Bayesian Methods in Risk Assessment

Article · January 2005

1 author:

Scott Ferson
University of Liverpool

Bayesian methods in risk assessment

Prepared for

Dominique GUYONNET
Head of Waste & Storage Unit
Service Environnement & Procédés
Bureau de Recherches Géologiques et Minières (BRGM)
3, avenue Claude GUILLEMIN
BP 6009
45060 Orléans Cedex 2
FRANCE
+33(0)2.38.64.38.17, fax +33(0)2.38.64.30.62
d.guyonnet@brgm.fr

by

Scott Ferson
Applied Biomathematics
100 North Country Road
Setauket, New York 11733 USA
scott@ramas.com
1-631-751-4350, fax 1-631-751-3435

Table of contents

Executive summary

Empirical data are almost always lacking in real-world risk analyses. In fact, some risk analysis problems try to forecast what risks may be associated with situations that are, at the time of the assessment, only hypothetical. It may therefore be impractical, unethical, or even impossible to collect relevant empirical data. To make matters worse for the analyst, the situations of concern in risk analyses are often novel and have never been studied before. This means that scientific understanding of the underlying processes may itself be in doubt. Because of these facts, computational problems in risk analysis are often characterized by three issues:

(i) there may be little or even no empirical data available for some variables,
(ii) it may be necessary to employ subjective information from the analyst's judgment or expert opinion, and
(iii) uncertainty about the mathematical model used in the assessment may be substantial.

These issues complicate and impede the assessment, and they can call into question any conclusions or inferences drawn from the assessment.

A Bayesian approach might be useful in addressing these issues. By design, Bayesian methods natively consider the uncertainty associated with the parameters of a probability model (even if those uncertain parameters are believed to be fixed numbers). Bayesian methods are often recommended as the proper way to make formal use of subjective information such as expert opinion and personal judgments or beliefs of an analyst. An important advantage of Bayesian methods, unlike frequentist methods with which they are often contrasted, is that they can always yield a precise answer, even when no data at all are available. Finally, recent Bayesian literature has focused on the potential significance of model uncertainty and how it can be incorporated into quantitative analyses.

This report reviews the technical and interpretational limitations of using Bayesian methods in risk and uncertainty analyses. It argues that the methods produce sometimes extreme overconfidence and arbitrariness in the computed answers. These deficiencies are illustrated using various computational problems including projecting uncertainty through predictive expressions, analyzing sample data, updating estimates based on new information, and accounting for model uncertainty. There are three main causes of this overconfidence and arbitrariness: (1) misuse of equiprobability as a model for incertitude, (2) overuse of averaging to aggregate information, which tends to erase variation rather than propagate it, and (3) reliance on precise values and particular distributions when available information does not justify such specificity.

Frequentists use two-dimensional Monte Carlo (2MC) simulation to account for uncertainty associated with the parameters of a probability model that Bayesian methods handle natively. Because of their mathematical simplicity, many risk analysts now routinely use and recommend 2MC simulation as a convenient and natural approach to account for incertitude—as distinguished from variability—in an assessment. It is clear, however, that 2MC has many of the same disadvantages as Bayesian methods, and harbors some problems of its own as well.

The methods most commonly employed by Bayesians and users of 2MC often rely on untenable or dubious assumptions. Robust methods are needed that can produce reliable projections even without precise and complete statistical information. Such methods produce bounding answers which seem more appropriate for the risk analysis problems considered.

Introduction

There are three ways that Bayesian methods might be employed in risk analysis for scientific or regulatory purposes. The first way is to take over the assessment and decision process entirely. The Bayesian approach would be used to **frame the scientific or regulatory decision problems**. The decision problems faced in risk analysis include questions like

- Should the manufacture of a dangerous product be banned or limited?
- Should we clean up a site of environmental contamination?
- How much should it be cleaned up? and
- Should the *Cassini* spacecraft be allowed to fly?

Bayesians would argue that their approach is a comprehensive scheme for inference and decision making, and that the only way to guarantee rational and coherent decisions to the problems is by expressing and solving them within a Bayesian framework. This use of Bayesian methods essentially puts the Bayesian analyst in charge of the entire endeavor. Some decision makers might prefer to use a formal infrastructure to arrive at decisions, but such uses have apparently been rare among prominent risk assessments.

On the other hand, Bayesian methods could be used merely to **estimate risk distributions**. This use of Bayesian methods puts the Bayesian analysts in the center of the issue, although not in charge of the entire process. Bayesian methods could be used to answer questions like

- What are the probabilities that contaminant doses will be large?
- What is the uncertainty about the mean radiation exposure?
- How likely is the stress to exceed a critical value?
- How long is an endangered species likely to persist?

The distributions and quantities that result from the analyses would be the primary fodder for the decision making, but the decision process itself would be insulated from the Bayesian analysis. Some decision makers might think this preferable if the decision involves horse trading or sensitive political implications.

Finally, Bayesian methods could be used as a tool to **select or parameterize input distributions** for a risk model. This use relegates Bayesian analysts to the roles of technicians and support analysts, because the form of the risk model and the overarching decision process are developed without appeal to Bayesian methods. In practice, it appears that Bayesian methods are invading risk analysis in a gradual way from this technical level first. In such uses, the methods are applied to estimation problems too, but in these cases one is estimating the inputs to models rather than estimating the answers directly. The questions that Bayesian methods might answer would be similar to

- What is the distribution for contaminant concentrations a receptor will experience?
- How much Perrier do people drink per day?
- What is the uncertainty about the probability of a critical event? and
- How can we characterize uncertainty about the mean of the body mass distribution?

Bayesian methods certainly provide an important way to select inputs for a risk analysis. Many would argue that they are the only formal way to make use of subjective information such as the analyst's judgments or experts' opinions in fashioning inputs. Because data are commonly so scarce in risk analysis that there may not be any other kind of information, the proper handling of subjective information can be critical.

This document will review the advantages and disadvantages of Bayesian methods, focusing mostly on the second and third ways they can be applied in risk analysis.

The subjectivist versus frequentist debate

Probability theory was invented three and a half centuries ago in a long series of letters between French mathematicians Blaise Pascal and Pierre de Fermat as a means to settle questions involving gambling. The first flowering of the discipline occurred two hundred years ago, when Pierre-Simon de Laplace (1820) realized that the methods of inference developed for questions about the possible outcomes of throwing dice and drawing playing cards could also be used to answer other kinds of questions. This realization allowed probability theory to accumulate a great variety of applications in many areas beyond gambling, particularly in mathematics, science and engineering. From the earliest correspondence between Pascal and Fermat, and essentially in the turn made by Laplace, there has been an element of controversy about just what 'probability' should mean. The crucial question is whether probability should measure something with a *physical reality*, or something that reflects a *person's beliefs* about the physical world.

The debate about the meaning of probability has continued to the present day. Neapolitan (1992) reviewed the use of three* different interpretations of probability to address three kinds of questions. The classical theory of probability applies to situations in which all possible outcomes are equally likely. It is used to answer questions like "What is the probability that two dice will land showing at least eight dots?" Everyone accepts this interpretation of probability and agrees that it is appropriate to answer such questions. But the success of the theory in handling situations in which uncertainty plays a central role tantalizes people who, naturally, want to extend this capacity. Two schools of thought emerge about how this can be done. The first school is composed of the frequentists, who maintain that probability can be applied to all repeatable, random phenomena, even if the possible outcomes are not equally likely. The frequentist theory of probability can be used to answer more complicated questions such as "What is the probability that we'd see farm yields like those observed if the fertilizer we used had no effect?"

The second school of thought is composed of the subjectivists who maintain that probability can be applied to any event about which one is uncertain. The subjectivist theory of probability can be applied to questions like "What is the probability that the fertilizer has no effect?" Note that this question is actually closely akin to questions like "Is there a God?" and "Is OJ guilty?" in that the fact of the matter is, presumably, either true or false. That is, there is no random process in which there is sometimes a God and sometimes not. Frequentists think such questions are meaningless, or at least that they cannot be addressed by probability theory because there is no underlying random process. God either exists or doesn't exist; the fertilizer either has an effect or it doesn't. By asking about the probability that the fertilizer has an effect (when it either does or doesn't), subjectivists are not asking a question about the real world, but rather one about their own brains. They are really asking "How much should I believe that the fertilizer has an effect?" The probabilist de Finetti (1970, page x) famously proclaimed "probability does not exist" by which he meant that, because probability measures a rational agent's degree of belief, there is no single, correct probability. Naturally, frequentists question the relevance of such questions, but subjectivists counter that the tortured way frequentists must approach simple questions (like whether the fertilizer is any good) make their approach unworkable. Subjectivists feel there is no compelling reason to limit the powerful methods of probability to asking about random events, subjectivists can ask about *any* event.

---

*Other possible interpretations of probability have also been proposed (see Hájek 2003).

In statistics, a frequentist draws inferences about data given an unknown parameter, while a subjectivist updates beliefs about uncertain parameters given new data. It happens that a frequentist analysis of a data set often agrees in large part with a parallel analysis based on a subjectivist interpretation of probability (Lindley 1965). Such situations appear to conform with the old joke that the reason academic arguments are so intense is because there is so little at stake. One might certainly regard the debate between the two schools as purely philosophical, and therefore of no real consequence to engineers or any practically minded person, if the numerical results of the analyses based on the different interpretations do not differ substantially. However, there can be numerical differences—and sometimes big ones—between the two kinds of analyses. The differences tend to be greatest when data are the scarcest, which is of course the typical case in risk assessment problems. For this reason, it seems that modern-day risk analysts do have to enter the fray of this debate and consider which school they will subscribe to or how they will try to bridge between the two of them. However, given its perennial nature, we shouldn't expect to completely resolve the dispute, even for a highly circumscribed domain such as risk analysis.

Bayesians are like snowflakes

Despite assertions that the Bayesian approach provides a unified theory for inferences and decision making, in fact there is a great diversity among Bayesian practitioners. Like proverbial snowflakes, it seems that each Bayesian is a unique assemblage of opinions about the proper methodology for statistics. For instance, most Bayesians regard all statistical problems as part of decision theory and consider the decision context to be an essential element of a properly framed problem. But not all Bayesians hold to this opinion. Although most Bayesians subscribe to a subjectivist interpretation of probability and affirm the fundamental importance of subjectivity within a general scheme for inference, there are some "objective Bayesians" who reject the subjective interpretation of probability. Most Bayesians feel obliged to reduce the distributions of parameters to their means to obtain single values, but many Bayesians have no qualms about reporting entire distributions to characterize the uncertainty about parameters. Most Bayesians believe that a precise probability model can and should be elicited, but practitioners of robust Bayes methods (Berger 1985) permit the use of classes of probabilities rather than insisting on a particular probability model. Perhaps most surprisingly, not all Bayesians regularly employ Bayes' rule to update priors with incoming data. Some analysts (e.g., Vose 2000, page 150) prefer to re-evaluate the prior whenever new data becomes available, and tend to defer the application of Bayes rule to the indefinite future. Because of this wide variation in practice, comments and criticisms about Bayesian methods must to a large degree follow a moving target. Indeed, some schools of Bayesians would agree with certain of the criticisms spelled out in this report, and they would offer their own approach as a fix for the problem or limitation. The review below intends to describe standard Bayesian methods (if such exist) as they might be used in risk analysis problems.

