)$ may sequentially be obtained for each ordered parametrization $\{\theta_1(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$ of interest, and these are invariant under re-parametrization of any of the $\theta_i(\omega)$'s. The choice of the ordered parametrization $\{\theta_1, \ldots, \theta_m\}$ precisely describes the particular prior required.

Flat priors

Mathematical convenience often leads to the use of 'flat' priors, typically some limiting form of a convenient family of priors; this may, however, have devastating consequences. Consider, for instance, that in a normal setting $p(\boldsymbol{x}|m) = N_k(\bar{x}_i|m, n^{-1}I)$, inferences are desired on $\theta = \sum_{i=1}^k \mu_i^2$, the squared distance of the unknown mean $\boldsymbol{\mu}$ to the origin. It is easily verified that the posterior distribution of θ based on a uniform prior on $\boldsymbol{\mu}$ (or in any 'flat' proper approximation) is strongly inconsistent (Stein's paradox). This is due to the fact that a uniform (or nearly uniform) prior on $\boldsymbol{\mu}$ is highly informative about θ, introducing a severe bias on its marginal posterior. The reference prior which corresponds to a parametrization of the form $\{\theta, \boldsymbol{\lambda}\}$ produces, however, for any choice of the nuisance parameter vector $\boldsymbol{\lambda}$, a reference posterior $\pi(\theta|\boldsymbol{x}, \mathrm{P}_0) \propto \theta^{-1/2}\chi^2(nt|k, n\theta)$, where $t = \sum_{i=1}^k \bar{x}_i^2$, with appropriate consistency properties. Far from being specific to Stein's example, the inappropriate behaviour in problems with many parameters of specific marginal posterior distributions derived from multivariate 'flat' priors (proper or improper) is indeed very frequent. Hence, sloppy, uncontrolled use of 'flat' priors (rather than the relevant reference priors) should be very strongly discouraged.

Inference summaries

From a Bayesian perspective, the final outcome of a problem of inference about any unknown quantity is the corresponding posterior distribution. Thus, given some data D and conditions H, all that can be said about any function $q = q(\boldsymbol{\omega})$ of the parameters which govern the model is contained in the posterior distribution $p(q|D, H)$, and all that can be said about some function $\boldsymbol{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\boldsymbol{y}|D, H)$. 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 (*a*) providing values of the quantity of interest which, in the light of the data, are likely to be a good proxy for its true (unknown) value, and by (*b*) 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. The Bayesian counterparts of those of traditional problems of estimation and hypothesis testing are now briefly considered.

Point estimation

Let D be the available data, which are assumed to have been generated by a probability model $\{p(D|\boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and let $q = q(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be the quantity of interest. A *point estimator* of $\boldsymbol{\theta}$ is some function of the data $\tilde{q} = \tilde{q}(D)$ 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 $\boldsymbol{\Theta}$ of possible $\boldsymbol{\theta}$ values. As dictated by the foundations of decision theory, to solve this decision problem it is necessary to specify a loss function $\boldsymbol{\ell}(\tilde{q}, q)$ measuring the consequences of acting as if the true value of the quantity of interest were $\tilde{q}$, when it is actually $\boldsymbol{\theta}$. The expected posterior loss if $\tilde{q}$ were used is

$$l(\tilde{q}|D) = \int_\Theta \boldsymbol{\ell}(\tilde{q}, q)\, p(q|D)\, dq, \tag{5}$$

and the corresponding *Bayes estimator* is that function of the data, $q^* = q^*(D)$, which minimizes $l(\tilde{q}|D)$.

For any given model, data and prior, the Bayes estimator obviously depends on the loss function which has been chosen. The loss function is context specific, and should be selected in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for scientific communication. These loss functions produce estimates which may often be regarded as simple descriptions of the location of the posterior distribution. If the loss function is quadratic, so that $\boldsymbol{\ell}(\tilde{q}, q) = (\tilde{q} - q)^t(\tilde{q} - q)$, the corresponding Bayes estimator is the posterior mean $E[q|D]$ (on the assumption that the mean exists). Similarly, if the loss function is a zero-one function, so that $\boldsymbol{\ell}(\tilde{q}, q) = 0$ if $\tilde{q}$ belongs to a ball or radius ε centred in $\boldsymbol{\theta}$ and $\boldsymbol{\ell}(\tilde{q}, q) = 1$ otherwise, the corresponding Bayes estimator converges to the posterior mode as the ball radius ε tends to zero (on the assumption that a unique mode exists). If θ is

univariate and the loss function is linear, so that $\boldsymbol{\ell}(\tilde{\theta}, \theta) = c_1(\tilde{\theta} - \theta)$ if $\tilde{\theta} \geq \theta$, and $\boldsymbol{\ell}(\tilde{\theta}, \theta) = c_2(\theta - \tilde{\theta})$ otherwise, 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 corresponding Bayes estimator is the posterior median. The results quoted for linear loss functions clearly illustrate the fact that any possible parameter value may turn out be a Bayes estimator: it all depends on the loss function characterizing the consequences of the anticipated uses of the estimate.

