Bayesian Statistics

David Draper
Department of Applied Mathematics and Statistics,
Baskin School of Engineering, University of California,
Santa Cruz, USA

Article Outline

Glossary

Bayes' theorem; prior, likelihood and posterior distributions Given (a) θ, something of interest which is unknown to the person making an uncertainty assessment, conveniently referred to as You, (b) y, an information source which is relevant to decreasing Your uncertainty about θ, (c) a desire to learn about θ from y in a way that is both internally and externally logically consistent, and (d) $\mathcal{B}$, Your background assumptions and judgments about how the world works, as these assumptions and judgments relate to learning about θ from y, it can be shown that You are compelled in this situation to reason within the standard rules of probability as the basis of Your *inferences* about θ, *predictions* of future data y^*, and *decisions* in the face of uncertainty (see below for contrasts between inference, prediction and decision-making), and to quantify Your uncertainty about any unknown quantities through conditional probability distributions. When inferences about θ are the goal, Bayes' Theorem provides a means of combining all relevant information internal and external to y:

$$p(\theta|y, \mathcal{B}) = c\, p(\theta|\mathcal{B})\, l(\theta|y, \mathcal{B})\,. \tag{1}$$

Here, for example in the case in which θ is a real-valued vector of length k, (a) $p(\theta|\mathcal{B})$ is Your *prior distribution* about θ given $\mathcal{B}$ (in the form of a probability density function), which quantifies all relevant information available to You about θ external to y, (b) c is a positive normalizing constant, chosen to make the density on the left side of the equation integrate to 1, (c) $l(\theta|y, \mathcal{B})$ is Your *likelihood distribution* for θ given y and $\mathcal{B}$, which is defined to be a density-normalized multiple of Your *sampling distribution* $p(\cdot|\theta, \mathcal{B})$ for future data values y^* given θ and $\mathcal{B}$, but re-interpreted as a function of θ for fixed y, and (d) $p(\theta|y, \mathcal{B})$ is Your *posterior distribution* about θ given y and $\mathcal{B}$, which summarizes Your current total information about θ and solves the basic inference problem.

Bayesian parametric and non-parametric modeling (1) Following de Finetti [23], a Bayesian statistical model is a joint predictive distribution $p(y_1, \ldots, y_n)$ for observable quantities y_i that have not yet been observed, and about which You are therefore uncertain. When the y_i are real-valued, often You will not regard them as probabilistically *independent* (informally, the y_i are independent if information about any of them does not help You to predict the others); but it may be possible to identify a *parameter vector* $\theta = (\theta_1, \ldots, \theta_k)$ such that You would judge the y_i *conditionally independent* given θ, and would therefore be willing to model them via the relation

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta)\,. \tag{2}$$

When combined with a prior distribution $p(\theta)$ on θ that is appropriate to the context, this is *Bayesian parametric modeling*, in which $p(y_i|\theta)$ will often have a standard distributional form (such as binomial, Poisson or Gaussian). (2) When a (finite) parameter vector that induces conditional independence cannot be found, if You judge your uncertainty about the real-valued y_i *exchangeable* (see below), then a representation theorem of de Finetti [21] states informally that all internally logically consistent predictive distributions $p(y_1, \ldots, y_n)$ can be expressed in a way that is equivalent to the *hierarchical model* (see below)

$$\begin{aligned} (F|\mathcal{B}) &\sim p(F|\mathcal{B}) \\ (y_i|F, \mathcal{B}) &\stackrel{\text{IID}}{\sim} F\,, \end{aligned} \tag{3}$$

where (a) F is the cumulative distribution function (CDF) of the underlying process $(y_1, y_2, \ldots)$ from which You are willing to regard $p(y_1, \ldots, y_n)$ as (in effect) like a random sample and (b) $p(F|\mathcal{B})$ is Your prior distribution on the space $\mathcal{F}$ of all CDFs on the real line. This (placing probability distributions on infinite-dimensional spaces such as $\mathcal{F}$) is *Bayesian nonparametric modeling*, in which priors involving *Dirichlet processes* and/or *Pólya trees* (see Sect. "Inference: Parametric and Non-Parametric Modeling of Count Data") are often used.

Exchangeability A sequence $y = (y_1, \ldots, y_n)$ of random variables (for $n \geq 1$) is *(finitely) exchangeable* if the joint probability distribution $p(y_1, \ldots, y_n)$ of the elements of y is invariant under permutation of the indices $(1, \ldots, n)$, and a countably infinite sequence $(y_1, y_2, \ldots)$ is *(infinitely) exchangeable* if every finite subsequence is finitely exchangeable.

Hierarchical modeling Often Your uncertainty about something unknown to You can be seen to have a *nested* or *hierarchical* character. One class of examples arises in *cluster sampling* in fields such as education and medicine, in which students (level 1) are nested within classrooms (level 2) and patients (level 1) within hospitals (level 2); cluster sampling involves random samples (and therefore uncertainty) at two or more levels in such a data hierarchy (examples of this type of hierarchical modeling are given in Sect. "Strengths and Weaknesses of the Two Approaches"). Another, quite different, class of examples of Bayesian hierarchical modeling is exemplified by equation (3) above, in which is was helpful to decompose Your overall predictive uncertainty about $(y_1, \ldots, y_n)$ into (a) uncertainty about F and then (b) uncertainty about the y_i given F (examples of this type of hierarchical modeling appear in Sect. "Inference and Prediction: Binary Outcomes with No Covariates" and "Inference: Parametric and Non-Parametric Modeling of Count Data").

Inference, prediction and decision-making; samples and populations Given a data source y, *inference* involves drawing probabilistic conclusions about the underlying process that gave rise to y, *prediction* involves summarizing uncertainty about future observable data values y^*, and decision-making involves looking for optimal behavioral choices in the face of uncertainty (about either the underlying process, or the future, or both). In some cases inference takes the form of reasoning backwards from a *sample* of data values to a *population*: a (larger) universe of possible data values from which You judge that the sample has been drawn in a manner that is *representative* (i. e., so that the sampled and unsampled values in the population are (likely to be) similar in relevant ways).

Mixture modeling Given y, unknown to You, and $\mathcal{B}$, Your background assumptions and judgments relevant to y, You have a choice: You can either model (Your uncertainty about) y directly, through the probability distribution $p(y|\mathcal{B})$, or (if that is not feasible) You can identify a quantity x upon which You judge y to depend and model y hierarchically, in two stages: first by modeling x, through the probability distribution $p(x|\mathcal{B})$, and then by modeling y given x, through the probability distribution $p(y|x, \mathcal{B})$:

$$p(y|\mathcal{B}) = \int_{\mathcal{X}} p(y|x, \mathcal{B})\, p(x|\mathcal{B})\, \mathrm{d}x\,, \tag{4}$$

where $\mathcal{X}$ is the space of possible values of x over which Your uncertainty is expressed. This is *mixture modeling*, a special case of hierarchical modeling (see above). In hierarchical notation (4) can be re-expressed as

$$y = \left\{ \begin{array}{c} x \\ (y|x) \end{array} \right\} . \tag{5}$$

Examples of mixture modeling in this article include (a) equation (3) above, with F playing the role of x; (b) the basic equation governing Bayesian prediction, discussed in Sect. "The Bayesian Statistical Paradigm"; (c) Bayesian model averaging (Sect. "The Bayesian Statistical Paradigm"); (d) de Finetti's representation theorem for binary outcomes (Sect. "Inference and Prediction: Binary Outcomes with No Covariates"); (e) random-effects parametric and non-parametric modeling of count data (Sect. "Inference: Parametric and Non-Parametric Modeling of Count Data"); and (f) integrated likelihoods in Bayes factors (Sect. "Decision-Making: Variable Selection in Generalized Linear Models; Bayesian Model Selection").

Probability – frequentist and Bayesian In the *frequentist* probability paradigm, attention is restricted to phenomena that are inherently repeatable under (essentially) identical conditions; then, for an event A of interest, $P_f(A)$ is the limiting relative frequency with which A occurs in the (hypothetical) repetitions, as the number of repetitions $n \to \infty$. By contrast, Your Bayesian probability $P_B(A|\mathcal{B})$ is the numerical weight of evidence, given Your background information $\mathcal{B}$ relevant to A, in favor of a true-false proposition A whose truth status is uncertain to You, obeying a series of reasonable axioms to ensure that Your Bayesian probabilities are internally logically consistent.

Utility To ensure internal logical consistency, optimal decision-making proceeds by (a) specifying a *utility* function $U(a, \theta_0)$ quantifying the numerical value associated with taking action a if the unknown is really θ_0 and (b) *maximizing expected utility*, where the expectation is taken over uncertainty in θ as quantified by the posterior distribution $p(\theta|y, \mathcal{B})$.

Definition of the Subject and Introduction

Statistics may be defined as the study of uncertainty: how to measure it, and how to make choices in the face of it. Uncertainty is quantified via *probability*, of which there are two leading paradigms, *frequentist* (discussed in Sect. "Comparison with the Frequentist Statistical Paradigm") and *Bayesian*. In the Bayesian approach to probability the primitive constructs are true-false *propositions A* whose truth status is uncertain, and the probability of A is the numerical weight of evidence in favor of A, constrained to obey a set of axioms to ensure that Bayesian probabilities are *coherent* (internally logically consistent).

The discipline of statistics may be divided broadly into four activities: *description* (graphical and numerical summaries of a data set y, without attempting to reason outward from it; this activity is almost entirely non-probabilistic and will not be discussed further here), *inference* (drawing probabilistic conclusions about the underlying process that gave rise to y), *prediction* (summarizing uncertainty about future observable data values y^*), and *decision-making* (looking for optimal behavioral choices in the face of uncertainty). Bayesian statistics is an approach to inference, prediction and decision-making that is based on the Bayesian probability paradigm, in which uncertainty about an unknown θ (this is the inference problem) is quantified by means of a conditional probability distribution $p(\theta|y, \mathcal{B})$; here y is all available relevant data and $\mathcal{B}$ summarizes the background assumptions and judgments of the person making the uncertainty assessment. Prediction of a future y^* is similarly based on the conditional probability distribution $p(y^*|y, \mathcal{B})$, and optimal decision-making proceeds by (a) specifying a *utility* function $U(a, \theta_0)$ quantifying the numerical reward associated with taking action a if the unknown is really θ_0 and (b) *maximizing expected utility*, where the expectation is taken over uncertainty in θ as quantified by $p(\theta|y, \mathcal{B})$.

The Bayesian Statistical Paradigm

Statistics is the branch of mathematical and scientific inquiry devoted to the study of uncertainty: its consequences, and how to behave sensibly in its presence. The subject draws heavily on *probability*, a discipline which predates it by about 100 years: basic probability theory can be traced [48] to work of Pascal, Fermat and Huygens in the 1650s, and the beginnings of statistics [34,109] are evident in work of Bayes published in the 1760s.

The Bayesian statistical paradigm consists of three basic ingredients:

- θ, something of interest which is unknown (or only partially known) to the person making the uncertainty assessment, conveniently referred to, in a convention proposed by Good (1950), as You. Often θ is a *parameter* vector of real numbers (of finite length k, say) or a matrix, but it can literally be almost anything: for example, a function (three leading examples are a cumulative distribution function (CDF), a density, or a regression surface), a phylogenetic tree, an image of a region on the surface of Mars at a particular moment in time, . . .
- y, an information source which is relevant to decreasing Your uncertainty about θ. Often y is a vector of real numbers (of finite length n, say), but it can also literally be almost anything: for instance, a time series, a movie, the text in a book, . . .
- A desire to learn about θ from y in a way that is both *coherent* (internally consistent: in other words, free of internal logical contradictions; Bernardo and Smith [11] give a precise definition of coherence) and *well-calibrated* (externally consistent: for example, capable of making accurate predictions of future data y^*).