The next two sections introduce Bayes' rule for events and distributions. They include seven numerical examples that illustrate how a Bayesian approach would work in risk analysis. The next section briefly reviews the advantages of the Bayesian approach. The section after that reviews the disadvantages, using the numerical examples as touchstones to argue the Bayesian solutions are inadequate or overconfident in particular ways.

Derivation of Bayes' rule

For two events A and B, having probabilities $P(A)$ and $P(B) \neq 0$ respectively, the definition of the conditional probability of A given B is

$$P(A \mid B) = P(A \,\&\, B) / P(B)$$

where $A \,\&\, B$ denotes the event where both A and B occur and $P(A \,\&\, B)$ denotes that event's probability. The quantity $P(A \mid B)$ describes* the chance that event A will occur, given the fact that B has already occurred.

Consider the Venn diagram below. So long as neither A nor B is impossible, i.e., if $P(A) \neq 0$ and $P(B) \neq 0$, then symmetry implies that the probability of the conjunction $A \,\&\, B$ is equal to two things,

$$P(A \mid B)\, P(B) = P(A \,\&\, B) = P(B \mid A)\, P(A),$$

which must therefore equal each other. Rearranging a little bit, we see immediately that

$$P(A \mid B) = P(A)\, P(B \mid A) / P(B)$$

which is Bayes' rule. It is just a way of converting a probability like $P(B \mid A)$ into one like $P(A \mid B)$, that is, the probability that B occurs given A has occurred to the probability that A occurs given B. This means, for instance, that the probability that you have a disease given that you've tested positive for it can be computed from the probability that you'd test positive for the disease if you had it. The confusion of these two quantities is not only common among beginning students of probability theory, but has found its way into the publications by some of the greatest names in its history, including de Morgan**, Boole, and even Laplace himself. Bayes' rule converts from $P(B \mid A)$ to $P(A \mid B)$ simply by multiplying by $P(A)$ and dividing by $P(B)$.

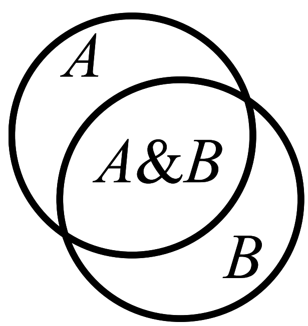


Figure 1. Venn diagram of two events.

---

*Jeffreys introduced the notation using the vertical line, but it was popularized by Feller. Like many things in probability theory, it is somewhat illogical. One expects the argument of $P(\cdot)$ to be an event as A, B or $A\&B$ are, but $A \mid B$ is not an event, i.e., it is not a subset of the sample space.

**Rice (2004) argued that de Morgan, who wrote the first treatise on probability theory in the English language, completely *abandoned* the field of probability because of its difficulty and because such errors are so easy to make. Peirce acknowledged the intrinsic and unique difficulty of probability among mathematical disciplines (quoted in Hamming 1991, page 41).

In this formulation, $\mathrm{P}(A)$ is called the prior and $\mathrm{P}(B \mid A)$ is called the likelihood. $\mathrm{P}(B)$ is called the normalization factor. It is often computed using the law of total probability as

$$\mathrm{P}(B) = \sum_i \mathrm{P}(B \mid C_i) P(C_i)$$

where C_i is some partition of the sample space. For instance, it might be computed as

$$\mathrm{P}(B) = \mathrm{P}(B \mid A)\ \mathrm{P}(A) + \mathrm{P}(B \mid \text{not } A)\ \mathrm{P}(\text{not } A) = \mathrm{P}(B \mid A)\ \mathrm{P}(A) + (1-\mathrm{P}(\text{not } B \mid \text{not } A))\ (1-\mathrm{P}(A)).$$

The calculation of the normalization factor is often the most computationally demanding aspect of the use of Bayes' rule.

It is important to understand that there is nothing whatever controversial about Bayes' rule in itself. It is a theorem based on the axioms of probability theory. Probabilists of all stripes consider it valid and routinely use it their arguments and calculations. The controversy around Bayes' rule arises, not from the rule itself, but from the applications that have sometimes been made of it. In the next and following sections, we shall review some of these applications.

Numerical example: disease testing

Suppose that the prevalence of a disease in the general population is 1 in 10,000 people (0.01%). Suppose that a test for the disease has been developed which has excellent characteristics. In particular, the sensitivity of the test, which is the probability that the test is positive when it is applied to a person with the disease, is 99.9%. The specificity of the test, which is the probability that the test is negative if the person does not have the disease, is 99.99%. If a person from the population with no other known risk factors tests positive for the disease, what is the chance that the person actually has the disease? When the question is asked this way, almost everyone, including medical professionals, think the answer is some value over 99%, even though this answer is completely wrong (Casscells et al. 1978; Cosmides and Tooby 1996; Gigerenzer 1991; 2002).

Bayes' rule allows us to compute the probability P(disease | positive) that the person has the disease given a positive test as the ratio

$$\mathrm{P}(\text{disease} \mid \text{positive}) = \frac{\mathrm{P}(\text{disease}) \times \mathrm{P}(\text{positive} \mid \text{disease})}{\mathrm{P}(\text{positive})}$$

where P(positive) is computed as the sum

$$\mathrm{P}(\text{disease}) \times \mathrm{P}(\text{positive} \mid \text{disease}) + (1 - \mathrm{P}(\text{disease})) \times (1 - \mathrm{P}(\text{negative} \mid \text{healthy})).$$

We use this formula by plugging in the numbers

prevalence = 0.01% = P(disease),
sensitivity = 99.9% = P(positive | disease), and
specificity = 99.99% = P(negative | healthy),

and computing

$$\text{prevalence} \times \text{sensitivity} / (\text{prevalence} \times \text{sensitivity} + (1 - \text{prevalence}) \times (1 - \text{specificity}))$$
$$= 0.01\% \times 99.9\% / (0.01\% \times 99.9\% + (1 - 0.01\%) \times (1 - 99.99\%))$$

which yields 0.4998. Thus, the probability that the person who tested positive for the disease actually has it is only about 50%. To see why this is so, imagine a population of 10,000 people without any specific risk factors. The prevalence of 0.01% suggests that about one person from this group has the disease. The sensitivity of the test (99.9%) suggests that this person will test positive almost certainly. The specificity of 99.99% suggests that, of the 9,999 people who do not have the disease, another will also test positive. Thus, we'd expect roughly two people to test positive for the disease, but only one of them actually has it.

Numerical example: paternity

Berry (1997) offers an example of using Bayes' rule to compute the probability that a man is the father of a girl given the results of (non-DNA) blood tests on the individuals involved. The girl has type B blood and her mother has type O. Genetics of the ABO allele system tell us that the girl could not have inherited the B allele from her mother. She must have inherited it from her father. The alleged father has type AB and, therefore, could have been the donor of the B allele. To apply Bayes' rule to compute the probability of the man's paternity given the blood results, we need

P(B | F)	probability the girl would inherit the B allele if the man were her father,
P(B | not F)	probability she could inherit the B allele from another man, and
P(F)	prior probability the man is the father of the girl.

Because the man's blood type is AB, he will contribute either an A or a B allele to his progeny, with equal frequency, so basic Mendelian genetics tells us that P(B | F) = ½. The frequency of the B allele in the general population is about 0.09, so that means P(B | not F) = 0.09. If we don't know what the prior should be, we might set P(F) = ½. Applying Bayes' rule, we compute

$$P(F \mid B) = \frac{P(F) \times P(B \mid F)}{P(F) \times P(B \mid F) + (1 - P(F)) \times P(B \mid \text{not } F)}$$

$$= \frac{\tfrac{1}{2} \times \tfrac{1}{2}}{\tfrac{1}{2} \times \tfrac{1}{2} + (1 - \tfrac{1}{2}) \times 0.09},$$

so P(F | B) is about 0.847. If we were unsure about the paternity before seeing the blood results, we ought to be somewhat more sure afterward. Certainly, this would seem to be a relevant argument for civil litigation alleging paternity.

Numerical example: waiting for a bus

A rather less biological example concerns the problem of whether you should continue to wait for an already late bus, or give up on it and start walking. Suppose that you arrived at the bus stop in plenty of time before its scheduled stop and have been waiting so that, now, it is 10 minutes late. Buses sometimes don't show up at all, and the next scheduled bus is two hours away. To make your appointment, you might have to walk. You're told that 90% of buses observe their rounds, and that 80% of those that do are no more than 10 minutes late. This means that there's a 10% chance that bus won't show up at all, and, if it does, it might be more than 10 minutes late. Given that it's already 10 minutes late, what is the probability the bus will come, or, to phrase it as a risk analysis question, what is the probability that the bus won't come and you'll need to walk?

Consider two events B and W. Event B is that the bus does come, and event W is that it never does and you have to walk to your appointment. The prior probabilities are

$$P(B) = 90\%,$$
$$P(W) = 10\%.$$

The likelihoods of the observation that the bus is 10 minutes late given these two events are

$$P(10 \text{ min} \mid B) = 20\%,$$
$$P(10 \text{ min} \mid W) = 100\%.$$

Because 80% of buses that show up are no more than 10 minutes late, our assuming that the bus will eventually come would mean that it would be one of the 20% that are even later. The second likelihood comes from the fact that, if the bus isn't coming at all, then we're sure it will be 10 minutes late. Applying Bayes' rule, we can see the probability that the bus will come given the observation that it's already 10 minutes late is

$$\begin{aligned} P(B \mid 10 \text{ min}) &= P(B) \times P(10 \text{ min} \mid B) / (P(B) \times P(10 \text{ min} \mid B) + P(W) \times P(10 \text{ min} \mid W)) \\ &= 90\% \times 20\% / (90\% \times 20\% + 10\% \times 100\%) \\ &= 18 / 28 \\ &\approx 0.643. \end{aligned}$$

The probability the bus won't come given the observation is

$$\begin{aligned} P(W \mid 10 \text{ min}) &= P(W) \times P(10 \text{ min} \mid W) / (P(B) \times P(10 \text{ min} \mid B) + P(W) \times P(10 \text{ min} \mid W)) \\ &= 10\% \times 100\% / (90\% \times 20\% + 10\% \times 100\%) \\ &= 10 / 28 \\ &\approx 0.357. \end{aligned}$$

So, despite your growing frustration with the bus company, it looks like it's still more likely that the bus will come than not.