Conventional loss functions are typically non-invariant under re-parametrization, so that the Bayes estimator ϕ^* of a one-to-one transformation $\phi = \phi(q)$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\phi(\boldsymbol{\theta}^*)$ (the univariate posterior median, which is invariant, is an interesting exception). Moreover, conventional loss functions focus on the discrepancy between the estimate $\tilde{q}$ and the true value $\boldsymbol{\theta}$, rather then on the more relevant discrepancy between the probability models which they label. Intrinsic losses directly focus on the discrepancy between the probability distributions $p(D|q)$ and $p(D|\tilde{q})$, and typically produce invariant solutions. An attractive example is the intrinsic discrepancy $\delta(\tilde{q}, q)$, defined as the minimum logarithmic divergence between a probability model labelled by $\boldsymbol{\theta}$ and a probability model labelled by $\tilde{q}$. When there are no nuisance parameters, this is

$$\delta(\tilde{q}, q) = \min\left\{\varkappa(\tilde{q}|q), \varkappa(q|\tilde{q})\right\}, \varkappa(q_i|\tilde{q}_j) = \int_{\mathrm{T}} p(\boldsymbol{t}|q_j) \log \frac{p(\boldsymbol{t}|q_j)}{p(\boldsymbol{t}|q_i)} d\boldsymbol{t}, \tag{6}$$

where $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathrm{T}$is *any* sufficient statistic (which may well be the whole data set D). The definition is easily extended to problems with nuisance parameters. The Bayes estimator is obtained by minimizing the corresponding posterior expected loss. An objective estimator, the *intrinsic estimator* $\tilde{q}_{int} = \tilde{q}_{int}(D)$, is obtained by minimizing the expected intrinsic discrepancy with respect to the *reference* posterior distribution,

$$d(\tilde{q}|d) = \int_{\boldsymbol{\Theta}} \delta(\tilde{q}, q)\pi(q|D)dq \tag{7}$$

Since the intrinsic discrepancy is invariant under re-parametrization, minimizing its posterior expectation produces *invariant* estimators. Thus, the intrinsic estimator of say, the log of the speed of a galaxy is simply log of the intrinsic estimator of the speed of the galaxy.

Region estimation

To describe the inferential content of the posterior distribution of the quantity of interest $p(q|D)$ it is often convenient to quote *credible* regions, defined as subsets of the parameter space $\boldsymbol{\Theta}$ of given posterior probability. For example, the identification of regions containing 50, 90, 95, or 99 per cent of the probability under the posterior may be sufficient to convey the general quantitative messages implicit in $p(q|D)$. Indeed, this is the intuitive basis of graphical representations of univariate distributions like those provided by boxplots. A posterior *q-credible region* for $\boldsymbol{\theta}$ is any region $C \subset \boldsymbol{\Theta}$ such that $\int_C p(q|D)dq = q$. Notice that this provides immediately a direct intuitive statement about the unknown quantity of interest $\boldsymbol{\theta}$ in probability terms, in marked contrast to the circumlocutory statements provided by conventional confidence intervals. A credible region is invariant under re-parametrization; thus, for any q-credible region C for $\boldsymbol{\theta}$, $\phi(C)$ is a q-credible region for $\phi = \phi(q)$.

Clearly, for any given q there are generally infinitely many credible regions. Credible regions are often 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 re-parametrization: the image $\phi(C)$ of an HPD region C 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 $100q$ per cent posterior quantile of $\theta \in \Theta \subset \Re$, then $C = \{\theta;\ \theta \leq \theta_q\}$ is a one-sided, typically unique q-credible region, and it is invariant under re-parametrization; the similarly invariant probability centred q-credible regions of the form $C = \{\theta;\ \theta_{(1-q)/2} \leq \theta \leq \theta_{(1+q)/2}\}$ are easier to compute than HPD regions; this notion, however, does not extend to multivariate problems.