It turns out [23,53] that You are compelled in this situation to reason within the standard rules of probability (see below) as the basis of Your inferences about θ, predictions of future data y^*, and decisions in the face of uncertainty, and to quantify Your uncertainty about any unknown quantities through conditional probability distributions, as in the following three basic equations of Bayesian statistics:

$$\begin{aligned} p(\theta|y, \mathcal{B}) &= c\, p(\theta|\mathcal{B})\, l(\theta|y, \mathcal{B}) \\ p(y^*|y, \mathcal{B}) &= \int_{\Theta} p(y^*|\theta, \mathcal{B})\, p(\theta|y, \mathcal{B})\, d\theta \\ a^* &= \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, E_{(\theta|y,\mathcal{B})} \left[U(a, \theta) \right] . \end{aligned} \tag{6}$$

(The basic rules of probability [71] are: for any true-false propositions A and B and any background assumptions and judgments $\mathcal{B}$, (*convexity*) $0 \le P(A|\mathcal{B}) \le 1$, with equality at 1 iff A is known to be true under $\mathcal{B}$; (*multiplication*) $P(A \text{ and } B|\mathcal{B}) = P(A|\mathcal{B})\, P(B|A, \mathcal{B}) = P(B|\mathcal{B})\, P(A|B, \mathcal{B})$; and (*addition*) $P(A \text{ or } B|\mathcal{B}) = P(A|\mathcal{B}) + P(B|\mathcal{B}) - P(A \text{ and } B|\mathcal{B})$.)

The meaning of the equations in (6) is as follows.

- $\mathcal{B}$ stands for Your background (often not fully stated) assumptions and judgments about how the world works, as these assumptions and judgments relate to learning about θ from y. $\mathcal{B}$ is often omitted from the basic equations (sometimes with unfortunate consequences), yielding the simpler-looking forms

$$\begin{aligned} p(\theta|y) &= c\, p(\theta)\, l(\theta|y) \\ p(y^*|y) &= \int_{\Theta} p(y^*|\theta)\, p(\theta|y)\, d\theta \\ a^* &= \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, E_{(\theta|y)} \left[U(a, \theta) \right] . \end{aligned} \tag{7}$$

- $p(\theta|\mathcal{B})$ is Your *prior information* about θ given $\mathcal{B}$, in the form of a probability density function (PDF) or probability mass function (PMF) if θ lives continuously or discretely on $\mathbb{R}^k$ (this is generically referred to as

Your *prior distribution*), and $p(\theta|y,\mathcal{B})$ is Your *posterior distribution* about θ given y and $\mathcal{B}$, which summarizes Your current total information about θ and solves the basic inference problem. These are actually not very good names for $p(\theta|\mathcal{B})$ and $p(\theta|y,\mathcal{B})$, because (for example) $p(\theta|\mathcal{B})$ really stands for all (relevant) information about θ (given $\mathcal{B}$) external to y, whether that information was obtained before (or after) y arrives, but (a) they do emphasize the sequential nature of learning and (b) through long usage it would be difficult for more accurate names to be adopted.

- c (here and throughout) is a generic positive normalizing constant, inserted into the first equation in (6) to make the left-hand side integrate (or sum) to 1 (as any coherent distribution must).
- $p(y^*|\theta,\mathcal{B})$ is Your *sampling distribution* for future data values y^* given θ and $\mathcal{B}$ (and presumably You would use the same sampling distribution $p(y|\theta,\mathcal{B})$ for (past) data values y, mentally turning the clock back to a point before the data arrives and thinking about what values of y You might see). This assumes that You are willing to regard Your data as like random draws from a *population* of possible data values (an heroic assumption in some cases, for instance with observational rather than randomized data; this same assumption arises in the frequentist statistical paradigm, discussed below in Sect. "Comparison with the Frequentist Statistical Paradigm").
- $l(\theta|y,\mathcal{B})$ is Your *likelihood function* for θ given y and $\mathcal{B}$, which is defined to be any positive constant multiple of the sampling distribution $p(y|\theta,\mathcal{B})$ but re-interpreted as a function of θ for fixed y:

$$l(\theta|y,\mathcal{B}) = c\,p(y|\theta,\mathcal{B})\,. \tag{8}$$

The likelihood function is also central to one of the main approaches to frequentist statistical inference, developed by Fisher [37]; the two approaches are contrasted in Sect."Comparison with the Frequentist Statistical Paradigm".

All of the symbols in the first equation in (6) have now been defined, and this equation can be recognized as *Bayes' Theorem*, named after Bayes [5] because a special case of it appears prominently in work of his that was published posthumously. It describes how to pass coherently from information about θ external to y (quantified in the prior distribution $p(\theta|\mathcal{B})$) to information both internal and external to y (quantified in the posterior distribution $p(\theta|y,\mathcal{B})$), via the likelihood function $l(\theta|y,\mathcal{B})$: You multiply the prior and likelihood pointwise in θ and normalize so that the posterior distribution $p(\theta|y,\mathcal{B})$ integrates (or sums) to 1.

- According to the second equation in (6), $p(y^*|y,\mathcal{B})$, Your (posterior) *predictive distribution* for future data y^* given (past) data y and $\mathcal{B}$, which solves the basic prediction problem, must be a weighted average of Your sampling distribution $p(y^*|\theta,\mathcal{B})$ weighted by Your current best information $p(\theta|y,\mathcal{B})$ about θ given y and $\mathcal{B}$; in this integral Θ is the space of possible values of θ over which Your uncertainty is expressed. (The second equation in (6) contains a simplifying assumption that should be mentioned: in full generality the first term $p(y^*|\theta,\mathcal{B})$ inside the integral would be $p(y^*|y,\theta,\mathcal{B})$, but it is almost always the case that the information in y is redundant in the presence of complete knowledge of θ, in which case $p(y^*|y,\theta,\mathcal{B}) = p(y^*|\theta,\mathcal{B})$; this state of affairs could be described by saying that *the past and future are conditionally independent given the truth*. A simple example of this phenomenon is provided by coin-tossing: if You are watching a Bernoulli(θ) process unfold (see Sect. "Inference and Prediction: Binary Outcomes with No Covariates") whose success probability θ is unknown to You, the information that 8 of the first 10 tosses have been heads is definitely useful to You in predicting the 11th toss, but if instead You somehow knew that θ was 0.7, the outcome of the first 10 tosses would be irrelevant to You in predicting any future tosses.)
- Finally, in the context of making a choice in the face of uncertainty, $\mathcal{A}$ is Your set of possible actions, $U(a,\theta_0)$ is the numerical value (*utility*) You attach to taking action a if the unknown is really θ_0 (specified, without loss of generality, so that large utility values are preferred by You), and the third equation in (6) says that to make the choice coherently You should find the action a^* that *maximizes expected utility* (MEU); here the expectation

$$E_{(\theta|y,\mathcal{B})}\left[U(a,\theta)\right] = \int_{\Theta} U(a,\theta)\,p(\theta|y,\mathcal{B})\,\mathrm{d}\theta \tag{9}$$

is taken over uncertainty in θ as quantified by the posterior distribution $p(\theta|y,\mathcal{B})$.

This summarizes the entire Bayesian statistical paradigm, which is driven by the three equations in (6). Examples of its use include clinical trial design [56] and analysis [105]; spatio-temporal modeling, with environmental applications [101]; forecasting and dynamic linear models [115]; non-parametric estimation of receiver operating characteristic curves, with applications in medicine and agriculture [49]; finite selection models, with health policy applications [79]; Bayesian CART model search, with applications in breast cancer research [16]; construction of radiocarbon calibration curves, with archaeological applications [15]; factor regression models, with applications to gene expression data [114]; mixture modeling for high-density genotyping arrays, with bioinformatic applications [100]; the EM algorithm for Bayesian fitting of latent process models [76]; state-space modeling, with applications in particle-filtering [92]; causal inference [42,99]; hierarchical modeling of DNA sequences, with genetic and medical applications [77]; hierarchical Poisson regression modeling, with applications in health care evaluation [17]; multiscale modeling, with engineering and financial applications [33]; expected posterior prior distributions for model selection [91]; nested Dirichlet processes, with applications in the health sciences [96]; Bayesian methods in the study of sustainable fisheries [74,82]; hierarchical non-parametric meta-analysis, with medical and educational applications [81]; and structural equation modeling of multilevel data, with applications to health policy [19].

Challenges to the paradigm include the following:

- **Q:** How do You specify the sampling distribution/likelihood function that quantifies the information about the unknown θ internal to Your data set y? **A:** (1) The solution to this problem, which is common to all approaches to statistical inference, involves imagining future data y^* from the same process that has yielded or will yield Your data set y; often the variability You expect in future data values can be quantified (at least approximately) through a standard parametric family of distributions (such as the Bernoulli/binomial for binary data, the Poisson for count data, and the Gaussian for real-valued outcomes) and the parameter vector of this family becomes the unknown θ of interest. (2) Uncertainty in the likelihood function is referred to as *model uncertainty* [67]; a leading approach to quantifying this source of uncertainty is *Bayesian model averaging* [18,25,52], in which uncertainty about the models M in an ensemble $\mathcal{M}$ of models (specifying $\mathcal{M}$ is part of $\mathcal{B}$) is assessed and propagated for a quantity, such as a future data value y^*, that is common to all models via the expression

$$p(y^*|y,\mathcal{B}) = \int_{\mathcal{M}} p(y^*|y,M,\mathcal{B})\, p(M|y,\mathcal{B})\, \mathrm{d}M\,. \quad (10)$$

  In other words, to make coherent predictions in the presence of model uncertainty You should form a weighted average of the conditional predictive distributions $p(y^*|y,M,\mathcal{B})$, weighted by the posterior model probabilities $p(M|y,\mathcal{B})$. Other potentially useful approaches to model uncertainty include *Bayesian non-parametric modeling*, which is examined in Sect. "Inference: Parametric and Non-Parametric Modeling of Count Data", and methods based on *cross-validation* [110], in which (in Bayesian language) part of the data is used to specify the prior distribution on $\mathcal{M}$ (which is an input to calculating the posterior model probabilities) and the rest of the data is employed to update that prior.
- **Q:** How do You quantify information about the unknown θ external to Your data set y in the prior probability distribution $p(\theta|\mathcal{B})$? **A:** (1) There is an extensive literature on *elicitation* of prior (and other) probabilities; notable references include O'Hagan et al. [85] and the citations given there. (2) If θ is a parameter vector and the likelihood function is a member of the *exponential family* [11], the prior distribution can be chosen in such a way that the prior and posterior distributions for θ have the same mathematical form (such a prior is said to be *conjugate* to the given likelihood); this may greatly simplify the computations, and often prior information can (at least approximately) be quantified by choosing a member of the conjugate family (see Sect."Inference and Prediction: Binary Outcomes with No Covariates" for an example of both of these phenomena).
  In situations where it is not precisely clear how to quantify the available information external to y, two sets of tools are available:
  - *Sensitivity analysis* [30], also known as *pre-posterior analysis* [4]: Before the data have begun to arrive, You can (a) generate data similar to what You expect You will see, (b) choose a plausible prior specification and update it to the posterior on the quantities of greatest interest, (c) repeat (b) across a variety of plausible alternatives, and (d) see if there is substantial stability in conclusions across the variations in prior specification. If so, fine; if not, this approach can be combined with hierarchical modeling [68]: You can collect all of the plausible priors and add a layer hierarchically to the prior specification, with the new layer indexing variation across the prior alternatives.
  - *Bayesian robustness* [8,95]: If, for example, the context of the problem implies that You only wish to specify that the prior distribution belongs to an infinite-dimensional class (such as, for priors on $(0,1)$, the class of monotone non-increasing functions) with (for instance) bounds on the first two moments, You can in turn quantify bounds on summaries of the resulting posterior distribution, which may be narrow enough to demonstrate that Your

uncertainty in specifying the prior does not lead to differences that are large in practical terms.

Often context suggests specification of a prior that has relatively little information content in relation to the likelihood information; for reasons that are made clear in Sect. "Inference and Prediction: Binary Outcomes with No Covariates", such priors are referred to as relatively *diffuse* or *flat* (the term *non-informative* is sometimes also used, but this seems worth avoiding, because any prior specification takes a particular position regarding the amount of relevant information external to the data). See Bernardo [10] and Kass and Wasserman [60] for a variety of formal methods for generating diffuse prior distributions.