Bayes' rule for distributions

Bayes' rule for events can be extended to define a Bayes' rule for random variables and their distribution functions in a straightforward way. It can be used to combine a prior distribution and a likelihood function to produce a posterior distribution. The posterior distribution might then be used as an input in a risk analysis. We can write Bayes' rule as

$$P(\theta \mid E) = P(\theta)\, P(E \mid \theta) / P(E)$$

where P denotes probability mass (or density), θ is a value of the random variable in question, and E denotes the evidence being considered. $P(\theta)$ is the prior probability that the random variable takes on the value θ. Because $P(\theta)$ is a distribution of θ, its integral with respect to θ is one. P(E | θ) is the conditional likelihood function that expresses the probability of the evidence given a particular value of θ. The likelihood function is the probability of observing a value if the value were actually θ, interpreted as a function of θ. It is not a distribution, so its integral need not be one. The normalizing factor P(E) corresponds to the probability of having obtained the observed evidence. This divisor is the sum (or integral) with respect to θ of the product of the prior and the likelihood. Bayes' rule is applied for all values of θ to obtain P(θ | E), which is the posterior distribution of θ given the evidence. Both the prior and the likelihood are functions of θ, and Bayes' rule for distributions is essentially their product for each possible value of θ. The normalizing factor is not a function of θ, but has a single value such that the resulting posterior distribution integrates to unity.

For most Bayesians, the prior distribution is derived from the opinions or beliefs of the analyst. It is intended to represent, at least initially, the analyst's subjective knowledge before any specific evidence is considered. Thus, it may be the result of amorphous preconceptions, mechanistic reasoning, hearsay, or some combination of these things. The likelihood function represents a model, also perhaps taken from the subjective knowledge of the analyst, of what data implies about the variable in question. The normalizing factor is often difficult to compute analytically, but the use of conjugate pairs (see page 29) may greatly simplify the problem. When these computational shortcuts can't be used, recently developed software programs may be able to solve the problem using computer-intensive methods.

Although statistics and probabilistic modeling in general are focused primarily on assessing or accounting for uncertainty about parameters, there may also be substantial uncertainty about the *model* in which the parameters are bundled. In analytical parlance, a 'model' consists of those aspects of the description of a problem or calculation which are regarded as given and not at present in question (Edwards 1972, page 3). This is not to suggest, however, that a model is forever beyond doubt. Indeed, the correctness of a model in its choices of distributions and structural assumptions may sometimes come to be the most dubious element of an analysis. Bayesian model averaging (Draper 1995; Raftery et al. 1997; Hoeting et al. 1999; Clyde and George 2003) is used to characterize model uncertainty so it can be incorporated into a risk analysis. Until very recently, analysts chose a single model and then acted as though it had generated the data. Bayesian model averaging recognizes that conditioning on a single selected model ignores model uncertainty, and therefore can lead to underestimation of uncertainty in forecasts. The Bayesian strategy to overcome the problem involves averaging over all possible models when making inferences about quantities of interest. Draper (1995) suggested employing standard techniques of data analysis, but when a good model is found, embedding it in a richer family of models. By assigning prior probabilities for the parameters of this family of model and treating model selection like other Bayesian parameter estimation problems, this approach produces a weighted average of the predictive distributions from each model, where the weights are given by the posterior probabilities for each model. By averaging over many different competing models, this approach incorporates model uncertainty into conclusions about parameters and predictions. In practice, however, this approach is often not computationally feasible because it can be difficult to enumerate all possible models for problems with a large number of variables. However, a variety of methods for implementing the approach for specific kinds of statistical models have been developed. The approach has been applied to many classes of statistical models including several kinds of regression models (Hoeting et al. 1999).

The sections below illustrate four more applications of Bayes' rule. In the sections that follow the examples, we will first summarize the advantages of the Bayesian approach and then review the limitations and deficiencies of such applications in risk analysis.

Numerical example: updating with sample values

Suppose that that your prior for some parameter θ is a normal distribution with a mean of 6 and unit variance, and that you have an observation $x = 13$ which is normally distributed with mean θ equal to the parameter of interest and a variance of 2. (The central limit theorem is often used as a justification for assuming that observations might be normally distributed.) Recall that a normal distribution on y with mean μ and variance σ^2 has probability density

$$\mathrm{P}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2} \right).$$

In this case, we have the prior

$$\mathrm{P}(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\theta-6)^2}{2} \right)$$

and the likelihood

$$\mathrm{P}(x \mid \theta) = \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{(x-\theta)^2}{4} \right),$$

so that, when we multiply these functions, expand the squares, and omit constants of proportionality, we obtain

$$\mathrm{P}(\theta \mid x) \propto \mathrm{P}(\theta)\mathrm{P}(x \mid \theta) \propto \exp\left(\frac{-3\theta^2 + 2(12+x)\theta}{4} \right).$$

Plugging in the observed value $x = 13$, we find the posterior density is proportional to $\exp((-3\theta^2+50\theta)/4)$. It can be shown that this is a normal distribution with a mean of 8□ ≈ 8.333 and a variance of □ ≈ 0.667. The graph below depicts the prior distribution, the likelihood function and the posterior distribution they yield.

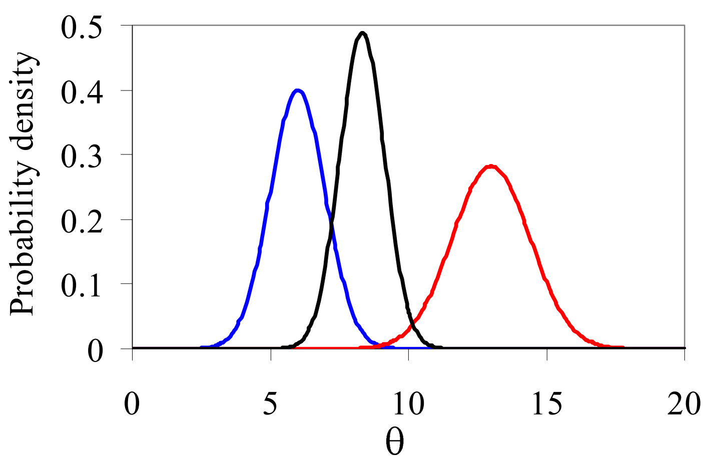


Figure 2. Prior distribution (blue), likelihood function (red) and the resulting posterior distribution (black) for θ.

This application of Bayes' rule is called updating because we have updated the prior to account for the evidence represented by the sample value. The calculation can be repeated for additional sample values by using the posterior as the prior for each next application of Bayes' rule for new sample values. The order in which the sample data arrive generally does not matter, nor does whether the sample values come singly or in groups. The posterior distribution we eventually obtain is the same, regardless of order and grouping of the data. As data accumulates during sequential updates, the initial choice of the first prior has a smaller and smaller influence on the final posterior, assuming the prior was reasonable to start with (see the section below on zero preservation). Because analysts typically have different prior information, their posteriors will likely differ. But as new data are collected and shared among the analysts, their posteriors will tend to converge.

Numerical example: model uncertainty

Suppose we're not sure whether a binary function f should be modeled as an addition or a multiplication. This would mean that either

$$f(A, B) = f_1(A, B) = A + B$$

or

$$f(A, B) = f_2(A, B) = A \times B$$

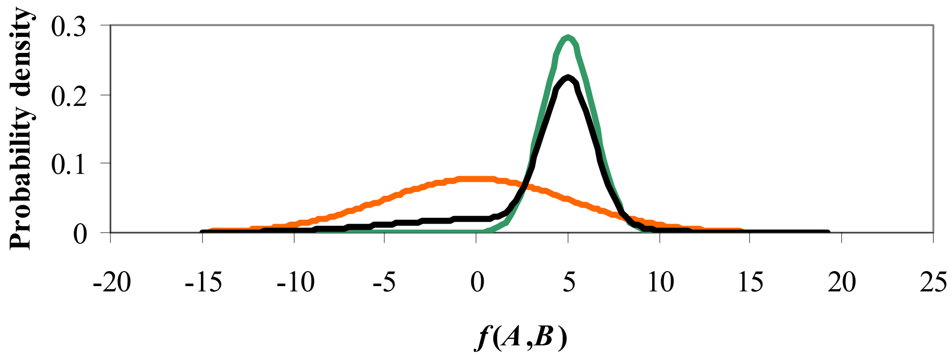
where, say, A is normally distributed with mean zero and unit standard deviation, and B is normally distributed with mean 5 and unit standard deviation, which we denote as

$$A \sim \text{normal}(0, 1),$$
$$B \sim \text{normal}(5, 1).$$

One sample of $f(A, B)$ has been taken and observed to have the value 7.59. This single piece of evidence can be used to compute the likelihood for each of the two models. The likelihood associated with the first model is the probability density associated with the value 7.59 for the convolution $A+B \sim \text{normal}(5, \sqrt{2})$. This value can be computed in Excel with the expression "=NORMDIST(7.59, 5, SQRT(2), FALSE)" and is about 0.05273. The likelihood associated with the other model is the density at the same x-value of the distribution for the multiplicative convolution $A+B$, which is approximately normal in shape with zero mean and standard deviation of $\sqrt{26}$. This value is about 0.02584. These likelihoods are then to be multiplied by the respective priors. Suppose the prior probabilities are given as 0.6 that f is the additive model and 0.4 that it is multiplicative. From these, we can compute the (normalized) Bayes factors for the two models,

Additive:	$0.6 \times 0.05273 / (0.6 \times 0.05273 + 0.4 \times 0.02584) = 0.7538$
Multiplicative:	$0.4 \times 0.02584 / (0.6 \times 0.05273 + 0.4 \times 0.02584) = 0.2462$

In this case, the sample datum tends to confirm that the additive model is more likely, which was the prior belief as well. These values represent the weights that we should associate with the additive and multiplicative models respectively. Bayesians would use the Bayes factors as weights to construct a stochastic mixture distribution of the two convolutions $A+B$ and $A \times B$. The mixture would thus combine the additive model with probability of about ¾ and the multiplicative model with probability of about ¼. The resulting mixture distribution is depicted as the black curves in Figure 3. This distribution is the posterior distribution for $f(A,B)$ that incorporates model uncertainty, the relative beliefs about the two models, and the available sample data. If there had been no sample data at all available, the posteriors would be taken to be the unmodified priors for the two models, so the mixture weights would have been 0.6 and 0.4.



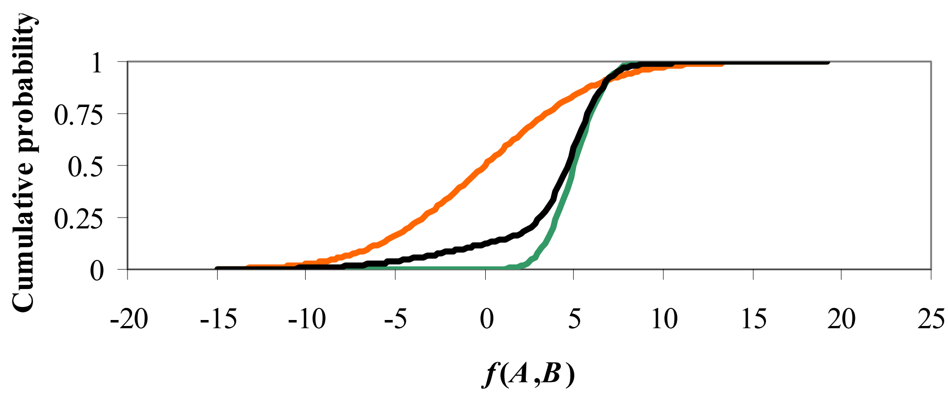


Figure 3. Mixture (black) of *f*(*A*, *B*) distributions for additive model (green) and multiplicative model (orange), with Bayes factors used as weights. Both density and cumulative displays are given.