Choosing a p-credible region may be seen as a decision problem where the action space is the class of all p-credible regions. Foundations then dictate that a loss function $\boldsymbol{\ell}(\tilde{q}, q)$ must be specified, and that the region chosen should consist of those $\tilde{\boldsymbol{\theta}}$ values with the lowest expected posterior loss $l(\tilde{q}|D) = \int_{\boldsymbol{\Theta}} \boldsymbol{\ell}(\tilde{q}|q)\, p(q|D)dq$. By definition, lowest posterior loss (LPL) regions are credible regions where all points in the region have smaller expected posterior loss than all points outside. If the loss function is quadratic, so that $\boldsymbol{\ell}(\tilde{q}, q) = (\tilde{q} - q)^t(\tilde{q} - q)$, the LPL p-credible region is a Euclidean sphere centred at the posterior mean $E[\boldsymbol{\theta}|D]$. Like HPD regions, LDL quadratic credible regions are not invariant under re-parametrization; however, LDL intrinsic regions, which minimize the posterior expectation of the invariant intrinsic discrepancy loss (6) are obviously invariant. *Intrinsic p-credible* regions are LDL intrinsic regions which minimize the expected intrinsic discrepancy with respect to the reference posterior distribution. These provide a general, invariant, objective solution to multivariate region estimation. The notions of point and region parameter estimation described above may easily extended to prediction problems by using the posterior predictive rather than the posterior of the parameter.

Hypothesis testing

The posterior distribution $p(q|D)$ 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* $q \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ of the possible values of the quantity of interest (where $\boldsymbol{\Theta}_0$ 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 $\boldsymbol{\Theta}_0$ would greatly simplify the model or because there are additional, context-specific arguments suggesting that $q \in \boldsymbol{\Theta}_0$. Intuitively, the *hypothesis* $H_0 \equiv \{q \in \boldsymbol{\Theta}_0\}$ should be judged to be *compatible* with the observed data D if there are elements in $\boldsymbol{\Theta}_0$ with a relatively high posterior density; however, a more precise conclusion is often required and, once again, this is possible with a decision-oriented approach. Formally, testing the hypothesis $H_0 \equiv \{q \in \boldsymbol{\Theta}_0\}$ 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, $\boldsymbol{\ell}(a_i, q)$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value $\boldsymbol{\theta}$ of the vector of interest. The optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting, $\int_{\boldsymbol{\Theta}} \boldsymbol{\ell}(a_0, q)\ p(q|D)\ dq$, is larger than the expected posterior loss of rejecting, $\int_{\boldsymbol{\Theta}} \boldsymbol{\ell}(a_1, \boldsymbol{\theta})\ p(\boldsymbol{\theta}|D)\ d\boldsymbol{\theta}$, that is, if (and only if)

$$\int_{\boldsymbol{\Theta}} [\boldsymbol{\ell}(a_0, q) - \boldsymbol{\ell}(a_1, q)]\, p(q|D)dq = \int_{\boldsymbol{\Theta}} \Delta\boldsymbol{\ell}(q)\, p(q|D)dq > 0$$

(8)

Therefore, only the loss difference $\Delta\boldsymbol{\ell}(q) = \boldsymbol{\ell}(a_0, q) - \boldsymbol{\ell}(a_1, q)$, which measures the *advantage* of rejecting H_0 as a function of $\boldsymbol{\theta}$, has to be specified: the hypothesis H_0 should be rejected whenever the expected advantage of rejecting H_0 is positive.

The simplest loss structure has the zero-one form given by $\{\boldsymbol{\ell}(a_0, q) = 0,\ \boldsymbol{\ell}(a_1, q) = 1\}$ if $q \in \boldsymbol{\Theta}_0$ and, similarly, $\{\boldsymbol{\ell}(a_0, q) = 1,\ \boldsymbol{\ell}(a_1, q) = 0\}$ if $q \notin \boldsymbol{\Theta}_0$, so that the *advantage* $\Delta\boldsymbol{\ell}(q)$ of rejecting H_0 is 1 if $q \notin \boldsymbol{\Theta}_0$ and it is –1 otherwise. With this, rather naive, loss function the optimal action is to reject H_0 if (and only if) $\Pr(q \notin \boldsymbol{\Theta}_0|D) > \Pr(q \in \boldsymbol{\Theta}_0|D)$. Notice that this formulation requires that $\Pr(q \in \boldsymbol{\Theta}_0) > 0$, that is, that the hypothesis H_0 has a strictly positive prior probability. If $\boldsymbol{\theta}$ is a continuous parameter and $\boldsymbol{\Theta}_0$ consists of a single point $\boldsymbol{\theta}_0$ (sharp null problems), this requires the use of a non-regular highly informative prior which places a positive probability mass at $\boldsymbol{\theta}_0$. This posterior probability approach is therefore only appropriate if it is sensible to condition on the assumption that $\boldsymbol{\theta}$ is indeed concentrated around $\boldsymbol{\theta}_0$.