- **Q:** How do You quantify Your utility function $U(a, \theta)$ for optimal decision-making? **A:** There is a rather less extensive statistical literature on elicitation of utility than probability; notable references include Fishburn [35,36], Schervish et al. [103], and the citations in Bernardo ans Smith [11]. There is a parallel (and somewhat richer) economics literature on utility elicitation; see, for instance, Abdellaoui [1] and Blavatskyy [12]. Sect. "Decision-Making: Variable Selection in Generalized Linear Models; Bayesian Model Selection" provides a decision-theoretic example.
- Suppose that $\theta = (\theta_1, \ldots, \theta_k)$ is a parameter vector of length k. Then (a) computing the normalizing constant in Bayes' Theorem

$$c = \left(\int \cdots \int p(y|\theta_1, \ldots, \theta_k, \mathcal{B}) \cdot p(\theta_1, \ldots, \theta_k|\mathcal{B}) \, d\theta_1 \cdots d\theta_k \right)^{-1} \quad (11)$$

involves evaluating a k-dimensional integral; (b) the predictive distribution in the second equation in (6) involves another k-dimensional integral; and (c) the posterior $p(\theta_1, \ldots, \theta_k|y, \mathcal{B})$ is a k-dimensional probability distribution, which for $k > 2$ can be difficult to visualize, so that attention often focuses on the *marginal* posterior distributions

$$p(\theta_j|y, \mathcal{B}) = \int \cdots \int p(\theta_1, \ldots, \theta_k|y, \mathcal{B}) \, d\theta_{-j} \quad (12)$$

for $j = 1, \ldots k$, where θ_{-j} is the θ vector with component j omitted; each of these marginal distributions involves a $(k-1)$-dimensional integral. If k is large these integrals can be difficult or impossible to evaluate exactly, and a general method for computing accurate approximations to them proved elusive from the time of Bayes in the eighteenth century until recently (in the late eighteenth century Laplace [63]) developed an analytical method, which today bears his name, for approximating integrals that arise in Bayesian work [11], but his method is not as general as the computationally-intensive techniques in widespread current use). Around 1990 there was a fundamental shift in Bayesian computation, with the belated discovery by the statistics profession of a class of techniques – *Markov chain Monte Carlo* (MCMC) methods [41,44] – for approximating high-dimensional Bayesian integrals in a computationally-intensive manner, which had been published in the chemical physics literature in the 1950s [78]; these methods came into focus for the Bayesian community at a moment when desktop computers had finally become fast enough to make use of such techniques.

MCMC methods approximate integrals associated with the posterior distribution $p(\theta|y, \mathcal{B})$ by (a) creating a Markov chain whose equilibrium distribution is the desired posterior and (b) sampling from this chain from an initial $\theta_{(0)}$ (i) until equilibrium has been reached (all draws up to this point are typically discarded) and (ii) for a sufficiently long period thereafter to achieve the desired approximation accuracy. With the advent and refinement of MCMC methods since 1990, the Bayesian integration problem has been solved for a wide variety of models, with more ambitious sampling schemes made possible year after year with increased computing speeds: for instance, in problems in which the dimension of the parameter space is not fixed in advance (an example is *regression change-point* problems [104], where the outcome y is assumed to depend linearly (apart from stochastic noise) on the predictor(s) x but with an unknown number of changes of slope and intercept and unknown locations for those changes), ordinary MCMC techniques will not work; in such problems methods such as *reversible-jump MCMC* [47,94] and *Markov birth-death processes* [108], which create Markov chains that permit trans-dimensional jumps, are required.

The main drawback of MCMC methods is that they do not necessarily scale well as n (the number of data observations) increases; one alternative, popular in the machine learning community, is *variational* methods [55], which convert the integration problem into an optimization problem by (a) approximating the posterior distribution of interest by a family of distributions yielding a closed-form approximation to the integral and (b) finding the member of the family that maximizes the accuracy of the approximation.

- Bayesian decision theory [6], based on maximizing expected utility, is unambiguous in its normative recommendation for how a single agent (You) should make a choice in the face of uncertainty, and it has had widespread success in fields such as economics (e. g., [2,50]) and medicine (e. g., [87,116]). It is well known, however [3,112]), that Bayesian decision theory (or indeed any other formal approach that seeks an optimal behavioral choice) can be problematic when used normatively for group decision-making, because of conflicts in preferences among members of the group. This is an important unsolved problem.

Three Examples

Inference and Prediction: Binary Outcomes with No Covariates

Consider the problem of measuring the quality of care at a particular hospital H. One way to do this is to examine the outcomes of that care, such as mortality, after adjusting for the burden of illness brought by the patients to H on admission. As an even simpler version of this problem, consider just the n binary mortality observables $y = (y_1, \ldots, y_n)$ (with mortality measured within 30 days of admission, say; 1 = died, 0 = lived) that You will see from all of the patients at H with a particular admission diagnosis (heart attack, say) during some prespecified future time window. You acknowledge Your uncertainty about which elements in the sequence will be 0s and which 1s, and You wish to quantify this uncertainty using the Bayesian paradigm. As de Finetti [20] noted, in this situation Your fundamental imperative is to construct a predictive distribution $p(y_1, \ldots, y_n|\mathcal{B})$ that expresses Your uncertainty about the future observables, rather than – as is perhaps more common – to reach immediately for a standard family of parametric models for the y_i (in other words, to posit the existence of a vector $\theta = (\theta_1, \ldots, \theta_k)$ of parameters and to model the observables by appeal to a family $p(y_i|\theta, \mathcal{B})$ of probability distributions indexed by θ).

Even though the y_i are binary, with all but the smallest values of n it still seems a formidable task to elicit from Yourself an n-dimensional predictive distribution $p(y_1, \ldots, y_n|\mathcal{B})$. De Finetti [20] showed, however, that the task is easier than it seems. In the absence of any further information about the patients, You notice that Your uncertainty about them is *exchangeable*: if someone (without telling You) were to rearrange the order in which their mortality outcomes become known to You, Your predictive distribution would not change. This still seems to leave $p(y_1, \ldots, y_n|\mathcal{B})$ substantially unspecified (where $\mathcal{B}$ now includes the judgment of exchangeability of the y_i), but de Finetti [20] proved a remarkable theorem which shows (in effect) that all exchangeable predictive distributions for a vector of binary observables are representable as *mixtures* of Bernoulli sampling distributions: if You're willing to regard $(y_1, \ldots, y_n)$ as the first n terms in an *infinitely exchangeable* binary sequence $(y_1, y_2, \ldots)$ (which just means that every finite subsequence is exchangeable), then to achieve coherence Your predictive distribution must be expressible as

$$p(y_1, \ldots, y_n|\mathcal{B}) = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n}\, p(\theta|\mathcal{B})\, d\theta\,, \quad (13)$$

where $s_n = \sum_{i=1}^n y_i$. Here the quantity θ on the right side of (13) is more than just an integration variable: the equation says that in Your predictive modeling of the binary y_i You may as well proceed *as if*

- There is a quantity called θ, interpretable both as the marginal death probability $p(y_i = 1|\theta, \mathcal{B})$ for each patient and as the long-run mortality rate in the infinite sequence $(y_1, y_2, \ldots)$ (which serves, in effect, as a population of values to which conclusions from the data can be generalized);
- Conditional on θ and $\mathcal{B}$, the y_i are independent identically distributed (IID) Bernoulli (θ); and
- θ can be viewed as a realization of a random variable with density $p(\theta|\mathcal{B})$.

In other words, exchangeability of Your uncertainty about a binary process is functionally equivalent to assuming the simple Bayesian *hierarchical* model [27]

$$\begin{aligned} (\theta|\mathcal{B}) &\sim p(\theta|\mathcal{B}) \\ (y_i|\theta, \mathcal{B}) &\stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta)\,, \end{aligned} \quad (14)$$

and $p(\theta|\mathcal{B})$ is recognizable as Your prior distribution for θ, the underlying death rate for heart attack patients similar to those You expect will arrive at hospital H during the relevant time window.

Consider now the problem of quantitatively specifying prior information about θ. From (13) and (14) the likelihood function is

$$l(\theta|y, \mathcal{B}) = c\, \theta^{s_n}(1-\theta)^{n-s_n}\,, \quad (15)$$

which (when interpreted in the Bayesian manner as a density in θ) is recognizable as a member of the Beta family of probability distributions: for $\alpha, \beta > 0$ and $0 < \theta < 1$,

$$\theta \sim \text{Beta}(\alpha, \beta) \text{ iff } p(\theta) = c\, \theta^{\alpha-1}(1-\theta)^{\beta-1}\,. \quad (16)$$

Moreover, this family has the property that the product of two Beta densities is another Beta density, so by Bayes'

Theorem if the prior $p(\theta|\mathcal{B})$ is chosen to be Beta(α, β) for some (as-yet unspecified) $\alpha > 0$ and $\beta > 0$, then the posterior will be Beta$(\alpha + s_n, \beta + n - s_n)$: this is conjugacy (Sect. "The Bayesian Statistical Paradigm") of the Beta family for the Bernoulli/binomial likelihood. In this case the conjugacy leads to a simple interpretation of α and β: the prior acts like a data set with α 1s and β 0s, in the sense that if person 1 does a Bayesian analysis with a Beta(α, β) prior and sample data $y = (y_1, \ldots, y_n)$ and person 2 instead merges the corresponding "prior data set" with y and does a maximum-likelihood analysis (Sect. "Comparison with the Frequentist Statistical Paradigm") on the resulting merged data, the two people will get the same answers. This also shows that the *prior sample size* n_0 in the Beta–Bernoulli/binomial model is $(\alpha + \beta)$. Given that the mean of a Beta(α, β) distribution is $\alpha / (\alpha + \beta)$, calculation reveals that the posterior mean $(\alpha + s_n) / (\alpha + \beta + n)$ of θ is a weighted average of the prior mean and the data mean $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, with prior and data weights n_0 and n, respectively:

$$\frac{\alpha + s_n}{\alpha + \beta + n} = \frac{n_0 \left(\frac{\alpha}{\alpha + \beta} \right) + n \bar{y}}{n_0 + n} . \tag{17}$$

These facts shed intuitive light on how Bayes' Theorem combines information internal and external to a given data source: thinking of prior information as equivalent to a data set is a valuable intuition, even in non-conjugate settings.

The choice of α and β naturally depends on the available information external to y. Consider for illustration two such specifications:

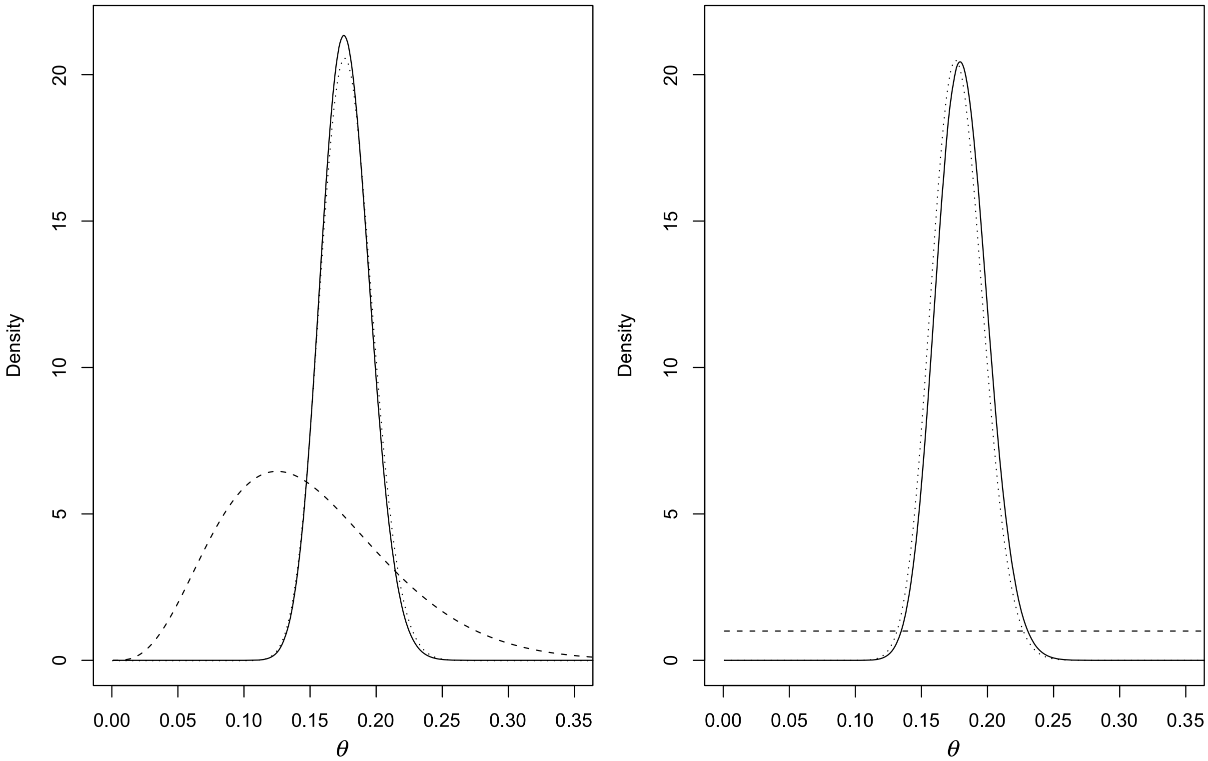
- Analyst 1 does a web search and finds that the 30-day mortality rate for heart attack (given average quality of care and average patient sickness at admission) in her country is 15%. The information she has about hospital H is that its care and patient sickness are not likely to be wildly different from the country averages but that a mortality deviation from the mean, if present, would be more likely to occur on the high side than the low. Having lived in the community served by H for some time and having not heard anything either outstanding or deplorable about the hospital, she would be surprised to find that the underlying heart attack death rate at H was less than (say) 5% or greater than (say) 30%. One way to quantify this information is to set the prior mean to 15% and to place (say) 95% of the prior mass between 5% and 30%.
- Analyst 2 has little information external to y and thus wishes to specify a relatively diffuse prior distribution that does not dramatically favor any part of the unit interval.