Numerical example: updating with only constraints

Bayesian updating does not require the information be in the form of sample data. Suppose we have three continuous random variables W, H, and A, and there is very little information available about these variables, except it is known that their ranges are surely constrained to the intervals

$W \in [23, 33]$,
$H \in [112, 150]$, and
$A \in [2000, 3200]$.

Suppose these variables correspond to width, height and their product, area*. Does knowing these variables are related as $W \times H = A$ allow us to say more about the variables?

To use Bayes' rule to answer this question, we need to construct a likelihood function that expresses what the knowledge about their relationship says about the random variables. This can be done with the likelihood

$$L(A = W \times H \mid W, H, A) = \delta(A - W \times H)$$

where $\delta(\cdot)$ is the Dirac delta function that is positive when its argument is zero and zero everywhere else. This likelihood function concentrates all of the probability mass onto the manifold of feasible combinations of W, H, and A. Any triples such that $W \times H \neq A$ have a likelihood of zero, so they will be excluding by the updating process.

Because of the paucity of specific information about the random variables, a Bayesian analysis would use an "uninformative" prior to model them that doesn't prefer any values within the respective intervals. The most common choice is the uniform distribution over the volume described by the three intervals. This uniform distribution is

$$\Pr(W, H, A) = \frac{I(W \in [23,33])}{33 - 23} \times \frac{I(H \in [112,150])}{150 - 112} \times \frac{I(A \in [2000,3200])}{3200 - 2000}$$

where $I(\cdot)$ is the indicator function equaling one when its argument is true and zero otherwise. The probability density is equal for all triples within the constraints, and zero for all triples such that any of the random variables lies outside its range.

The density of the posterior distribution is proportional to the product of the prior and the likelihood

$$f(W, H, A \mid A = W \times H) \propto \delta(A - W \times H) \times \Pr(W, H, A) .$$

The integration necessary to solve for the posterior is cumbersome but straightforward. The cumulative distribution functions of the marginal distributions of the three random variables implied by the resulting posterior are displayed in Figure 4. The ordinate for each of the three graphs is cumulative probability. In addition to the distribution functions shown as black curves, also shown are intervals resulting from a naïve constraint analysis that about the possible range of each variable fiven what was known about the ranges of the other two. It is clear that the distributions are much more precise than intervals are. Because the distributions are not uniform within the intervals, they offer a much more detailed picture of how using knowledge of how variables are related has improved their estimates.

---

*Or they might be electrical current, resistance and voltage, or perhaps the concentration of a environmental contaminant, a receptor's intake rate and the corresponding exposure.

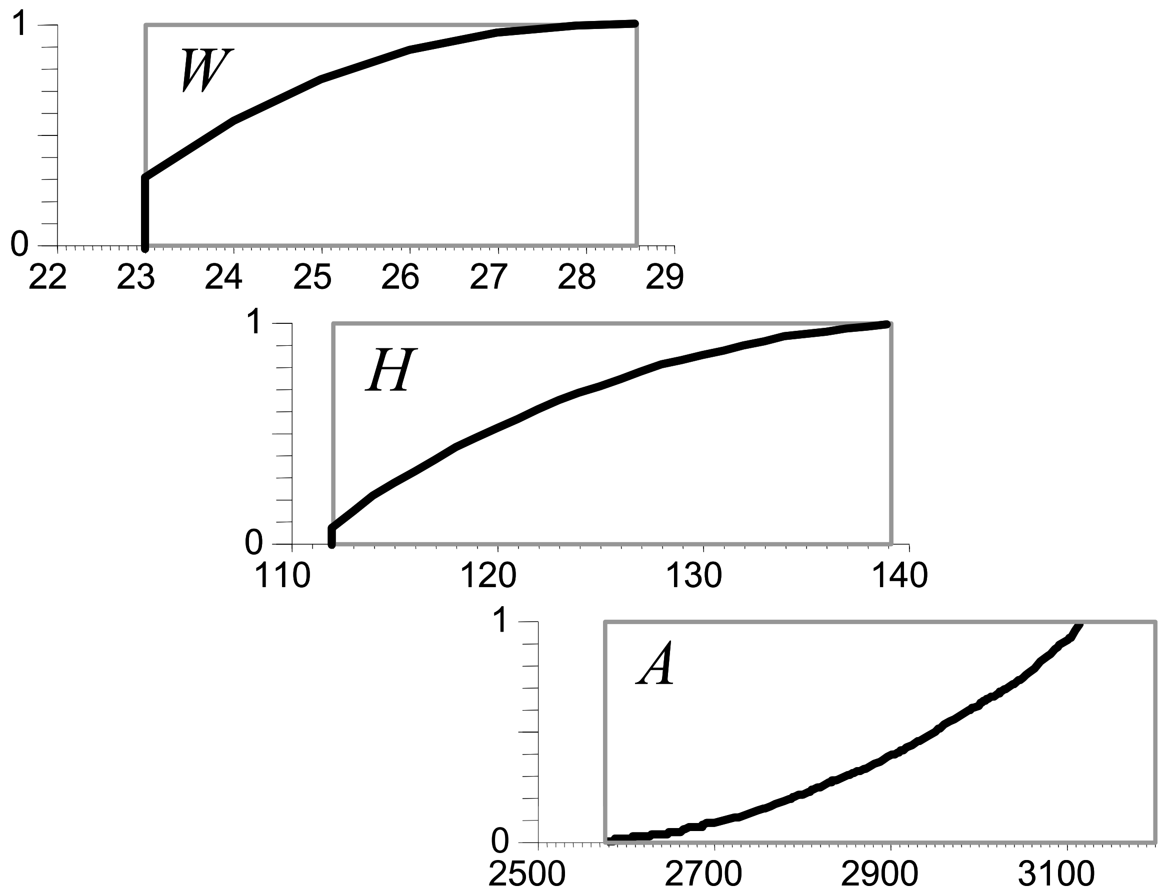


Figure 4. Posterior marginal cumulative distributions (black) for the random variables width, height and area along with their possible ranges (gray).

Numerical example: multinomial sampling

Suppose there is a bag of colored marbles and we want to know the probability that a marble drawn at random from the bag is red. After each marble is drawn from the bag, it is placed back in the bag which is then thoroughly shaken. If we could examine all the marbles in the bag and see what color each is, we could compute the probability as the number of red marbles divided by the total number of marbles in the bag. If, however, we can't examine all the marbles at the beginning, and can only see the color of each marble as it is drawn from the bag, we will accumulate sample data about the bag's marbles that will enable us to estimate the probability that the next marble drawn from it would be red.*

Consider a model with N random observations sampled with replacement. As each of the N marbles is drawn, we see its color $\omega_j \in \Omega = \{\omega_1, \ldots, \omega_k\}$, where k is the number of colors of the marbles in the bag. This is a multinomial model with the probability of drawing a marble colored ω_j is $P(\omega_j) = \theta_j$, where $j = 1,\ldots,k$, $0 \leq \theta_j$, and $\Sigma\theta_j = 1$. The θ_j are unknown. Given the observed

*This problem is an exemplar for a wide variety of statistical and risk-analytic problems involving independent and identically distributed trials with finitely many possible outcomes.

number n_j of marbles that were colored ω_j in the first $N = \sum n_j$ trials, the statistical likelihood function is proportional to

$$\prod_{j=1}^{k} \theta_j^{n_j} .$$

A Dirichlet prior is convenient to use here. The Dirichlet distribution is a multivariate generalization of a beta distribution. It has density proportional to

$$\prod_{j=1}^{k} \theta_j^{st_j - 1}$$

where s is some fixed positive number, $0 < t_j < 1$ for each j, and $\sum t_j = 1$. This distribution can be denoted more compactly as Dirichlet(s, t_j). The values t_j are the means for the different dimensions θ_j. When this prior and the likelihood are multiplied, the posterior has density proportional to

$$\prod_{j=1}^{k} \theta_j^{n_j + st_j - 1} ,$$

which corresponds to the Dirichlet(N+s, $(n_j+st_j)/(N+s)$) distribution.

Suppose the sample space is Ω = {red, blue, green, white, yellow}, so there are $k = 5$ possible colors. If we haven't been able to peek in the bag, we initially have no information about the relative probabilities of drawing a marble of any particular color. In this case, Bayesians might use Laplace's "principle of insufficient reason" to select an uninformative prior that doesn't have much influence on the result. The Bayes-Laplace prior is uniform over the simplex for the θ_j. In the five-dimensional case, the prior is Dirichlet(s=5, t_j=1/5). The corresponding posterior is Dirichlet(N+5, $(n_j+1)/(N+5)$). The vector of means for a Dirichlet(s, t_j) distribution is t_j, so the posterior expected value of θ_j is $(n_j+1)/(N+5)$, where n_j is the number of sampled marbles colored ω_j and N is the total number of marbles sampled. This is the "predictive probability" that the next marble is colored ω_j. If the observed sequence of marble colors from the first seven independent draws happened to be

blue, green, blue, blue, red, green, red,

the probability the next marble drawn is red would be computed as (2 + 1) / (7 + 5) = 0.25. This value can be compared to the observed frequency of red marbles 2/7 $\approx$ 0.2857.

What can be said before any marbles have been drawn? An important feature of the Bayesian approach is that it yields a precise value for the probability the next marble is red even *before any data have been collected*, that is, when $N = 0$. Of course, frequentists cannot begin to answer this sort of question, but being able to could be a great advantage in the context of risk analysis where empirical data for variables are sometimes nonexistent. The probability that we'll get a red marble on the very first sampling is (0 + 1) / (0 + 5) = 0.2, which is just the mean of the prior probability for θ_{red}.

Advantages of the Bayesian approach in risk analysis

Many proponents of the Bayesian approach praise it with almost religious zeal, and, to be sure, the approach has several significant advantages that recommend it for wide use. This section summarizes the most important of these advantages for its use in risk analyses. Insofar as statistics is part of risk analysis, we should also consider the advantages of Bayesian statistics over traditional frequentist statistics against which it is usually contrasted and with which scientists are most familiar. Although the frequentist school of statistical inference developed by Fisher, Neyman, Pearson, Gosset and others is still dominant in science, the Bayesian school has grown enormously over the last quarter century. Berry (1996; 1997) compares Bayesian statistics with the frequentist perspective, and finds many advantages for the former. The following paragraphs highlight some of these advantages. They are listed in order of increasing interest to risk analysts.