Frequently, however, the compatibility of the observed data with H_0 is to be judged without assuming such a sharp prior knowledge. In those situations, the advantage $\Delta\boldsymbol{\ell}(q)$ of rejecting H_0 as a function of $\boldsymbol{\theta}$ may be typically assumed to be of the general form $\Delta\boldsymbol{\ell}(q) = \delta(\boldsymbol{\Theta}_0, q) - d^*$, for some $d^* > 0$, where $\delta(\boldsymbol{\Theta}_0, q)$ is some measure of the discrepancy between the assumed model $p(D|q)$ and its closest approximation within the class $\{p(D|q_0),\ q_0 \in \boldsymbol{\Theta}_0\}$ and such that $\delta(\Theta_0, q) = 0$ whenever $q \in \boldsymbol{\Theta}_0$, and d^* is a context dependent *utility constant* which measures the (necessarily positive) advantage of being able to work with the restricted model when it is true. For reasons similar to those supporting its use in estimation, an attractive choice for the loss function $\delta(\boldsymbol{\Theta}_0, q)$ is an appropriate extension of the intrinsic discrepancy loss; when there are no nuisance parameters, this is given by $\delta(\boldsymbol{\Theta}_0, q) = \inf_{q_0 \in \boldsymbol{\Theta}_0} \delta(q_0, q)$ where $\delta(q_0, q)$ is the intrinsic discrepancy loss defined by (6). The corresponding optimal strategy, called the 'Bayesian reference criterion' (BRC), is then to reject H_0 if, and only if,

$$d(\boldsymbol{\Theta}_0|D) = \int_{\Theta} \delta(\boldsymbol{\Theta}_0, q)\pi(q|D)\, dq > d^*.$$

(9)

The choice of d^* plays a similar role to the choice of the significance level in conventional hypothesis testing. Standard choices for scientific communication may be of the form $d^* = \log k$ for, in view of (6) and of (7), this means that the data D are expected to be at least k times more likely under the true model than under H_0. This is actually equivalent to rejecting H_0 if $\boldsymbol{\Theta}_0$ is not contained in an intrinsic q_k-credible region for $\boldsymbol{\theta}$ whose size q_k depends on k. Under conditions for asymptotic posterior normality,

$$q_k \approx 2\,\Phi[(2 \log\ k - 1)^{1/2}] - 1,$$

where Φ is the standard normal distribution function. For instance, if $k = 100$, $q_k \approx 0.996$, while if $k = 11.25$, $q_k \approx 0.95$. The Bayesian reference criterion provides a general objective procedure for multivariate hypothesis testing which is invariant under re-parametrization.

Example 2.: **Exponential data, continued**. The intrinsic discrepancy loss for an exponential model is $\delta(\tilde{\theta}, \theta) = g(\phi)$, if $\phi \leq 1$, and $\delta(\tilde{\theta}, \theta) = g(1/\phi)$, if $\phi > 1$, where $g(\phi) = \phi - 1 - \log \phi$, and $\phi = \tilde{\theta}/\theta$. Using (7) with the data from Example 1, the expected intrinsic loss $d(\tilde{\theta}|x)$ is the function represented in the lower panel of Figure 1. The intrinsic estimate is the value which minimizes $d(\tilde{\theta}|x)$, $\tilde{\theta}_{int} = 1.546$ (marked with a solid dot in the figure), and the intrinsic 0.90-credible set is (0.720,3.290), the set of parameter values with expected loss below 1.407 (corresponding to the shaded area in the upper panel of the figure).

See Also

- Bayes, Thomas
- Bayesian econometrics
- Bayesian methods in macroeconometrics
- Bayesian nonparametrics
- Bayesian time series analysis
- de Finetti, Bruno
- Savage, Leonard J. (Jimmie)
- statistical decision theory

Bibliography

Berger, J. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Bernardo, J. 2005. Reference analysis. In *Handbook of Statistics* 25, ed. D. Dey and C. Rao. Amsterdam: North-Holland, 17–90.

Bernardo, J.M. and Smith, A.F.M. 1994. *Bayesian Theory*. Chichester: Wiley (2nd edn available in 2007).

Berry, D. 1996. *Statistics: A Bayesian Perspective*. Belmont, CA: Wadsworth.

Box, G. and Tiao, G. 1973. *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. 2003. *Bayesian Data Analysis*, 2nd edn. London: Chapman and Hall.

Jeffreys, H. 1961. *Theory of Probability*, 3rd edn. Oxford: University Press.

Lee, P. 2004. *Bayesian Statistics: An Introduction*, 3rd edn. London: Arnold.

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

O'Hagan, A. 2004. *Bayesian Inference*, 2nd edn. London: Arnold.

Robert, C. 2001. *The Bayesian Choice*, 2nd edn. New York: Springer.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

How to cite this article

Bernardo, José M. "Bayesian statistics." The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 22 April 2009 <http://www.dictionaryofeconomics.com /article?id=pde2008_B000314> doi:10.1057/9780230226203.0111(available via http://dx.doi.org/)