Numerical integration reveals that $(\alpha, \beta) = (4.5, 25.5)$, with a prior sample size of 30.0, satisfies Analyst 1's constraints. Analyst 2's diffuse prior evidently corresponds to a rather small prior sample size; a variety of positive values of α and β near 0 are possible, all of which will lead to a relatively flat prior.

Suppose for illustration that the time period in question is about four years in length and H is a medium-size US hospital; then there will be about $n = 385$ heart attack patients in the data set y. Suppose further that the observed mortality rate at H comes out $\bar{y} = s_n / n = 69 / 385 \doteq 18\%$. Figure 1 summarizes the prior-to-posterior updating with this data set and the two priors for Analysts 1 (left panel) and 2 (right panel), with $\alpha = \beta = 1$ (the Uniform distribution) for Analyst 2. Even though the two priors are rather different – Analyst 1's prior is skewed, with a prior mean of 0.15 and $n_0 \doteq 30$; Analyst 2's prior is flat, with a prior mean of 0.5 and $n_0 = 2$ – it is evident that the posterior distributions are nearly the same in both cases; this is because the data sample size $n = 385$ is so much larger than either of the prior sample sizes, so that the likelihood information dominates. With both priors the likelihood and posterior distributions are nearly the same, another consequence of $n_0 \ll n$. For Analyst 1 the posterior mean, standard deviation, and 95% central posterior interval for θ are $(0.177, 0.00241, 0.142, 0.215)$, and the corresponding numerical results for Analyst 2 are $(0.181, 0.00258, 0.144, 0.221)$; again it is clear that the two sets of results are almost identical. With a large sample size, careful elicitation – like that undertaken by Analyst 1 – will often yield results similar to those with a diffuse prior.

The posterior predictive distribution $p(y_{n+1}|y_1, \ldots y_n, \mathcal{B})$ for the next observation, having observed the first n, is also straightforward to calculate in closed form with the conjugate prior in this model. It is clear that $p(y_{n+1}|y, \mathcal{B})$ has to be a Bernoulli(θ^*) distribution for some θ^*, and intuition says that θ^* should just be the mean $\alpha^* / (\alpha^* + \beta^*)$ of the posterior distribution for θ given y, in which $\alpha^* = \alpha + s_n$ and $\beta^* = \beta + n - s_n$ are the parameters of the Beta posterior. To check this, making use of the fact that the normalizing constant in the Beta(α, β) family is $\Gamma(\alpha + \beta) / \Gamma(\alpha) \, \Gamma(\beta)$, the second equation in (6) gives

$$\begin{aligned} &p(y_{n+1}|y_1, \ldots y_n, \mathcal{B}) \\ &\quad = \int_0^1 \theta^{y_{n+1}} (1 - \theta)^{1 - y_{n+1}} \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*) \, \Gamma(\beta^*)} \theta^{\alpha^* - 1} \\ &\qquad \cdot (1 - \theta)^{\beta^* - 1} \, \mathrm{d}\theta \end{aligned}$$



Bayesian Statistics, Figure 1

Prior-to-posterior updating with two prior specifications in the mortality data set (in both panels, prior: *long dotted lines*; likelihood: *short dotted lines*; posterior: *solid lines*). The left and right panels give the updating with the priors for Analysts 1 and 2, respectively

$$
\begin{aligned}
&= \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \int_0^1 \theta^{\alpha^* + y_{n+1} - 1} \\
&\quad \cdot (1-\theta)^{(\beta^* - y_{n+1} + 1) - 1}\, \mathrm{d}\theta \\
&= \left[\frac{\Gamma(\alpha^* + y_{n+1})}{\Gamma(\alpha^*)} \right] \left[\frac{\Gamma(\beta^* - y_{n+1} + 1)}{\Gamma(\beta^*)} \right] \\
&\quad \cdot \left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)} \right] ;
\end{aligned} \tag{18}
$$

this, combined with the fact that $\Gamma(x+1)/\Gamma(x) = x$ for any real x, yields, for example in the case $y_{n+1} = 1$,

$$
\begin{aligned}
p(y_{n+1} = 1 | y, \mathcal{B}) &= \left[\frac{\Gamma(\alpha^* + 1)}{\Gamma(\alpha^*)} \right] \left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)} \right] \\
&= \frac{\alpha^*}{\alpha^* + \beta^*} ,
\end{aligned} \tag{19}
$$

confirming intuition.

Inference: Parametric and Non-Parametric Modeling of Count Data

Most elderly people in the Western world say they would prefer to spend the end of their lives at home, but many instead finish their lives in an institution (a nursing home or hospital). How can elderly people living in their communities be offered health and social services that would help to prevent institutionalization? Hendriksen et al. [51] conducted an experiment in the 1980s in Denmark to test the effectiveness of *in-home geriatric assessment* (IHGA), a form of preventive medicine in which each person's medical and social needs are assessed and acted upon individually. A total of $n = 572$ elderly people living in non-institutional settings in a number of villages were randomized, $n_C = 287$ to a control group, who received standard health care, and $n_T = 285$ to a treatment group, who received standard care plus IHGA. The number of hospitalizations during the two-year life of the study was an outcome of particular interest.

The data are presented and summarized in Table 1. Evidently IHGA lowered the mean hospitalization rate per two years (for the elderly Danish people in the study, at least) by $(0.944 - 0.768) \doteq 0.176$, which is about an 18% reduction from the control level, a clinically large difference. The question then becomes, in Bayesian inferential language: what is the posterior distribution for the treatment effect in the entire population $\mathcal{P}$ of patients judged exchangeable with those in the study?

Bayesian Statistics, Table 1
Distribution of number of hospitalizations in the IHGA study

Group	Number of Hospitalizations									Sample	
	0	1	2	3	4	5	6	7	n	Mean	Variance
Control	138	77	46	12	8	4	0	2	287	0.944	1.54
Treatment	147	83	37	13	3	1	1	0	285	0.768	1.02

Continuing to refer to the relevant analyst as You, with a binary outcome variable and no covariates in Sect. "Inference and Prediction: Binary Outcomes with No Covariates" the model arose naturally from a judgment of exchangeability of Your uncertainty about all n outcomes, but such a judgment of *unconditional* exchangeability would not be appropriate initially here; to make such a judgment would be to assert that the treatment and control interventions have the same effect on hospitalization, and it was the point of the study to see if this is true. Here, at least initially, it would be more scientifically appropriate to assert exchangeability separately and in parallel within the two experimental groups, a judgment de Finetti [22] called *partial exchangeability* and which has more recently been referred to as *conditional exchangeability* [28,72] given the treatment/control status covariate.

Considering for the moment just the control group outcome values C_i, $i = 1, \dots, n_C$, and seeking as in Sect. "Inference and Prediction: Binary Outcomes with No Covariates" to model them via a predictive distribution $p(C_1, \dots, C_{n_C}|\mathcal{B})$, de Finetti's previous representation theorem is not available because the outcomes are real-valued rather than binary, but he proved [21] another theorem for this situation as well: if You're willing to regard $(C_1, \dots, C_{n_C})$ as the first n_C terms in an infinitely exchangeable sequence $(C_1, C_2, \dots)$ of values on $\mathbb{R}$ (which plays the role of the population $\mathcal{P}$, under the control condition, in this problem), then to achieve coherence Your predictive distribution must be expressible as

$$p(C_1, \dots, C_{n_C}|\mathcal{B}) = \int_{\mathcal{F}} \prod_{i=1}^{n_C} F(C_i)\, \mathrm{d}G(F|\mathcal{B})\,; \qquad (20)$$

here (a) F has an interpretation as $F(t) = \lim_{n_C \to \infty} F_{n_C}(t)$, where F_{n_C} is the empirical CDF based on $(C_1, \dots, C_{n_C})$; (b) $G(F|\mathcal{B}) = \lim_{n_C \to \infty} p(F_{n_C}|\mathcal{B})$, where $p(\cdot|\mathcal{B})$ is Your joint probability distribution on $(C_1, C_2, \dots)$; and (c) $\mathcal{F}$ is the space of all possible CDFs on $\mathbb{R}$. Equation (20) says informally that exchangeability of Your uncertainty about an observable process unfolding on the real line is functionally equivalent to assuming the Bayesian hierarchical model

$$\begin{aligned} (F|\mathcal{B}) &\sim p(F|\mathcal{B}) \\ (y_i|F, \mathcal{B}) &\overset{\text{IID}}{\sim} F\,, \end{aligned} \qquad (21)$$

where $p(F|\mathcal{B})$ is a prior distribution on $\mathcal{F}$. Placing distributions on functions, such as CDFs and regression surfaces, is the topic addressed by the field of *Bayesian nonparametric* (BNP) modeling [24,80], an area of statistics that has recently moved completely into the realm of day-to-day implementation and relevance through advances in MCMC computational methods. Two rich families of prior distributions on CDFs about which a wealth of practical experience has recently accumulated include (mixtures of) *Dirichlet processes* [32] and *Pólya trees* [66].

Parametric modeling is of course also possible with the IHGA data: as noted by Krnjajić et al. [62], who explore both parametric and BNP models for data of this kind, Poisson modeling is a natural choice, since the outcome consists of counts of relatively rare events. The first Poisson model to which one would generally turn is a *fixed-effects* model, in which $(C_i|\lambda_C)$ are IID Poisson(λ_C) $(i = 1, \dots, n_C = 287)$ and $(T_j|\lambda_T)$ are IID Poisson(λ_T) $(j = 1, \dots, n_T = 285)$, with a diffuse prior on (λ_C, λ_T) if little is known, external to the data set, about the underlying hospitalization rates in the control and treatment groups. However, the last two columns of Table 1 reveal that the sample variance is noticeably larger than the sample mean in both groups, indicating substantial Poisson over-dispersion. For a second, improved, parametric model this suggests a *random-effects* Poisson model of the form

$$\begin{aligned} (C_i|\lambda_{iC}) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_{iC}) \\ \left[\log(\lambda_{iC})|\beta_{0C}, \sigma_C^2\right] &\overset{\text{IID}}{\sim} N(\beta_{0C}, \sigma_C^2)\,, \end{aligned} \qquad (22)$$

and similarly for the treatment group, with diffuse priors for $(\beta_{0C}, \sigma_C^2, \beta_{0T}, \sigma_T^2)$. As Krnjajić et al. [62] note, from a medical point of view this model is more plausible than the fixed-effects formulation: each patient in the control group has his/her own *latent* (unobserved) underlying rate of hospitalization λ_{iC}, which may well differ from the underlying rates of the other control patients because of unmeasured differences in factors such as health status at the beginning of the experiment (and similarly for the treatment group).