Naturalness. By redefining probability as a subjective quantity rather than a measure of limiting frequencies, Bayesians can compute "credibility intervals" to characterize the uncertainty about parameter estimates. They argue that credibility intervals are more natural and easier to work with than traditional Neyman-Pearson confidence intervals whose interpretations make them hard to understand and for which there is no calculus permitting their use in subsequent calculations. Bayesians use probability distributions for both the data and the parameters of their models. In contrast, frequentists do not permit distributions to be used as models of fixed quantities, so they usually cannot use distributions to represent their model parameters. Frequentists can only consider only the probabilities for random data they collect. As a result, frequentist models and inferences have to be fashioned in a way that seems contorted to many. For example, the measures that frequentists compute depend on the experimental design and can only take account of information collected in that experiment. Bayesians can use all the available information, even if that information came from sources outside the experiment. Frequentist measures consider the probabilities of data sets that were possible but that didn't actually occur. (This is how they can tell whether the data they have are surprising enough to reject the null hypothesis.) The Bayesian posterior depends on data only through the likelihood. The likelihood is calculated from real data which were actually observed. Frequentist measures depend on experimental design and require the experimental design be followed. Results measure the probability of observing data as or more extreme than those actually seen and therefore depend on the rule employed to terminate data collection. Inference is impossible within a frequentist approach when one uses the "scientific stopping rule" (stop collecting data when you have enough information to draw a conclusion). Bayesians, of course, update continually as data accumulates; they don't need to chose their sample sizes in advance and can stop collecting data whenever they like. Unlike frequentist inference, their calculations and the interpretation of their results do not depend on their having carried out the experiment according to its original plan. This means that Bayesian methods enjoy great flexibility and can benefit from unplanned windfalls and serendipity.

Data mining. Another important advantage of the Bayesian approach is that it allows peeking at the data. The classical Neyman-Pearson school of statistics holds that data mining is scientifically improper. Because one cannot legitimately compute *p*-values after having examined the data, one should not peek at the data without having already formed an a priori hypothesis. In this conception, data contaminates the mind of the scientist. Under this philosophy, a person should take a graduate-level course in statistics as preparation, then collect one data set, compute the *p*-value for a single a priori hypothesis, and then retire from science. The Neyman-Pearson school also holds that it is inappropriate to create models with a large number of parameters for data sets with relatively few samples. Likewise, one should not try to estimate more than a few parameters at a time. Of course, scientists in practice do not—and could not—follow these strictures. Data are always in short supply and, in the process of collecting them, scientists inevitably peek at the data values. Hypotheses occur to us only as we see the world revealed in new data, and models invariably get more and more complex as we know more about the world they represent. Every practicing scientist engages in multiple hypothesis tests, and we generally do not make use of any multiple-test corrections (such as suggested by Bonferroni or Scheffé) simply because doing so would make statistical significance impossibly elusive. Users of Bayesian statistics have no guilt about peeking at the data. This is perfectly proper under the Bayesian approach. They don't hesitate to estimate a lot of parameters all at once, and they don't worry too much about having few data samples.

Decision making. Bayesian methods are tailored to decision analysis which in principle could allow analysts and decision makers to construct a consistent set of decisions about the risk assessments and their proper management. In fact, it is possible to guarantee that decisions are sensible in that they meet the axioms of coherent decision theory by expressing all uncertainties with probabilities and employing the Bayesian approach (Savage 1954; Fishburn 1981; 1986). Traditionally, the only decision made by frequentists is whether a data set is sufficiently extreme to justify rejecting a null hypothesis or not. In making this decision, frequentists don't usually balance the twin costs of making a wrong decision either way, but control only the so-called Type I error of falsely rejecting the null hypothesis and thereby admitting an unsupported (positive) conclusion into the scientific canon. There are two costs: the cost of thinking something is true when it's actually not, and the cost of thinking something is false when it's actually true. In practical decision making, it is generally essential to consider both these costs and balance them against each other in making reasoned decisions. Decision analysis does just this and Bayesian methods are integrated into decision analysis at a fundamental level. As mentioned in the introduction, it is possible express the entirety of risk management within a Bayesian decision context. Although risk analysts have not yet called for this, the intellectual infrastructure necessary to do it already exists in the Bayesian approach.

Rationality. Bayesian often tout the "rationality" of their approach. Their claims refer to a series of rather technical mathematical theorems (Cox 1948; Savage 1954; de Finetti 1970; Fishburn 1981; 1986; Dawid 1982; inter alia). Bayesian rationality is tied to the idea of maximizing expected utility, and, to operationalize this idea, probabilities are interpreted in terms of betting. A gambler is said to be incoherent if bookies who cover his bets can choose terms that assure they can always reap a positive payoff from the gambler whatever way the outcome of the bet. (Remember that having a probability means you have to take all bets at some odds.) Essentially, such gamblers are falling for the line "heads I win, tails you lose". The beliefs and preferences that led the gambler to this fate are said to be irrational because they cause the gambler to act against his own interests. This sorry state of affairs is called "sure loss" and, in principle, the gambler can be made into a money pump. If their probability assignments are coherent, gamblers can avoid sure loss. This is one of the basic features of rationality. The estimates could be wrong, but they ought to at least make sense by not contradicting each other. Many adherents (e.g., Jaynes 2003) suggest that Bayesian methods provide the *only* consistent approach possible for handling uncertainty in calculation and inference. Bayesians also argue that updating via Bayes' rule supports the accumulation of knowledge. Different people will typically have different preconceptions, and would therefore be likely to draw rather different inferences when data are sparse. Applying Bayes' rule to incoming (well behaved) data creates an almost Millsian convergence to the truth as information accumulates. Dawid's theorem suggests that rational people given the same information will come to agree about probabilities for convergent sequences of trials, and that these probabilities will be the same as the observed frequencies.

Although different analysts might start out with very different priors, if they both see the same data, usually their posteriors will eventually come to agree about their conclusions.

Subjective information. The Bayesian approach formalizes—and to some extent legitimizes—the explicit use of subjective information, including personal judgments made by the analyst and expert opinions that the analyst may elect to adopt. It gives the analyst a way to factor subjective judgments into the objective equations. This is an extremely attractive feature to risk analysts because, in many practical assessments, most of the available information is partly or entirely subjective. Although it would probably be preferable to have validated empirical information, it seems unreasonable to deny the usefulness of the knowledge in the heads of the analysts and their consulting experts. Ignoring this subjective knowledge would be profligate in situations where information of any provenance is highly limited. Bayesians acknowledge that frequentists also use subjective knowledge, but complain that they do so in a cryptic way, such as by modifying the model structure or by altering parameter choices. Bayesians argue that an explicit accounting for such information is better than the frequentists' adopting it implicitly in a way that prevents the tracing of its effects on the computational results.

Working without data. A seminal advantage of Bayesian methods over frequentist methods is that Bayesian methods can in principle always yield an answer, even in situations were frequentist methods cannot be used. Specifically, they can produce answers even when there is no sample data at all. The trick is of course to use the unmodified prior as the posterior. Because there are no data, the application of Bayes' rule would not alter the prior. In this sense, the initial prior is the zeroth posterior. The legitimacy of this feature of the Bayesian approach has been questioned on scientific grounds. The next section discusses this and various other criticisms of the Bayesian approach.

Limitations and disadvantages of Bayesian methods

This section reviews the drawbacks, disadvantages and limitations of using Bayes' rule in risk analysis and uncertainty propagation. It finds the most salient of these to be (*i*) the inadequacy of the Bayesian model of ignorance which doesn't distinguish between incertitude and equiprobability, (*ii*) the consequent overconfidence (or apparent overconfidence) of conclusions, and (*iii*) the acceptance (or necessity) of using subjectivity even in public policy decision making. On top of these issues, would-be users of Bayesian methods have to contend with the often considerable analytical and computational difficulty that hampers the derivation of solutions and computation of numerical answers, as well as the controversy engendered in some quarters by using Bayesian methods.

The subsections below review several of these issues in turn. Where possible, the discussions refer to the numerical examples offered earlier to explain precisely where the deficiencies of the methods appear. The topics of the first sections are those about which the most clamor has arisen, whereas the topics of the latter sections are those that represent perhaps the gravest problems for the use of Bayesian methods in risk analysis.

Controversy

There is nothing controversial about Bayes' rule per se. Bayes' rule is a *theorem* of probability theory that provides a computational formula by which conditional probabilities of different forms can be interconverted. There is nothing controversial about the need to convert conditional probabilities, nor about the propriety—and necessity—of using Bayes' rule to effect the conversion. The controversy arises in applications of Bayes' rule to situations in which there is no underlying random process. Such an application is called the "method of inverse probability" (Edwards 1972). According to this method, the deductive argument leading from hypothesis to probability of results is inverted to form an inductive argument from results to probability of the hypothesis. Especially when combined with the use of subjectively determined probabilities as inputs, this method is commonly called the "Bayesian approach". Although the mechanics of the inverse argument originated with Bayes (1763), the idea was really developed by Laplace (1820) and championed by Jeffreys (1961) and Jaynes (2003). Although it has never lacked proponents of great authority, the idea has always been controversial. In fact, Edwards (1972, page 51) suggests that Bayes himself probably would not have subscribed to the idea because he was well aware of the assumptions required to make it work. The idea withered under criticism during the 19th century (Cournot 1843; Boole 1854; Venn 1866, pages 146-166; Peirce 1878; 1933, volume 2, sections 669-693; Bertrand 1889, pages 171-174), but has enjoyed a renaissance during the past several decades. In the last fifteen years or so, vigorous proposals have been made to extend the Bayesian approach into various arenas of science, engineering and policy studies. Adherents of the Bayesian approach argue, often with great fervor, that it is the *only* rational scheme for probabilistic modeling and inference, and that it alone provides intuitive structures and natural reasoning schemes for accumulating and synthesizing knowledge (Jaynes 2003; see Malakof 1999; Jaffe 2003). This claim has left dumbstruck many scientists with traditional statistical training. Needless to say, some of the enthusiasm of Bayesians has been criticized by non-Bayesians as overstated, exaggerated, or downright misleading (Fisher 1956; 1973 page 24-38; Edwards 1972; Feller; Glymour 1980; Mayo 1996; Walley 1999; Simon 2003; see also Jeffreys 1961, pages 127-132). The wellspring for the controversy has not been the *philosophical* objections to the ideas behind Bayesianism, although they are numerous, but rather the many limitations and disadvantages of the approach as it has been applied *in practice*. The following sections describe these issues.

Bayes' rule doesn't tell you anything you didn't already believe

Scientists need theories that are rich enough to be capable of explaining or accounting for a wide variety of phenomena. On the other hand, a theory that says anything is possible is not scientifically useful. This section illustrates how setting the prior totally determines the outcome of a Bayesian analysis. In the numerical example estimating the probability of paternity on page 9, we used ½ as the prior probability the man was the girl's father. We chose this value mostly because we really weren't sure, in the absence of more information about the lives of the man and the woman, what other value to use. But what is the impact of using ½ for the prior probability? Certainly the prior could have many other values. Is setting it to one half really the proper way to denote ignorance? It is easy to compute the posteriors that could have been obtained by varying the prior over all possible values between zero and one. The result is the graph shown in Figure 5. This display is interesting and very telling about how Bayes' rule really works. It says that you can get any value for the probability of paternity at all just by wiggling the prior you use. Moreover, it also says that, if you genuinely don't know what the prior probability that he is father ought to be, then you don't know after considering the blood evidence either. In other words, if you'd say the prior could be any value between zero and one, then you are left with a vacuous posterior probability of paternity given the blood results, that is, it could likewise be any value between zero and one.