Model (22), when complemented by its analogue in the treatment group, specifies a Lognormal mixture of Poisson distributions for each group and is straightforward to fit by MCMC, but the Gaussian assumption for the mixing distribution is conventional, not motivated by the underlying science of the problem, and if the distribution of the latent variables is not Gaussian – for example, if it is multimodal or skewed – model (22) may well lead to incorrect inferences. Krnjajić et al. [62] therefore also examine several BNP models that are centered on the random-effects Poisson model but which permit learning about the true underlying distribution of the latent variables instead of assuming it is Gaussian. One of their models, when applied (for example) to the control group, was

$$
\begin{array}{rcl}
(C_i|\lambda_{iC}) & \stackrel{\text{indep}}{\sim} & \text{Poisson}(\lambda_{iC}) \\
\left[\log(\lambda_{iC})|G\right] & \stackrel{\text{IID}}{\sim} & G \\
(G|\alpha,\mu,\sigma^2) & \sim & \text{DP}[\alpha\, N(\mu,\sigma^2)]\,.
\end{array}
\tag{23}
$$

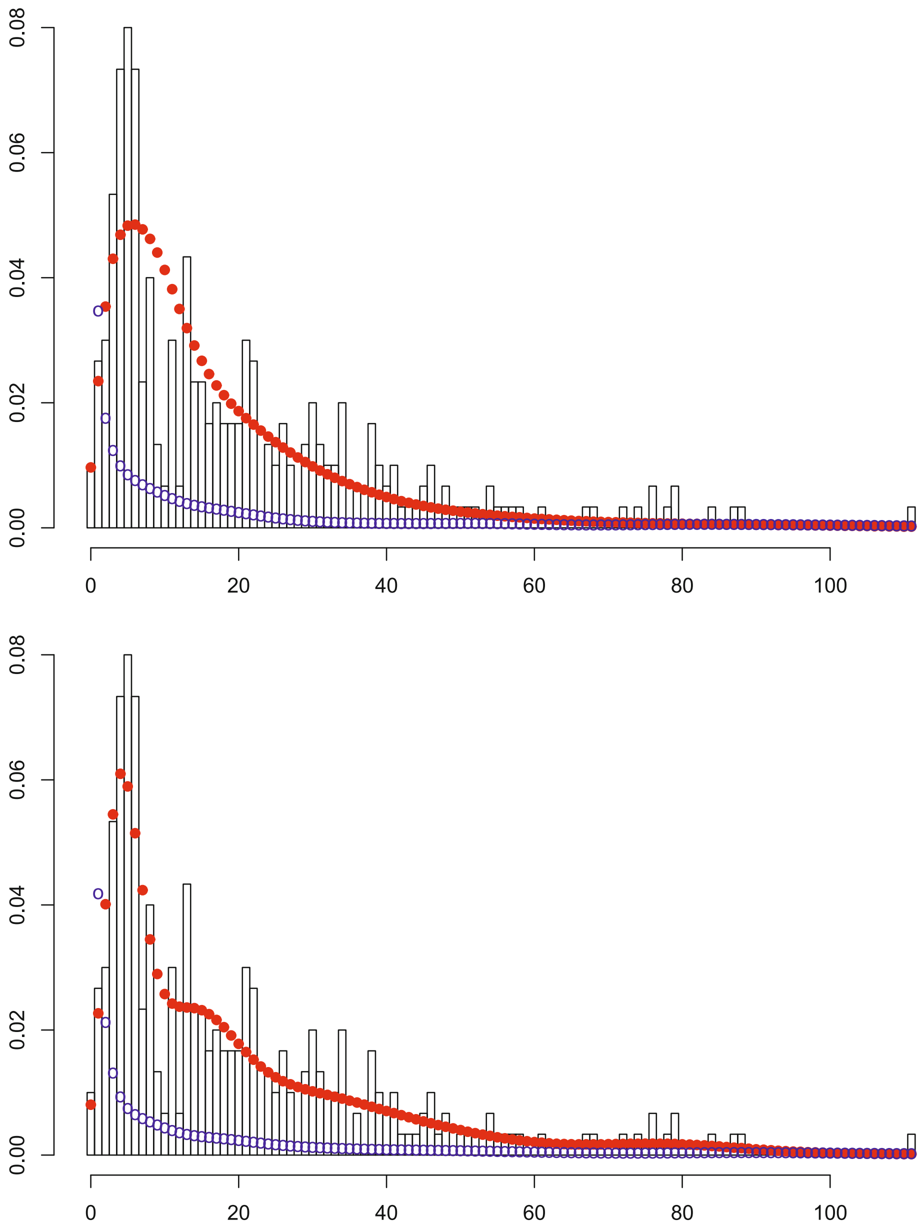
Here $\text{DP}[\alpha\, N(\mu,\sigma^2)]$ refers to a Dirichlet process prior distribution, on the CDF G of the latent variables, which is centered at the $N(\mu,\sigma^2)$ model with *precision parameter* α. Model (23) is an *expansion* of the random-effects Poisson model (22) in that the latter is a special case of the former (obtained by letting $\alpha \to \infty$). *Model expansion* is a common Bayesian analytic tool which helps to assess and propagate model uncertainty: if You are uncertain about a particular modeling detail, instead of fitting a model that assumes this detail is correct with probability 1, embed it in a richer model class of which it is a special case, and let the data tell You about its plausibility.

With the IHGA data, models (22) and (23) turned out to arrive at similar inferential conclusions – in both cases point estimates of the ratio of the treatment mean to the control mean were about 0.82 with a posterior standard deviation of about 0.09, and a posterior probability that the (population) mean ratio was less than 1 of about 0.95, so that evidence is strong that IHGA lowers mean hospitalizations not just in the sample but in the collection $\mathcal{P}$ of elderly people to whom it is appropriate to generalize. But the two modeling approaches need not yield similar results: if the latent variable distribution is far from Gaussian, model (22) will not be able to adjust to this violation of one of its basic assumptions. Krnjajić et al. [62] performed a simulation study in which data sets with 300 observations were generated from various Gaussian and non-Gaussian latent variable distributions and a variety of parametric and BNP models were fit to the resulting count data; Fig. 2 summarizes the prior and posterior predictive distributions from models (22; top panel) and (23; bottom panel) with a bimodal latent variable distribution. The parametric Gaussian random-effects model cannot fit the bimodality on the data scale, but the BNP model – even though centered on the Gaussian as the random-effects distribution – adapts smoothly to the underlying bimodal reality.

Decision-Making: Variable Selection in Generalized Linear Models; Bayesian Model Selection

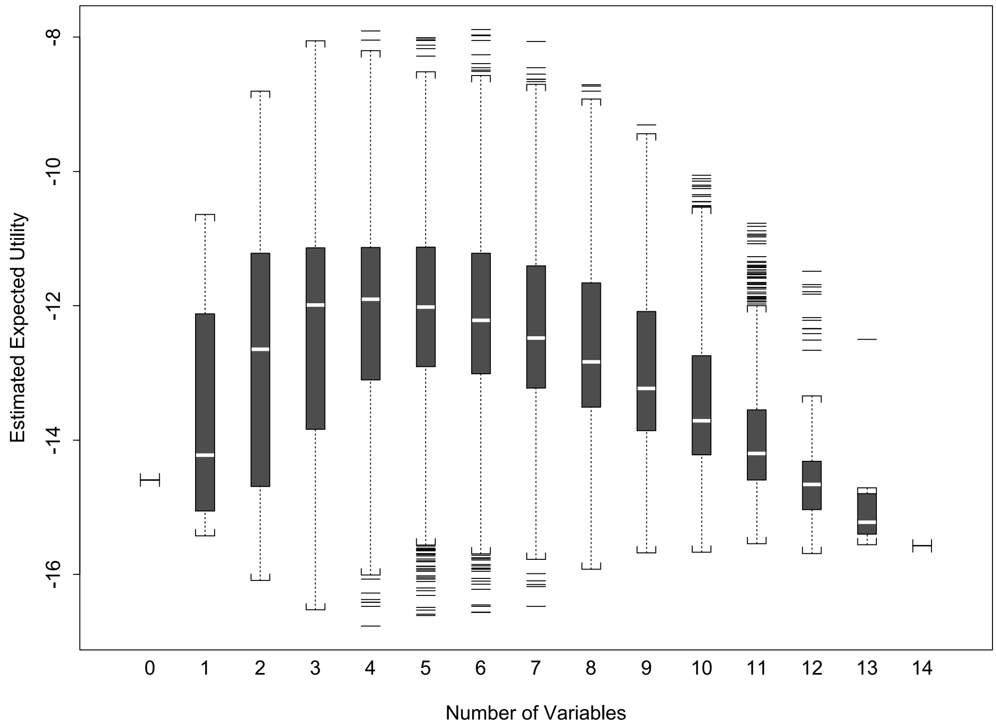
Variable selection (choosing the "best" subset of predictors) in generalized linear models is an old problem, dating back at least to the 1960s, and many methods [113] have been proposed to try to solve it; but virtually all of them ignore an aspect of the problem that can be important: the cost of data collection of the predictors. An example, studied by Fouskakis and Draper [39], which is an elaboration of the problem examined in Sect. "Inference and Prediction: Binary Outcomes with No Covariates", arises in the field of quality of health care measurement, where patient sickness at admission is often assessed by using logistic regression of an outcome, such as mortality within 30 days of admission, on a fairly large number of sickness indicators (on the order of 100) to construct a sickness scale, employing standard variable selection methods (for instance, backward selection from a model with all predictors) to find an "optimal" subset of 10–20 indicators that predict mortality well. The problem with such *benefit-only* methods is that they ignore the considerable differences among the sickness indicators in the *cost* of data collection; this issue is crucial when admission sickness is used to drive programs (now implemented or under consideration in several countries, including the US and UK) that attempt to identify substandard hospitals by comparing observed and expected mortality rates (given admission sickness), because such quality of care investigations are typically conducted under cost constraints. When both data-collection cost and accuracy of prediction of 30-day mortality are considered, a large variable-selection problem arises in which the only variables that make it into the final scale should be those that achieve a cost-benefit tradeoff.

Variable selection is an example of the broader process of *model selection*, in which questions such as "Is model M_1 better than M_2?" and "Is M_1 good enough?" arise. These inquiries cannot be addressed, however, without first answering a new set of questions: good enough (better than) for what purpose? Specifying this purpose [26,57,61,70] identifies model selection as a decision problem that should be approached by constructing a contextually relevant utility function and maximizing expected utility. Fouskakis and Draper [39] create a utility function, for variable selection in their severity of illness problem, with two components that are combined additively: a data-collection component (in monetary units, such as US$), which is simply the negative of the total amount of money required to collect data on a given set of patients with a given subset of the sickness indicators; and a predictive-accuracy component, in which a method



Bayesian Statistics, Figure 2
Prior (*open circles*) and posterior (*solid circles*) predictive distributions under models (22) and (23) (*top* and *bottom* panels, respectively) based on a data set generated from a bimodal latent variable distribution. In each panel, the histogram plots the simulated counts

is devised to convert increased predictive accuracy into decreased monetary cost by thinking about the consequences of labeling a hospital with bad quality of care "good" and vice versa. One aspect of their work, with a data set (from a RAND study: [58]) involving $p = 83$ sickness indicators gathered on a representative sample of $n = 2{,}532$ elderly American patients hospitalized in the period 1980–86 with pneumonia, focused only on the $p = 14$ variables in the original RAND sickness scale; this was chosen because $2^{14} = 16\,384$ was a small enough number of possible models to do brute-force enumeration of the estimated expected utility (EEU) of all the models. Figure 3 is a parallel boxplot of the EEUs of all 16 384 variable subsets, with the boxplots sorted by the number of variables in each model. The model with no predictors does poorly, with an EEU of about US$−14.5, but from a cost-benefit point of



Bayesian Statistics, Figure 3
Estimated expected utility of all 16 384 variable subsets in the quality of care study based on RAND data

view the RAND sickness scale with all 14 variables is even worse (US$−15.7), because it includes expensive variables that do not add much to the predictive power in relation to cheaper variables that predict almost as well. The best subsets have 4–6 variables and would save about US$8 per patient when compared with the entire 14–variable scale; this would amount to significant savings if the observed-versus-expected assessment method were applied widely.

Returning to the general problem of Bayesian model selection, two cases can be distinguished: situations in which the precise purpose to which the model will be put can be specified (as in the variable-selection problem above), and settings in which at least some of the end uses to which the modeling will be put are not yet known. In this second situation it is still helpful to reason in a decision-theoretic way: the hallmark of a good (bad) model is that it makes good (bad) predictions, so a utility function based on predictive accuracy can be a good general-purpose choice. With (a) a single sample of data y, (b) a future data value y^*, and (c) two models M_j $(j = 1, 2)$ for illustration, what is needed is a *scoring rule* that measures the discrepancy between y^* and its predictive distribution $p(y^*|y, M_j, \mathcal{B})$ under model M_j. It turns out [46,86] that the optimal (*impartial, symmetric, proper*) scoring rules are linear functions of $\log p(y^*|y, M_j, \mathcal{B})$, which has a simple intuitive motivation: if the predictive distribution is Gaussian, for example, then values of y^* close to the center (in other words, those for which the prediction has been good) will receive a greater reward than those in the tails. An example [65], in this one-sample setting, of a model selection criterion (a) based on prediction, (b) motivated by utility considerations and (c) with good model discrimination properties [29] is the *full-sample log score*

$$LS_{FS}(M_j|y, \mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p^*(y_i|y, M_j, \mathcal{B}) \,, \qquad (24)$$

which is related to the *conditional predictive ordinate* criterion [90]. Other Bayesian model selection criteria in current use include the following:

- *Bayes factors* [60]: Bayes' Theorem, written in odds form for discriminating between models M_1 and M_2, says that

$$\frac{p(M_1|y, \mathcal{B})}{p(M_2|y, \mathcal{B})} = \left[\frac{p(M_1|\mathcal{B})}{p(M_2|\mathcal{B})}\right] \cdot \left[\frac{p(y|M_1, \mathcal{B})}{p(y|M_2, \mathcal{B})}\right] ; \qquad (25)$$

here the prior odds in favor of M_1,

$$\frac{p(M_1|\mathcal{B})}{p(M_2|\mathcal{B})},$$

are multiplied by the *Bayes factor*

$$\frac{p(y|M_1,\mathcal{B})}{p(y|M_2,\mathcal{B})}$$

to produce the posterior odds

$$\frac{p(M_1|y,\mathcal{B})}{p(M_2|y,\mathcal{B})}.$$

According to the logic of this criterion, models with high posterior probability are to be preferred, and if all the models under consideration are equally plausible a priori this reduces to preferring models with larger Bayes factors in their favor. One problem with this approach is that – in parametric models in which model M_j has parameter vector θ_j defined on parameter space Θ_j – the *integrated likelihoods* $p(y|M_j,\mathcal{B})$ appearing in the Bayes factor can be expressed as

$$\begin{aligned} p(y|M_j,\mathcal{B}) &= \int_{\Theta_j} p(y|\theta_j,M_j,\mathcal{B})\,p(\theta_j|M_j,\mathcal{B})\,\mathrm{d}\theta_j \\ &= E_{(\theta_j|M_j,\mathcal{B})}\left[p(y|\theta_j,M_j,\mathcal{B})\right]. \end{aligned} \tag{26}$$

In other words, the numerator and denominator ingredients in the Bayes factor are each expressible as expectations of likelihood functions with respect to the *prior* distributions on the model parameters, and if context suggests that these priors should be specified diffusely the resulting Bayes factor can be unstable as a function of precisely how the diffuseness is specified. Various attempts have been made to remedy this instability of Bayes factors (for example, {partial, intrinsic, fractional} Bayes factors, well calibrated priors, conventional priors, intrinsic priors, expected posterior priors, …; [9]); all of these methods appear to require an appeal to ad-hockery which is absent from the log score approach.

- *Deviance Information Criterion* (DIC): Given a parametric model $p(y|\theta_j,M_j,\mathcal{B})$, Spiegelhalter et al. [106] define the *deviance information criterion* (*DIC*) (by analogy with other information criteria) to be a tradeoff between (a) an estimate of the model lack of fit, as measured by the *deviance* $D(\overline{\theta}_j)$ (where $\overline{\theta}_j$ is the posterior mean of θ_j under M_j; for the purpose of DIC, the deviance of a model [75] is minus twice the logarithm of the likelihood for that model), and (b) a penalty for model complexity equal to twice the effective number of parameters p_{Dj} of the model:

$$DIC(M_j|y,\mathcal{B}) = D(\overline{\theta}_j) + 2\,\hat{p}_{Dj}. \tag{27}$$

When p_{Dj} is difficult to read directly from the model (for example, in complex hierarchical models, especially those with random effects), Spiegelhalter et al. motivate the following estimate, which is easy to compute from standard MCMC output:

$$\hat{p}_{Dj} = \overline{D(\theta_j)} - D(\overline{\theta}_j); \tag{28}$$

in other words, $\hat{p}_{Dj}$ is the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters.

DIC is available as an option in several MCMC packages, including `WinBUGS` [107] and `MLwiN` [93]. One difficulty with DIC is that the MCMC estimate of p_{Dj} can be poor if the marginal posteriors for one or more parameters (using the parameterization that defines the deviance) are far from Gaussian; reparameterization (onto parameter scales where the posteriors are approximately Normal) helps but can still lead to mediocre estimates of p_{Dj}.

Other notable recent references on the subject of Bayesian variable selection include Brown et al. [13], who examine multivariate regression in the context of compositional data, and George and Foster [43], who use *empirical Bayes* methods in the Gaussian linear model.

Comparison with the Frequentist Statistical Paradigm

Strengths and Weaknesses of the Two Approaches

Frequentist statistics, which has concentrated mainly on inference, proceeds by (i) thinking of the values in a data set y as like a random sample from a *population* $\mathcal{P}$ (a set to which it is hoped that conclusions based on the data can validly be generalized), (ii) specifying a summary θ of interest in $\mathcal{P}$ (such as the population mean of the outcome variable), (iii) identifying a function $\hat{\theta}$ of y that can serve as a reasonable estimate of θ, (iv) imagining repeating the random sampling from $\mathcal{P}$ to get other data sets y and therefore other values of $\hat{\theta}$, and (v) using the random behavior of $\hat{\theta}$ across these repetitions to make inferential probability statements involving θ. A leading implementation of the frequentist paradigm [37] is based on using the value $\hat{\theta}_{\text{MLE}}$ that maximizes the likelihood function as the estimate of θ and obtaining a measure of uncertainty for $\hat{\theta}_{\text{MLE}}$ from the curvature of the logarithm of the likelihood function at its maximum; this is *maximum likelihood inference*.

Each of the frequentist and Bayesian approaches to statistics has strengths and weaknesses.

- The frequentist paradigm has the advantage that repeated-sampling calculations are often more tractable than manipulations with conditional probability distributions, and it has the clear strength that it focuses attention on the scientifically important issue of *calibration*: in settings where the true data-generating process is known (e. g., in simulations of random sampling from a known population $\mathcal{P}$), how often does a particular method of statistical inference recover known truth? The frequentist approach has the disadvantage that it only applies to inherently repeatable phenomena, and therefore cannot be used to quantify uncertainty about many true-false propositions of real-world interest (for example, if You are a doctor to whom a new patient (male, say) has just come, strictly speaking You cannot talk about the frequentist probability that *this* patient is HIV positive; he either is or he is not, and his arriving at Your office is not the outcome of any repeatable process that is straightforward to identify). In practice the frequentist approach also has the weaknesses that (a) model uncertainty is more difficult to assess and propagate in this paradigm, (b) predictive uncertainty assessments are not always straightforward to create from the frequentist point of view (the *bootstrap* [31] is one possible solution) and (c) inferential calibration may not be easy to achieve when the sample size n is small.
  An example of several of these drawbacks arises in the construction of *confidence intervals* [83], in which repeated-sampling statements such as
  $$P_f(\hat{\theta}_{\text{low}} < \theta < \hat{\theta}_{\text{high}}) = 0.95 \tag{29}$$
  (where P_f quantifies the frequentist variability in $\hat{\theta}_{\text{low}}$ and $\hat{\theta}_{\text{high}}$ across repeated samples from $\mathcal{P}$) are interpreted in the frequentist paradigm as suggesting that the unknown θ lies between $\hat{\theta}_{\text{low}}$ and $\hat{\theta}_{\text{high}}$ with 95% "confidence." Two difficulties with this are that (a) equation (29) looks like a probability statement about θ but is not, because in the frequentist approach θ is a fixed unknown constant that cannot be described probabilistically, and (b) with small sample sizes *nominal* 95% confidence intervals based on maximum likelihood estimation can have actual *coverage* (the percentage of time in repeated sampling that the interval includes the true θ) substantially less than 95%.
- The Bayesian approach has the following clear advantages: (a) It applies (at least in principle) to uncertainty about anything, whether associated with a repeatable process or not; (b) inference is unambiguously based on the first equation in (6), without the need to face questions such as what constitutes a "reasonable" estimate of θ (step (iii) in the frequentist inferential paradigm above); (c) prediction is straightforwardly and unambiguously based on the second equation in (6); and (d) in the problem of decision analysis a celebrated theorem of Wald [111] says informally that all good decisions can be interpreted as having been arrived at by maximizing expected utility as in the third equation of (6), so the Bayesian approach appears to be the way forward in decision problems rather broadly (but note the final challenge at the end of Sect. "The Bayesian Statistical Paradigm"). The principal disadvantage of the Bayesian approach is that coherence (internal logical consistency) by itself does not guarantee good calibration: You are free in the Bayesian paradigm to insert strong prior information in the modeling process (without violating coherence), and – if this information is seen after the fact to have been out of step with the world – Your inferences, predictions and/or decisions may also be off-target (of course, the same is true in the both the frequentist and Bayesian paradigms with regard to Your modeling of the likelihood information).

Two examples of frequentist inferences having poor calibration properties in small samples were given by Browne and Draper [14]. Their first example again concerns the measurement of quality of care, which is often studied with *cluster samples*: a random sample of J hospitals (indexed by j) and a random sample of N total patients (indexed by i) nested in the chosen hospitals is taken, and quality of care for the chosen patients and various hospital- and patient-level predictors are measured. With y_{ij} as the quality of care score for patient i in hospital j, a first step would often be to fit a *variance-components model* with random effects at both the hospital and patient levels, to assess the relative magnitudes of within- and between-hospital variability in quality of care:

$$\begin{gathered} y_{ij} = \beta_0 + u_j + e_{ij}, \;\; i = 1, \ldots, n_j, \;\; j = 1, \ldots, J; \\ \sum_{j=1}^{J} n_j = N, \;\; (u_j | \sigma_u^2) \overset{\text{IID}}{\sim} N(0, \sigma_u^2), \;\; (e_{ij} | \sigma_e^2) \overset{\text{IID}}{\sim} N(0, \sigma_e^2). \end{gathered} \tag{30}$$

Browne and Draper [14] used a simulation study to show that, with a variety of maximum-likelihood-based methods for creating confidence intervals for σ_u^2, the actual coverage of nominal 95% intervals ranged from 72–94% across realistic sample sizes and true parameter values in the fields of education and medicine, versus 89–94% for Bayesian methods based on diffuse priors.

Their second example involved a re-analysis of a Guatemalan National Survey of Maternal and Child Health [89,97], with three-level data (births nested within mothers within communities), working with the random-effects logistic regression model

$$(y_{ijk} \mid p_{ijk}) \overset{\text{indep}}{\sim} \text{Bernoulli}(p_{ijk}) \quad \text{with}$$
$$\text{logit}(p_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + v_k\,, \tag{31}$$

where y_{ijk} is a binary indicator of modern prenatal care or not and where $u_{jk} \sim N(0, \sigma_u^2)$ and $v_k \sim N(0, \sigma_v^2)$ were random effects at the mother and community levels (respectively). Simulating data sets with 2 449 births by 1 558 women living in 161 communities (as in the Rodríguez and Goldman study [97]), Browne and Draper [14] showed that things can be even worse for likelihood-based methods in this model, with actual coverages (at nominal 95%) as low as 0–2% for intervals for σ_u^2 and σ_v^2, whereas Bayesian methods with diffuse priors again produced actual coverages from 89–96%. The technical problem is that the marginal likelihood functions for random-effects variances are often heavily skewed, with maxima at or near 0 even when the true variance is positive; Bayesian methods, which integrate over the likelihood function rather than maximizing it, can have (much) better small-sample calibration performance as a result.