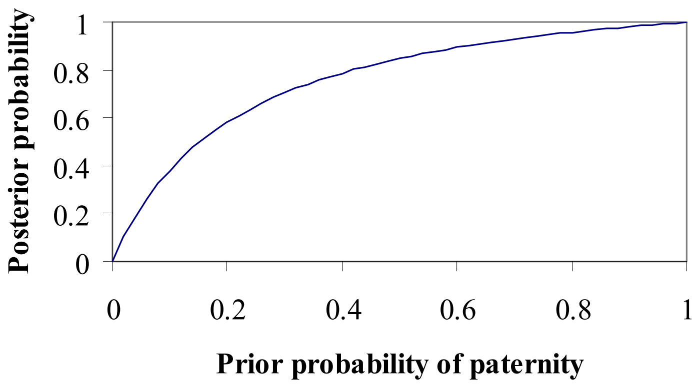


Figure 5. Posterior probability of paternity as a function of its prior probability.

Critics of Bayesian methods complain that Bayes' rule doesn't tell us anything new so much as merely suggest which direction our prejudices should skew. Bayesians are often lampooned as the mathematicians who ask what you think before an experiment in order to tell you what to think afterwards. Proponents of the Bayesian approach counter that this flexibility and ability to reflect preconceptions is entirely appropriate. If the man knows he didn't have sexual intercourse with the woman, then he knows he could not be the father, and the girl's blood type would be totally irrelevant to this conclusion. If the woman knows she's only ever had sexual intercourse with this one man, then she knows he is the father. Genetic testing that purported to prove he is not (such as finding he had type A blood and could not have donated the B allele to the girl) would probably lead her to doubt the genetic testing before doubting his paternity.

Bayesians maintain that if I know the coin is biased, my knowledge better be reflected in my decisions about how to bet on it. Kyburg (1999) argued that when the prior is justified, then Bayesian methods are just right for computing posterior probability, but "sometimes we don't have a prior probability, and sometimes we want a conclusion concerning frequencies or distributions or facts" rather than merely beliefs. Mayo (1996) has argued compellingly that Bayesian methods fail to provide the critical connection with empirical reality which is the hallmark of most sciences (if they deserve that name). She likens Bayesians to people who scrupulously balance their checkbooks but never bother to look at their bank statements. Cooke (2004) has conducted thousands of experiments to check the *consistency* of the experts whose opinions he employs in Bayesian analyses. Interestingly, however, he does not check whether the experts are actually *correct* in their predictions about the real world.

Of course, it's really only the degree of curvature of the function in Figure 5 that represents the evidentiary contribution of the blood data to the question about the man's paternity. If its second derivative is strongly negative, the evidence is saying that his paternity is likely. If it's strongly positive the evidence is telling us otherwise. It is important to understand, however, that the import of the evidence is entirely contained in the likelihood function, and, as Edwards (1972) has argued, we don't need to use Bayes' rule or Bayesian methods to wield likelihood as a way to understand data.

Need for a prior

The first issue about which novice Bayesian analysts become agitated is their responsibility to select a prior. The selection is supposed to be subjective, so, in principle, no one should be able to complain about one's choice for a prior, but even seasoned analysts may often be anxious about these selections. The reason they may be concerned is that the choice often *matters*. Although proponents of the Bayesian approach often argue that data will quickly overwhelm and obviate the choice of the prior (Jaynes 2003), this is not always so in practice. Analysts are left with a serious conundrum. Hamming (1991, page 298) wrote bluntly "*If the prior distribution, at which I am frankly guessing, has little or no effect on the result, then why bother; and if it has a large effect, then since I do not know what I am doing how would I dare act on the conclusions drawn?*" [emphasis in the original]. The advice given to Bayesian analysts on selecting priors is two-fold: (*i*) justify assumptions and evaluate their plausibility in view of what is known, and (*ii*) explore the sensitivity of the results of the analysis to the assumptions and the prior. (The sensitivity of the results can be explored comprehensively using robust Bayes techniques; see page 38.)

So where should priors come from? How should they be determined? To what extent can they be justified? Many approaches for the selection of priors have been suggested, including conjugate priors, maximum entropy priors, uninformative priors, such as uniform priors, Jeffreys' priors, and reference priors, empirical Bayes methods, empirical distribution functions, personal and subjective priors, and distributions based on a betting rate interpretation of probability. The following paragraphs consider each of these approaches.

Conjugate priors. One could adopt as one's prior any probabilistic characterization that leads to the most convenient mathematical properties. As explained on page 29, selecting a conjugate prior for a given sampling model allows us simply write down the characterization of the resulting posterior, without the necessity of performing any integrations or complicated mathematics. These choices for priors often have little to recommend them except their mathematical convenience, and are often deprecated for this reason.

Maximum entropy priors. Alternatively, one could employ the maximum entropy criterion (Jaynes 1957; 2003; Levine and Tribus 1976; Grandy and Schick 1991), in which constraints known for a random variable are used to identify a class of possible distributions and the distribution from the class with the greatest Shannon-Weaver entropy is selected as the prior. Using this approach, analysts can avoid the need to select the prior subjectively if the constraints on the random variable can be identified in some objective fashion. The maximum entropy criterion is a generalization of Laplace's "principle of insufficient reason" (sometimes called the "indifference principle") which was used to motivate the selection of uniform distributions when only the range of the random variable was known. The criterion uses whatever information may be available about the variable. It makes no shape assumptions and specifies the prior distribution in an optimal way. In summary, the approach suggests the following priors be used (Lee and Wright 1994):

When you know	*Use this shape**
{minimum = m, maximum = M}	uniform(m, M)
{mean = μ, variance = σ^2}	normal(μ, σ)
{minimum = 0, mean = μ}	exponential(μ)
{minimum = m, mean = μ}	exponential($\mu - m$) + m

*The formulations use the syntax of RAMAS Risk Calc (Ferson 2002).

{min = m, max = M, mean = μ, var= σ^2}	beta1((μ–m)/(M–m),σ/(M–m))(M–m)+m
{minimum = 0, p^{th} quantile = q}	exponential(–q / ln(1–p))
{minimum > 0, p^{th} quantile = q }	gamma1(–q / ln(1–p),2) + m
{discrete, minimum = m, maximum = M}	discreteuniform(M – m) + m
{minimum = m, maximum = M, mean = μ}	(M – m) × beta1(λ, ρ) + m

where $\lambda = (\mu-m)/(M-m)$ and $\rho = \lambda(0.99-0.4\lambda-1.15\lambda^2+0.55\lambda^3)$ are fitted to maximize entropy. Solana and Lind (1990) described an approach to formulating an empirical distribution function to a collection of sample data that is based on the maximum entropy criterion. Abbas (2003) derived the maximum entropy solution for a random variable about which only constraints on its percentiles are known. Such constraints define stepwise bounds on the distribution function in the form of a p-box (Ferson 2002). The solution is the “taut string” distribution whose shape follows the shortest path lying totally within a p-box and connects its far corners. Of course, if absolutely nothing is known about constraints on the random variable, then the maximum entropy criterion cannot specify a distribution to use as the prior.

Many objective Bayesians argue that using maximum entropy is better than simply guessing at a prior, and many consider it the state of the art. The maximum entropy criterion is controversial however. Among several different kinds of criticisms, perhaps the most serious problem with the approach is that the model of uncertainty it uses is inconsistent through changes of scale. For instance, suppose we are interested in a prior for the half-life of some environmental contaminant, which we know only to be within, say, 6 and 12 days. The maximum entropy criterion would suggest using a uniform distribution over this range to represent this state of knowledge. Now consider the related variable degradation rate. If all we know about the half-life is its range, then surely all we know about the degradation rate is its range, which would be between 1/12 days and 1/6 days. Yet, if we had employed the maximum entropy criterion to find a prior on the degradation rate directly, we’d have obtained another uniform distribution over its possible range. It is clear, however, that these two uniform distributions are not consistent with each other. (The reciprocal of a uniformly distributed variate is not itself uniformly distributed.) Other changes of scale, such as going to or from logs, produce similarly inconsistent characterizations of uncertainty. The difficulty with the practical use of the maximum entropy criterion, therefore, is that the distribution it recommends depends on the scale in which the random variable is expressed. The inconsistencies mean that analysts must arbitrarily pick a scale on which to express their uncertainty and resist comparing it across different scales.

There have been many other criticisms of the maximum entropy criterion. One fundamental criticism is that it seems wrong to treat some information such as range or moments as perfect in determining constraints while other kinds of information is regarded as entirely absent or completely unreliable (Uffink 1997). It would rather seem that all such information would be tentative to varying degrees. Other criticisms are of a more technical nature. For instance, Hamming (1991, page 269f) points out that there are distributions with infinite entropy. Cases exist for both discrete and continuous distributions with both unbounded or even finite supports (the latter case depending on Dirac-like infinite probabilities values). A criterion based on maximizing entropy could hardly exclude such distributions. If the class of constraints admits multiple distributions with infinite entropy, it is unclear how one would select the prior from among them.

Uninformative priors. Prior distributions are often selected to represent complete ignorance about the parameter. Such a prior is said to be “uninformative” or “noninformative”. They are also sometimes called “flat”, “conventional”, “default”, “neutral”, “non-subjective”, “objective”, "formal", or "reference" (Syversveen 1998; Irony and Singpurwalla 1997). Amusingly, Vose (2000) calls them "uninformed" priors. Typically, uniform distributions are used for this purpose, following Laplace's principle of insufficient reason. But the use of a uniform prior can be problematic. A uniform with an infinite range is called an improper prior. Although improper priors are not distributions themselves (because their areas are not unity), they can sometimes lead to posteriors that are proper distributions. Improper priors are commonly used in practice although some Bayesians frown on their use (Irony and Singpurwalla 1997). The main problem with uniform priors (whether they are proper or improper) was foreshadowed in the discussion of the maximum entropy criterion above. Complete ignorance about a parameter would seem to imply complete ignorance about functions* of the parameter (Fisher 1956, page 16f; Edwards 1972, page 58). What constitutes an uninformative prior for θ is typically not uninformative for transformations of the parameter such as $1/\theta$, θ^2, or $\sqrt{\theta}$. One may, however, be able to select a prior that is invariant to transformations of interest. For example, a loguniform prior (i.e., a distribution such that $\ln(\theta)$ is uniformly distributed) is invariant to any linear transformation of θ. A Jeffreys prior is constructed to be a prior that is invariant to any one-to-one transformation. For instance, the beta(½,½) distribution is the Jeffreys prior for the probability parameter of a binomial distribution. This beta distribution, whose density is shaped like a deep bathtub, certainly doesn't *seem* uninformative however. Paradoxically, which prior is uninformative depends on the likelihood the prior will be combined with. In general, there isn't a unique uninformative prior for a model; there will be different ones for different parameters of interest.

Struggling out of the conundrum that there is no such thing as *the* uninformative prior distribution, Bernardo (1979; Bernardo and Smith 1994, chapter 2; Irony and Singpurwalla 1997) argued that the purpose of an uninformative prior is not to represent ignorance anyway (no prior can do that). Indeed, it's not intended to describe the beliefs of the analyst in the first place. Rather, it is merely a technical device that allows us to minimize the influence of the prior and thus discover what the data themselves are saying about the posterior. In this sense, using an uninformative prior is just a part of a Bayesian sensitivity analysis (see page 38). This argument frees analysts in one way, but restricts them in others. First, it admits that no prior can reflect lack of knowledge. At best it can only represent ignorance *relative* to the information given by the data. Second, it means that uninformative priors should not be interpreted as probability distributions. In particular, they cannot be used to make predictions in the absence of data (although such predictions are legitimate when the prior reflects the analyst's personal beliefs). Bernardo devised the notion of "reference priors" as a way to compute model-based posteriors that describe the inferential content of the data for scientific communication in a non-subjective way. In special cases, these degenerate to Jeffreys' priors. Yang and Berger (1998) provide a partial catalog of uniform, Jeffreys, reference and other uniformative priors and their known properties.