Some Historical Perspective

The earliest published formal example of an attempt to do statistical inference – to reason backwards from effects to causes – seems to have been Bayes [5], who defined conditional probability for the first time and noted that the result we now call Bayes' Theorem was a trivial consequence of the definition. From the 1760s til the 1920s, all (or almost all) statistical inference was Bayesian, using the paradigm that Fisher and others referred to as *inverse probability*; prominent Bayesians of this period included Gauss [40], Laplace [64] and Pearson [88]. This Bayesian consensus changed with the publication of Fisher [37], which laid out a user-friendly program for maximum-likelihood estimation and inference in a wide variety of problems. Fisher railed against Bayesian inference; his principal objection was that in settings where little was known about a parameter (vector) θ external to the data, a number of prior distributions could be put forward to quantify this relative ignorance. He believed passionately in the late Victorian–Edwardian goal of scientific *objectivity*, and it bothered him greatly that two analysts with somewhat different diffuse priors might obtain somewhat different posteriors. (There is a Bayesian account of objectivity: a probability is objective if many different people more or less agree on its value. An example would be the probability of drawing a red ball from an urn known to contain 20 red and 80 white balls, if a sincere attempt is made to thoroughly mix the balls without looking at them and to draw the ball in a way that does not tend to favor one ball over another.)

There are two problems with Fisher's argument, which he never addressed:

1. He would be perfectly correct to raise this objection to Bayesian analysis if investigators were often forced to do inference based solely on prior information with no data, but in practice with even modest sample sizes the posterior is relatively insensitive to the precise manner in which diffuseness is specified in the prior, because the likelihood information in such situations is relatively so much stronger than the prior information; Sect. "Inference and Prediction: Binary Outcomes with No Covariates" provides an example of this phenomenon.
2. If Fisher had looked at the entire process of inference with an engineering eye to sensitivity and stability, he would have been forced to admit that uncertainty in how to specify the likelihood function has inferential consequences that are often an order of magnitude larger than those arising from uncertainty in how to specify the prior. It is an inescapable fact that subjectivity, through assumptions and judgments (such as the form of the likelihood function), is an integral part of any statistical analysis in problems of realistic complexity.

In spite of these unrebutted flaws in Fisher's objections to Bayesian inference, two schools of frequentist inference – one based on Fisher's maximum-likelihood estimation and *significance tests* [38], the other based on the confidence intervals and *hypothesis tests* of Neyman [83] and Neyman and Pearson [84] – came to dominate statistical practice from the 1920s at least through the 1980s. One major reason for this was practical: the Bayesian paradigm is based on integrating over the posterior distribution, and accurate approximations to high-dimensional integrals were not available during the period in question. Fisher's technology, based on differentiation (to find the maximum and curvature of the logarithm of the likelihood function) rather than integration, was a much more tractable approach for its time. Jeffreys [54], working in the field of astronomy, and Savage [102] and Lindley [69], building on de Finetti's results, advocated forcefully for the adoption of Bayesian methods, but prior to the advent of MCMC techniques (in the late 1980s) Bayesians were often in the position of saying that they knew the best way to solve statistical problems but the computations were beyond them. MCMC has removed this practical objection to the Bayesian paradigm for a wide class of problems.

The increased availability of affordable computers with decent CPU throughput in the 1980s also helped to overcome one objection raised in Sect. "Strengths and Weaknesses of the Two Approaches" against likelihood methods, that they can produce poorly-calibrated inferences with small samples, through the introduction of the bootstrap by Efron [31] in 1979. At this writing (a) both the frequentist and Bayesian paradigms are in vigorous inferential use, with the proportion of Bayesian articles in leading journals continuing an increase that began in the 1980s; (b) Bayesian MCMC analyses are often employed to produce meaningful predictive conclusions, with the use of the bootstrap increasing for frequentist predictive calibration; and (c) the Bayesian paradigm dominates decision analysis.

A Bayesian-Frequentist Fusion

During the 20th century the debate over which paradigm to use was often framed in such a way that it seemed it was necessary to choose one approach and defend it against attacks from people who had chosen the other, but there is nothing that forces an analyst to choose a single paradigm. Since both approaches have strengths and weaknesses, it seems worthwhile instead to seek a *fusion* of the two that makes best use of the strengths. Because (a) the Bayesian paradigm appears to be the most flexible way so far developed for quantifying all sources of uncertainty and (b) its main weakness is that coherence does not guarantee good calibration, a number of statisticians, including Rubin [98], Draper [26], and Little [73], have suggested a fusion in which inferences, predictions and decisions are formulated using Bayesian methods and then evaluated for their calibration properties using frequentist methods, for example by using Bayesian models to create 95% predictive intervals for observables not used in the modeling process and seeing if approximately 95% of these intervals include the actual observed values. Analysts more accustomed to the purely frequentist (likelihood) paradigm who prefer not to explicitly make use of prior distributions may still find it useful to reason in a Bayesian way, by integrating over the parameter uncertainty in their likelihood functions rather than maximizing over it, in order to enjoy the superior calibration properties that integration has been demonstrated to provide.

Future Directions

Since the mid- to late-1980s the Bayesian statistical paradigm has made significant advances in many fields of inquiry, including agriculture, archaeology, astronomy, bioinformatics, biology, economics, education, environmetrics, finance, health policy, and medicine (see Sect. "The Bayesian Statistical Paradigm" for recent citations of work in many of these disciplines). Three areas of methodological and theoretical research appear particularly promising for extending the useful scope of Bayesian work, as follows:

- *Elicitation of prior distributions and utility functions:* It is arguable that too much use is made in Bayesian analysis of diffuse prior distributions, because (a) accurate elicitation of non-diffuse priors is hard work and (b) lingering traces still remain of a desire to at least appear to achieve the unattainable Victorian–Edwardian goal of objectivity, the (false) argument being that the use of diffuse priors somehow equates to an absence of subjectivity (see, e. g., the papers by Berger [7] and Goldstein [45] and the ensuing discussion for a vigorous debate on this issue). It is also arguable that too much emphasis was placed in the 20th century on inference at the expense of decision-making, with inferential tools such as the Neyman–Pearson hypothesis testing machinery (Sect. "Some Historical Perspective") used incorrectly to make decisions for which they are not optimal; the main reason for this, as noted in Sect. "Strengths and Weaknesses of the Two Approaches" and "Some Historical Perspective", is that (a) the frequentist paradigm was dominant from the 1920s through the 1980s and (b) the high ground in decision theory is dominated by the Bayesian approach. Relevant citations of excellent recent work on elicitation of prior distributions and utility functions were given in Sect. "The Bayesian Statistical Paradigm"; it is natural to expect that there will be a greater emphasis on decision theory and non-diffuse prior modeling in the future, and elicitation in those fields of Bayesian methodology is an important area of continuing research.
- *Group decision-making:* As noted in Sect. "The Bayesian Statistical Paradigm", maximizing expected utility is an effective method for decision-making by a single agent, but when two or more agents are involved in the decision process this approach cannot be guaranteed to

yield a satisfying solution: there may be conflicts in the agents' preferences, particularly if their relationship is at least partly adversarial. With three or more possible actions, *transitivity* of preference – if You prefer action a_1 to a_2 and a_2 to a_3, then You should prefer a_1 to a_3 – is a criterion that any reasonable decision-making process should obey; informally, a well-known theorem by Arrow [3] states that even if all of the agents' utility functions obey transitivity, there is no way to combine their utility functions into a single decision-making process that is guaranteed to respect transitivity. However, Arrow's theorem is temporally static, in the sense that the agents do not share their utility functions with each other and iterate after doing so, and it assumes that all agents have the same set $\mathcal{A}$ of feasible actions. If agents A_1 and A_2 have action spaces $\mathcal{A}_1$ and $\mathcal{A}_2$ that are not identical and they share the details of their utility specification with each other, it is possible that A_1 may realize that one of the actions in $\mathcal{A}_2$ that (s)he had not considered is better than any of the actions in $\mathcal{A}_1$ or vice versa; thus a temporally dynamic solution to the problem posed by Arrow's theorem may be possible, even if A_1 and A_2 are partially adversarial. This is another important area for new research.

- *Bayesian computation:* Since the late 1980s, simulation-based computation based on Markov chain Monte Carlo (MCMC) methods has made useful Bayesian analyses possible in an increasingly broad range of application areas, and (as noted in Sect. "The Bayesian Statistical Paradigm") increases in computing speed and sophistication of MCMC algorithms have enhanced this trend significantly. However, if a regression-style data set is visualized as a matrix with n rows (one for each subject of inquiry) and k columns (one for each variable measured on the subjects), MCMC methods do not necessarily scale well in either n or k, with the result that they can be too slow to be of practical use with large data sets (e.g, at current desktop computing speeds, with n and/or k on the order of 10^5 or greater). Improving the scaling of MCMC methods, or finding a new approach to Bayesian computation that scales better, is thus a third important area for continuing study.

Bibliography

1. Abdellaoui M (2000) Parameter-free elicitation of utility and probability weighting functions. Manag Sci 46:1497–1512
2. Aleskerov F, Bouyssou D, Monjardet B (2007) Utility Maximization, Choice and Preference, 2nd edn. Springer, New York
3. Arrow KJ (1963) Social Choice and Individual Values, 2nd edn. Yale University Press, New Haven CT
4. Barlow RE, Wu AS (1981) Preposterior analysis of Bayes estimators of mean life. Biometrika 68:403–410
5. Bayes T (1764) An essay towards solving a problem in the doctrine of chances. Philos Trans Royal Soc Lond 53:370–418
6. Berger JO (1985) Statistical Decision Theory and Bayesian Analysis. Springer, New York
7. Berger JO (2006) The case for objective Bayesian analysis (with discussion). Bayesian Anal 1:385–472
8. Berger JO, Betro B, Moreno E, Pericchi LR, Ruggeri F, Salinetti G, Wasserman L (eds) (1995) Bayesian Robustness. Institute of Mathematical Statistics Lecture Notes-Monograph Series, vol 29. IMS, Hayward CA
9. Berger JO, Pericchi LR (2001) Objective Bayesian methods for model selection: introduction and comparison. In: Lahiri P (ed) Model Selection. Monograph Series, vol 38. Institute of Mathematical Statistics Lecture Notes Series, Beachwood, pp 135–207
10. Bernardo JM (1979) Reference posterior distributions for Bayesian inference (with discussion). J Royal Stat Soc, Series B 41:113–147
11. Bernardo JM, Smith AFM (1994) Bayesian Theory. Wiley, New York
12. Blavatskyy P (2006) Error propagation in the elicitation of utility and probability weighting functions. Theory Decis 60:315–334
13. Brown PJ, Vannucci M, Fearn T (1998) Multivariate Bayesian variable selection and prediction. J Royal Stat Soc, Series B 60:627–641
14. Browne WJ, Draper D (2006) A comparison of Bayesian and likelihood methods for fitting multilevel models (with discussion). Bayesian Anal 1:473–550
15. Buck C, Blackwell P (2008) Bayesian construction of radiocarbon calibration curves (with discussion). In: Case Studies in Bayesian Statistics, vol 9. Springer, New York
16. Chipman H, George EI, McCulloch RE (1998) Bayesian CART model search (with discussion). J Am Stat Assoc 93:935–960
17. Christiansen CL, Morris CN (1997) Hierarchical Poisson regression modeling. J Am Stat Assoc 92:618–632
18. Clyde M, George EI (2004) Model uncertainty. Stat Sci 19:81–94
19. Das S, Chen MH, Kim S, Warren N (2008) A Bayesian structural equations model for multilevel data with missing responses and missing covariates. Bayesian Anal 3:197–224
20. de Finetti B (1930) Funzione caratteristica di un fenomeno aleatorio. Mem R Accad Lincei 4:86–133
21. de Finetti B (1937). La prévision: ses lois logiques, ses sources subjectives. Ann Inst H Poincaré 7:1–68 (reprinted in translation as de Finetti B (1980) Foresight: its logical laws, its subjective sources. In: Kyburg HE, Smokler HE (eds) Studies in Subjective Probability. Dover, New York, pp 93–158)
22. de Finetti B (1938/1980). Sur la condition d'équivalence partielle. Actual Sci Ind 739 (reprinted in translation as de Finetti B (1980) On the condition of partial exchangeability. In: Jeffrey R (ed) Studies in Inductive Logic and Probability. University of California Press, Berkeley, pp 193–206)
23. de Finetti B (1970) Teoria delle Probabilità, vol 1 and 2. Eunaudi, Torino (reprinted in translation as de Finetti B (1974–75) Theory of probability, vol 1 and 2. Wiley, Chichester)
24. Dey D, Müller P, Sinha D (eds) (1998) Practical Nonparametric and Semiparametric Bayesian Statistics. Springer, New York

25. Draper D (1995) Assessment and propagation of model uncertainty (with discussion). J Royal Stat Soc, Series B 57:45–97
26. Draper D (1999) Model uncertainty yes, discrete model averaging maybe. Comment on: Hoeting JA, Madigan D, Raftery AE, Volinsky CT (eds) Bayesian model averaging: a tutorial. Stat Sci 14:405–409
27. Draper D (2007) Bayesian multilevel analysis and MCMC. In: de Leeuw J, Meijer E (eds) Handbook of Multilevel Analysis. Springer, New York, pp 31–94
28. Draper D, Hodges J, Mallows C, Pregibon D (1993) Exchangeability and data analysis (with discussion). J Royal Stat Soc, Series A 156:9–37
29. Draper D, Krnjajić M (2008) Bayesian model specification. Submitted
30. Duran BS, Booker JM (1988) A Bayes sensitivity analysis when using the Beta distribution as a prior. IEEE Trans Reliab 37:239–247
31. Efron B (1979) Bootstrap methods. Ann Stat 7:1–26
32. Ferguson T (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230
33. Ferreira MAR, Lee HKH (2007) Multiscale Modeling. Springer, New York
34. Fienberg SE (2006) When did Bayesian inference become "Bayesian"? Bayesian Anal 1:1–40
35. Fishburn PC (1970) Utility Theory for Decision Making. Wiley, New York
36. Fishburn PC (1981) Subjective expected utility: a review of normative theories. Theory Decis 13:139–199
37. Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans Royal Soc Lond, Series A 222:309–368
38. Fisher RA (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh
39. Fouskakis D, Draper D (2008) Comparing stochastic optimization methods for variable selection in binary outcome prediction, with application to health policy. J Am Stat Assoc, forthcoming
40. Gauss CF (1809) Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium, vol 2. Perthes and Besser, Hamburg
41. Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409
42. Gelman A, Meng X-L (2004) Applied Bayesian Modeling and Causal Inference From Incomplete-Data Perspectives. Wiley, New York
43. George EI, Foster DP (2000) Calibration and empirical Bayes variable selection. Biometrika 87:731–747
44. Gilks WR, Richardson S, Spiegelhalter DJ (eds) (1996) Markov Chain Monte Carlo in Practice. Chapman, New York
45. Goldstein M (2006) Subjective Bayesian analysis: principles and practice (with discussion). Bayesian Anal 1:385–472
46. Good IJ (1950) Probability and the Weighing of Evidence. Charles Griffin, London
47. Green P (1995) Reversible jump Markov chain Monte carlo computation and Bayesian model determination. Biometrika 82:711–713
48. Hacking I (1984) The Emergence of Probability. University Press, Cambridge
49. Hanson TE, Kottas A, Branscum AJ (2008) Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches. J Royal Stat Soc, Series C (Applied Statistics) 57:207–226
50. Hellwig K, Speckbacher G, Weniges P (2000) Utility maximization under capital growth constraints. J Math Econ 33:1–12
51. Hendriksen C, Lund E, Stromgard E (1984) Consequences of assessment and intervention among elderly people: a three year randomized controlled trial. Br Med J 289:1522–1524
52. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14:382–417
53. Jaynes ET (2003) Probability Theory: The Logic of Science. Cambridge University Press, Cambridge
54. Jeffreys H (1931) Scientific Inference. Cambridge University Press, Cambridge
55. Jordan MI, Ghahramani Z, Jaakkola TS, Saul L (1999) An introduction to variational methods for graphical models. Mach Learn 37:183–233
56. Kadane JB (ed) (1996) Bayesian Methods and Ethics in a Clinical Trial Design. Wiley, New York
57. Kadane JB, Dickey JM (1980) Bayesian decision theory and the simplification of models. In: Kmenta J, Ramsey J (eds) Evaluation of Econometric Models. Academic Press, New York
58. Kahn K, Rubenstein L, Draper D, Kosecoff J, Rogers W, Keeler E, Brook R (1990) The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series (with discussion). J Am Med Assoc 264:1953–1955
59. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795
60. Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. J Am Stat Assoc 91:1343–1370
61. Key J, Pericchi LR, Smith AFM (1999) Bayesian model choice: what and why? (with discussion). In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian Statistics 6. Clarendon Press, Oxford, pp 343–370
62. Krnjajić M, Kottas A, Draper D (2008) Parametric and nonparametric Bayesian model specification: a case study involving models for count data. Comput Stat Data Anal 52: 2110–2128
63. Laplace PS (1774) Mémoire sur la probabilité des causes par les évenements. Mém Acad Sci Paris 6:621–656
64. Laplace PS (1812) Théorie Analytique des Probabilités. Courcier, Paris
65. Laud PW, Ibrahim JG (1995) Predictive model selection. J Royal Stat Soc, Series B 57:247–262
66. Lavine M (1992) Some aspects of Pólya tree distributions for statistical modelling. Ann Stat 20:1222–1235
67. Leamer EE (1978) Specification searches: Ad hoc inference with non-experimental data. Wiley, New York
68. Leonard T, Hsu JSJ (1999) Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers. Cambridge University Press, Cambridge
69. Lindley DV (1965) Introduction to Probability and Statistics. Cambridge University Press, Cambridge
70. Lindley DV (1968) The choice of variables in multiple regression (with discussion). J Royal Stat Soc, Series B 30:31–66
71. Lindley DV (2006) Understanding Uncertainty. Wiley, New York
72. Lindley DV, Novick MR (1981) The role of exchangeability in inference. Ann Stat 9:45–58
73. Little RJA (2006) Calibrated Bayes: A Bayes/frequentist roadmap. Am Stat 60:213–223

74. Mangel M, Munch SB (2003) Opportunities for Bayesian analysis in the search for sustainable fisheries. ISBA Bulletin 10:3–5
75. McCullagh P, Nelder JA (1989) Generalized Linear Models, 2nd edn. Chapman, New York
76. Meng XL, van Dyk DA (1997) The EM Algorithm: an old folk song sung to a fast new tune (with discussion). J Royal Stat Soc, Series B 59:511–567
77. Merl D, Prado R (2007) Detecting selection in DNA sequences: Bayesian modelling and inference (with discussion). In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian Statistics 8. University Press, Oxford, pp 1–22
78. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machine. J Chem Phys 21:1087–1091
79. Morris CN, Hill J (2000) The Health Insurance Experiment: design using the Finite Selection Model. In: Morton SC, Rolph JE (eds) Public Policy and Statistics: Case Studies from RAND. Springer, New York, pp 29–53
80. Müller P, Quintana F (2004) Nonparametric Bayesian data analysis. Stat Sci 19:95–110
81. Müller P, Quintana F, Rosner G (2004) Hierarchical meta-analysis over related non-parametric Bayesian models. J Royal Stat Soc, Series B 66:735–749
82. Munch SB, Kottas A, Mangel M (2005) Bayesian nonparametric analysis of stock-recruitment relationships. Can J Fish Aquat Sci 62:1808–1821
83. Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Philos Trans Royal Soc Lond A 236:333–380
84. Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20:175–240
85. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain Judgements. Eliciting Experts' Probabilities. Wiley, New York
86. O'Hagan A, Forster J (2004) Bayesian Inference, 2nd edn. In: Kendall's Advanced Theory of Statistics, vol 2B. Arnold, London
87. Parmigiani G (2002) Modeling in medical decision-making: A Bayesian approach. Wiley, New York
88. Pearson KP (1895) Mathematical contributions to the theory of evolution, II. Skew variation in homogeneous material. Proc Royal Soc Lond 57:257–260
89. Pebley AR, Goldman N (1992) Family, community, ethnic identity, and the use of formal health care services in Guatemala. Working Paper 92-12. Office of Population Research, Princeton
90. Pettit LI (1990) The conditional predictive ordinate for the Normal distribution. J Royal Stat Soc, Series B 52:175–184
91. Pérez JM, Berger JO (2002) Expected posterior prior distributions for model selection. Biometrika 89:491–512
92. Polson NG, Stroud JR, Müller P (2008) Practical filtering with sequential parameter learning. J Royal Stat Soc, Series B 70:413–428
93. Rashbash J, Steele F, Browne WJ, Prosser B (2005) A User's Guide to `MLwiN`, Version 2.0. Centre for Multilevel Modelling, University of Bristol, Bristol UK; available at www.cmm.bristol.ac.uk Accessed 15 Aug 2008
94. Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). J Royal Stat Soc, Series B 59:731–792
95. Rios Insua D, Ruggeri F (eds) (2000) Robust Bayesian Analysis. Springer, New York
96. Rodríguez A, Dunston DB, Gelfand AE (2008) The nested Dirichlet process. J Am Stat Assoc, 103, forthcoming
97. Rodríguez G, Goldman N (1995) An assessment of estimation procedures for multilevel models with binary responses. J Royal Stat Soc, Series A 158:73–89
98. Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Stat 12:1151–1172
99. Rubin DB (2005) Bayesian inference for causal effects. In: Rao CR, Dey DK (eds) Handbook of Statistics: Bayesian Thinking, Modeling and Computation, vol 25. Elsevier, Amsterdam, pp 1–16
100. Sabatti C, Lange K (2008) Bayesian Gaussian mixture models for high-density genotyping arrays. J Am Stat Assoc 103:89–100
101. Sansó B, Forest CE, Zantedeschi D (2008) Inferring climate system properties using a computer model (with discussion). Bayesian Anal 3:1–62
102. Savage LJ (1954) The Foundations of Statistics. Wiley, New York
103. Schervish MJ, Seidenfeld T, Kadane JB (1990) State-dependent utilities. J Am Stat Assoc 85:840–847
104. Seidou O, Asselin JJ, Ouarda TMBJ (2007) Bayesian multivariate linear regression with application to change point models in hydrometeorological variables. In: Water Resources Research 43, W08401, doi:10.1029/2005WR004835.
105. Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Wiley, New York
106. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J Royal Stat Soc, Series B 64:583–640
107. Spiegelhalter DJ, Thomas A, Best NG (1999) `WinBUGS` Version 1.2 User Manual. MRC Biostatistics Unit, Cambridge
108. Stephens M (2000) Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible-jump methods. Ann Stat 28:40–74
109. Stigler SM (1986) The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge
110. Stone M (1974) Cross-validation choice and assessment of statistical predictions (with discussion). J Royal Stat Soc, Series B 36:111–147
111. Wald A (1950) Statistical Decision Functions. Wiley, New York
112. Weerahandi S, Zidek JV (1981) Multi-Bayesian statistical decision theory. J Royal Stat Soc, Series A 144:85–93
113. Weisberg S (2005) Applied Linear Regression, 3rd edn. Wiley, New York
114. West M (2003) Bayesian factor regression models in the "large p, small n paradigm." Bayesian Statistics 7:723–732
115. West M, Harrison PJ (1997) Bayesian Forecasting and Dynamic Models. Springer, New York
116. Whitehead J (2006) Using Bayesian decision theory in dose-escalation studies. In: Chevret S (ed) Statistical Methods for Dose-Finding Experiments. Wiley, New York, pp 149–171