Subjectivists criticize the use of reference priors and uninformative priors generally. They feel that priors should reflect the analyst's prior knowledge, and should not be a function of the model, especially if this violates the likelihood principle that *all* the information provided by the data should be embodied in the likelihood function.

Empirical Bayes. Newman (2000) and others have suggested a method called "empirical Bayes" in which maximum likelihood estimates are used to determine model parameters for a parametric prior. The same data are then used to form the likelihood, so this approach obviously uses the data twice, which many Bayesians, as well as non-Bayesians, find objectionable.

---

*Interestingly, Jaynes (2003) denies this extension.

Empirical distribution functions. Thompson (2000) suggested that one could use the empirical distribution function summarizing available sampling data about a parameter as the prior distribution. Her suggestion does not make clear, however, why such empirical data wouldn't be considered the evidence to be used in the current updating, rather than the prior, or, if it is old data, why it would not already have been incorporated into a posterior generated by a previous application of Bayes' rule now to be used as the prior, and thus obviating the need to select a prior at all. In either case, it also seems to be a double use of the data. Even if there were some reasonable way to form a prior from an empirical distribution function, it would certainly be a poor idea to allow the prior to be zero outside the observed range of the data (see the section below on the zero preservation problem).

Personal and subjective priors. In principle, a prior is intended to reflect the analyst's personal beliefs about the parameter in question. Probability is used as a measure of a person's private beliefs. In fact, many Bayesians feel that this is the only appropriate kind of prior. Non-Bayesians, on the other hand, tend to throw up their hands when presented with such objects. The problem with a prior that is personal and purely subjective is that it will generally be hard to say what it means—or should mean—to anyone else. Even though I might proclaim that my subjective probability for event *A* is, say, 0.2, it is unclear what that really implies about my beliefs about the event. It is not necessarily clear that my beliefs about event *A* are commensurate with my beliefs about another event *B*, and merely assessing them on a single scale may not be entirely appropriate in the first place. Even subjectivists need some way to operationalize their definitions and interpret their statements. Historically, most analysts using Bayesian methods have declined or been reluctant to incorporate their personal beliefs into the priors for scientific work, even though the philosophy of the Bayesian approach gives them license to do precisely this.

Betting rate priors. A very old idea is that probability, that is, my belief about an event can be identified with how much I'd willing to bet that it occurs. So long as I'm rational, I should be willing to accept any bet for which the expected payoff is positive. In particular, if I believe the probability of event *A* is 0.2, then I should be willing to buy a gamble that pays one euro if *A* occurs (and nothing otherwise) for any amount up to 20 cents. Likewise, I should be willing to buy a gamble that pays one euro if *A does not* occur (and again nothing otherwise) for any amount up to 80 cents = (1−0.2) euros. In principle, by accepting or rejecting such gambles, my probability can be elicited perfectly. Probability distributions can be elicited similarly as a sequence of probabilities about events like $\theta < 1$, $\theta < 2$, and $\theta < 5.6$. The distribution is the collection of all such exceedance probabilities. Some probabilists consider betting rates to be the very definition of (subjective) probabilities, but others argue that betting quotients may fail to reflect true beliefs (Irony and Singpurwalla 1997). In any case, when it can be operationalized to yield precise values, it provides a clear interpretation of what probabilistic statements mean in terms of behavioral predictions about the agent.

Subjectivity

Within the Bayesian framework, the accumulation of knowledge occurs as an inferential chain of updates, each link of which is an application of Bayes' rule. Whenever new information arrives, it is added to the analyst's knowledge by treating a previously obtained posterior as a prior and conditioning it on the new knowledge. The inferential chain must have a beginning somewhere. This beginning represents the situation before any evidence or specific information is available. Such a state, by definition, must be purely a reflection of one's subjective judgments. Thus, subjectivity is an integral part of Bayesian inference.

Hamming (1991, page 298) argued that the Bayesian approach, although it sometimes yields reasonable answers, is not scientific because it traffics in subjective judgments. Although this conclusion may be rather facile, it is important to critically review the Bayesian embrace of subjectivity. Recognizing the inescapability of subjectivity does not mean that one shouldn't still strive for objectivity, at least when risk analyses are deployed in scientific or public policy uses. The fact that we cannot avoid every vestige of subjectivity does not justify the abandonment of objectivity altogether. The objectionable aspect of the Bayesian approach is not that it allows or even requires subjectivity, but that it seems to *wallow* in it. Of course modelers see the world through subjectively colored glasses, and certainly there is no escaping some degree of subjectivity in any non-trivial modeling exercise. But the attitudes and practices that (some) Bayesians bring to this issue seem scientifically perverse. It would be like physicists claiming that, because Heisenberg showed that no measurement could be made perfectly, one shouldn't bother to make measurements anymore. Surely we should struggle to maintain and increase objectivity* whenever possible. If different analysts are always getting different results, what is the benefit of analysis at all?

Objectivity is essential in any scientific enterprise because it makes possible the reproducibility that makes the analysis relevant. It also endows the results with an imprimatur of reasonablenss (if not correctness) by intersubjective validation. In short, it is the glue that holds any science together and justifies public confidence. Of course different analysts will produce different analyses, and different modelers will sculpt different models. But, in science, we must all strive for objectivity to minimize this volatility and identify real disagreements as the locus for further empirical inquiry. Insofar as risk analysis is a science, it disavows the radical multiculturalism that denies a right answer even exists. The danger of the Bayesian approach is that it gives—or appears to give—license to all manner of quirkiness. Leading Bayesians (e.g., Berger 1985; Bernardo and Smith 1995) recognize this danger and take pains to emphasize the importance of sensitivity analyses as a way to disperse any doubt associated with the choice of a specific prior. "Robust Bayes" methods (see page 38) have been developed for doing Bayesian sensitivity analysis. The idea is that an inference or decision is robust if it is not very sensitive to changes in the assumptions or the model that produced it. It is interesting to note that, when the charge to undertake sensitivity analyses is taken seriously enough to actually address the question of how much subjectivity quantitatively influences the inferences, the resulting analysis can no longer be described as Bayesian or even classically probabilistic because it has to abandon the idea that all beliefs are encapsulated into a single precise probability measure and all preferences are embodied in a single precise utility measure (Walley 1991; Pericchi 2000).

Some Bayesians themselves denigrate the use of subjectively selected priors. Most Bayesian statisticians throughout the last half century have been reluctant to use personal priors in scientific analyses. Many have sought alternatives that could make their results more objective, or at least *seem* more objective (Yang and Berger 1998). This is why a great deal of the Bayesian literature has been devoted to the topic of uninformative or non-subjective priors (see the previous section).

---

*It is objectivity and, by the way, measurement that create evidence. Bayesians sometimes argue that, so long as their beliefs and the inferences drawn from beliefs are coherent, their reasoning is rational and their conclusions sound. Walley (1991, pages 396f) acknowledged that standard Bayesian inferences are coherent, but complained that coherence is only a minimal requirement of rationality. He argued that beliefs ought to also *conform to evidence*, which Bayesian inferences rarely do. The real problem with the Bayesian approach is not merely that it admits subjectivity, but that, in abandoning objectivity, it has forsaken the essential connection to evidence and accountability in the real world (Mayo 2003).

Some analysts who call themselves "objective Bayesians" argue that one doesn't have to involve subjectivity in analyses, but most Bayesians think subjectivity is inescapable and that the ability of Bayesian methods to account for and legitimize its use in scientific inference is a crucial feature of the approach. In fact, many consider subjectivity even more central to "Bayesianism" that the use of Bayes' rule itself.

Other Bayesians stress the importance of responsible judgment by the analyst as the key to justifying the use of subjectivity. They argue that the analyst should undertake to justify the prior and be honest about its selection. In a sense, this is taking personal responsibility for one's personal beliefs. By combining candor with strong science and serious scholarship, one should be able to construct a prior that commands the respect of peers and public on its own merit. This idea is attractive to an undergraduate sensibility, but it seems antiquated if not foolish in our post-post-modern environment of gamesmanship and colliding worldviews. And, even if the idea could work, it still seems wrong. No matter how well trained and nobly motivated the analyst may be, it is not at all clear that it is proper or desirable to have the analyst's personal beliefs embodied in risk assessments undertaken as a part of public policy and decision making. If the analyst's subjective knowledge is important, why shouldn't this knowledge be published in the scientific literature where it can undergo the review process to regularize it for use in assessments?

Some Bayesians suggest that the issue of subjectivity is overblown because the numerical impact of the prior on the posterior can be minimized relative to the likelihood, and in any case it diminishes with the accumulation of data over time. This is perhaps the weakest apology the Bayesians offer for their use of subjectivity. The original prior for the quantity of interest represents a theory about it. In principle, the Bayesian approach keeps this theory forever in that it is only modified, and never simply replaced. Given the zero preservation problem and the possible inertia of the prior with respect to new data, this could be a serious concern. Some have complained that never replacing the theory is an unscientific feature of the Bayesian approach. This is true but largely unimportant, just because individual analysts don't live forever. The renewals of new generations of analysts, not to mention common sense, are fully capable of keeping the accumulation of knowledge on track. The concern should not be about the long term, but about the short term. Bayesian argue that the importance of the prior and its quantitative effect on the posteriors generally lessens as more and more evidence accumulates. Whether this is true in practice depends not only and the idiosyncrasy of the original priors, but on the nature and quantity of the incoming information. Claims that the prior is soon overwhelmed by data may generally be false in the context of risk analysis. The convergence arguments would have merit if risk analysis problems actually were accumulating data and routinely applying Bayes' rule to incorporate them. But risk analysis is intrinsically characterized by its paucity of data. However grave the underlying decisions, new data are not likely to be collected. No new information is coming that could stabilize the posterior (at least within the time horizon that would be relevant).

Critics (Edwards 1972; Hamming 1991) argue that, because the choice of the prior *matters*, the intrusion of subjectivity is important. Bayesians turn this argument around and insist that every non-trivial model has subjectivity. Subjective decisions include the structural form of the model, the dependence assumptions it includes, and distributional assumptions. Bayesians suggest that choosing the prior distribution is simply another component of the model that includes subjectivity. Bayesians assert that they are simply being honest about the importance of subjectivity in the modeling endeavor. Because their methods are formal, the influence of subjectivity can be tracked and monitored, which would be much harder to do in a non-Bayesian approach that cobbles together ideas of varying degrees of objectivity and subjectivity.

Determining the likelihood function

Most critics of Bayesian methods focus on the difficulty of selecting the prior distribution, but choosing the likelihood function may often be even more problematic. In some cases, telling risk analysts to "just specify the likelihood function" is a bit like telling the homeless to "just get a house". Although many strategies have been proposed to help analysts to select prior distributions, there is precious little guidance on the formulation of the likelihood function. This is a genuine problem not just because getting the likelihood is half the battle, but also because its derivation can often require a good measure of mathematical sophistication. The relevance and appropriateness of a probability model that justifies a particular likelihood function is always at best a matter of opinion (Edwards 1972, page 51). If there are data at all in a given situation, the risk analyst may have no clear idea about what sampling process would be appropriate to model their distribution. Yet the standard Bayesian approach assumes that the likelihood function can be specified precisely and the analyst knows the distribution from which sample data are drawn. This assumption is arguably the weakest because such knowledge may or may not be available.

Because selecting the likelihood appropriate for the analyst's unique case requires thought, several paradigmatic cases have become popular. These cases exploit lucky special cases in which the prior and the likelihood dovetail together conveniently such that there is no need to carry out integration to compute the normalization factor. In each line of the list below, the likelihood (not explicitly listed here) that is determined by the indicated sampling model will combine with the indicated prior to produce the corresponding posterior distribution. For these pairs of sampling model and prior, formulaic* updating rules permit the immediate specification of the posterior's parameters from those of the prior and statistics from the data. (Parameters of the sampling models not designated by θ are assumed to be known.)

Sampling model	*Prior*	*Posterior*
bernoulli(θ)	uniform(0, 1)	beta($1+\Sigma x_i$,$1+n-\Sigma x_i$)
bernoulli(θ)	beta(α, β)	beta($\alpha+\Sigma x_i$, $\beta+n-\Sigma x_i$)
poisson(θ)	gamma(α, β)	gamma($\alpha+\Sigma x_i$,$\beta+n$)
normal(θ, s)	normal(μ, σ)	normal($(s^2\mu+\sigma^2\Sigma x_i)/v$,$\surd((s^2+\sigma^2)/v)$), $v=s^2+n\sigma^2$
exponential(θ)	gamma(α, β)	gamma($\alpha+n$,$\beta+\Sigma x_i$)
binomial(k, θ)	beta(α, β)	beta
uniform(0, θ)	pareto(b, c)	pareto
negativebinomial	beta(α, β)	beta
normal(θ, s)	gamma(α, β)	gamma
exponential(θ)	inversegamma(α, β)	inversegamma
multinomial(k, θ_j)	dirichlet(s, t_j)	dirichlet($n+s$,$(x_j+st_j)/(n+s)$), $n=\Sigma x_j$, $j\in\{1,\ldots,k\}$

For the assumptions underlying the use of these conjugate pairs and details on exactly how the calculations are to be made, consult standard references on Bayesian methods (e.g., Lee 1997; Sander and Badoux 1991; Berger 1985; DeGroot 1970; Gelman et al. 1995).

Naturally, the existence of these conjugate pairs greatly simplifies the demands of applying Bayes' rule and are widely used for the sake of convenience. In some cases, it is clearly reasonable to fashion a probability model in which the analyst's prior happens to be one of the conjugate families, and the data come from a sampling model corresponding to the matching likelihood function. But this happy circumstance can hardly be assured in practice. The available

*The formulations use the syntax of RAMAS Risk Calc (Ferson 2002).

conjugate pairs are very restricted in scope and most Bayesians (as well as their critics) deprecate the use of the conjugate families if it is justified purely by mathematical convenience. Some Bayesians (e.g., Lee 1989, page 66) suggest nevertheless that a conjugate pair could be employed if the distribution and function can be selected to be "sufficiently close" to the actual ones.

Empirical likelihood. Owen (2001) describes a nonparametric way to generate likelihoods from sampling data. This approach, called empirical likelihood, could be an especially useful approach in risk analysis because there might be no way for an analyst to know the distribution for a newly encountered set of data, and there may be no reason to suppose in the first place that the distribution of new data belongs to any of the distribution families that have been well studied. Empirical likelihood uses the data themselves to determine the shape of the likelihood ratio function, without requiring the user to specify the family of distributions for the data. This means that, in statistical applications, the validity of the inference does not depend on the analyst's specifying a parametric model for the data. Because it is a likelihood method, it can be used to pool information from different data sources, or from exogenous constraints. The method may be extended to account for censored, truncated and biased sampling. Different approaches could be used to develop empirical likelihood functions, including, for example, the use of kernel techniques and smoothing operations. The behavior of empirical likelihoods with Bayes' rule has not yet been well studied (Owen 2001, page 188), but care must be taken to prevent the empirical likelihood from being zero for values beyond the observed range of the data, for the reason discussed in the next section. The most important practical limitation of the approach in risk analysis will no doubt be the fact that empirical likelihood is an asymptotic theory. It depends on there being plenty of sample data, and, although this situation may occur in some risk analysis problems, it is likely to be rare.

Zero preservation problem

Any event for which the prior probability is given as zero will always have zero posterior probability, no matter how large its likelihood might be. Any values on the real line for which the prior distribution is surely zero will remain with zero probability in the posterior, no matter what the likelihood is and no matter what new data may arrive. Likewise, whenever the likelihood is zero, the posterior will be zero too, no matter how probable the prior said it was. These conclusions are consequences of the simple fact that any number multiplied by zero is zero. They imply that an analysis that uses a prior that categorically denies certain possibilities can never learn from new evidence anything to the contrary. Bayesians do not consider this issue to be much of a problem in practice; they are careful to advise students to employ nonzero priors for all events and values that are remotely possible.

Zero preservation is the extreme case of a more general problem of Bayesian methods having to do with their possible insensitivity to surprise in the form of unanticipated data (Hammitt 1995). For instance, consider the numerical example depicted in Figure 2. In this case, the posterior can substantially disagree with both the prior and the new data (Clemen and Winkler 1999). Analysts might prefer fidelity to one or the other when they are in conflict. Or they might prefer that the result at least not be more certain than the prior when such disagreements between expectation and evidence arise. Despite the apparently surprising nature of the evidence embodied in the likelihood function compared to the prior, the posterior distribution is tighter (that is, has smaller variance) than the prior distribution. In the case of extreme surprise such as this, one might hope that a methodology would always yield a result that represented more uncertainty, not less.

Computational difficulty

Bayesian methods are notoriously computationally difficult. The difficulty usually arises in the calculation of the normalizing factor in the denominator of Bayes' rule. In particular, there is usually no closed-form solution available for computing the integral, unless the prior and likelihood happen to constitute a conjugate pair (see page 29) for which the analytical details work out nicely. But using conjugate pairs can remove the computational difficulty only if the conjugate prior is a reasonable reflection of the analyst's personal prior. In non-conjugate problems, several methods have been advanced for computing results. Among these are non-interactive Monte Carlo methods including direct sampling techniques, and indirect methods such as importance sampling, rejection sampling and weighted bootstrap sampling. There are also Markov Chain Monte Carlo (MCMC) techniques, such as substitution sampling and data augmentation, Gibbs sampling, and the Metropolis-Hastings algorithms (Metropolis et al. 1953; Hastings 1970) which are general methods for optimizing functions under certain constraints. The essential feature of MCMC is that the transition kernel of the chain is constructed so that the steady-state distribution that it eventually reaches is precisely the posterior distribution resulting from conditioning on all the evidence. In other words, the MCMC process eventually gives samples drawn from the joint posterior probability distribution function of the model variables. This distribution is the same one that would be achieved by applying Bayes' rule directly, which would usually be computationally intractable otherwise.

In additional to their intrinsic computational complexity, there is another layer of intellectual difficulty associated with the application of Bayes' rule in practice. Part of the Bayesian mantra is that analysts should *think* about their problems individually. Part of the reason that there is no convenient software that makes Bayes' rule easy to apply in routine problems is that Bayesians believe that real problems are unique and do not admit facile or canned solutions. There simply is no formula or paradigm that can be applied to each situation to select the prior or fashion the likelihood function. Bayesians hold that these decisions are truly the special and inescapable responsibilities of the modeler/analyst. The Bayesian approach is not intended to yield a quick and dirty but adequate solution, and proponents seem to encourage the idea that the proper application of Bayesian methods requires admission to a priesthood of competence that might be beyond any particular analyst's abilities. Although it would not seem to be necessarily so, Bayesian methods have about them an aura of difficulty that goes beyond the computational difficulty described above. This aura is similar to, but a more focused version of, the general air of difficulty that has historically surrounded Western mathematics in general since the time of the Greeks (Kline 1980). Naturally, this creates a barrier for neophyte users and represents a notable disadvantage of the method, although it may be an advantage in other respects if it tends to make outsiders regard a Bayesian solution with a bit more respect.

Bayesian model averaging masks uncertainty

The numerical example on page 13 involved the use of Bayesian model averaging (Draper 1995; Raftery et al. 1997; Hoeting et al. 1999; Clyde and George 2003) to represent model uncertainty. The approach involved "Bayes factors" which are the posterior probabilities for each of the two models (addition and multiplication) used as weights in a mixture model that combined the distributions obtain under these two models. The posterior probabilities were obtained by updating subjective priors to account for an observed data point. The advantages of Bayesian model averaging include the fact that it acknowledges the existence of model uncertainty which has traditionally been ignored in formal assessments. It can express the analyst's beliefs about the relative likelihoods of the different models and it can take account of relevant data that might help choose between the models.

Bayesian model averaging also has several significant disadvantages. Perhaps its greatest limitation is that the analyst must be able to enumerate all the possible models. This may be possible in some situations, but in general this can be an insurmountable task in itself. Some forms of model uncertainty resist enumeration. For instance, when model uncertainty includes the choice of what parameters to use, the choices about the level of abstraction, and the depth of detail to incorporate into the model, the space of possible models may be too complex for the methods discussed here. Once the possible models are listed, the analyst must assign prior probabilities to each possible model. Again, this can be difficult in practice. When the number of models is large, some sort of equiprobability scheme is commonly used. The calculation of the Bayes factors may involve a complicated integral which can be computationally burdensome.

These disadvantages are obstacles to the practical use of Bayesian model averaging that analysts would try to overcome them if the result were sufficiently valuable, but it is not entirely clear that Bayesian model averaging produces results that are useful in risk analysis problems. The primary objection to the approach concerns its use of a stochastic mixture. Using a stochastic mixture does not seem at all reasonable if the model uncertainty exists because we don't know which model is correct, as opposed to there being multiple models in some kind of timeshare situation. The way Bayesian model averaging employs the mixture would be reasonable if it existed because of variability. If, for instance, there was some switching mechanism that caused the function f to sometimes behave like an additive operation and sometimes like a multiplicative one, then some of the random values would come from the green distribution in Figure 3 and some would come from the orange distribution. However, if the model uncertainty is due to scientific ignorance, measurement error or similar epistemic reasons—and when we call it 'model uncertainty' it almost always does—then a mixture model creates a Frankenstein distribution that is consistent with neither model. Finkel (1995) reviewed the problems with averaging together incompatible models. The greatest problem is that it tends to mask the extant uncertainty. If we don't know which model is correct, then we have real uncertainty about how many values of the random variable $f(A,B)$ will be negative. If the multiplicative model is correct, fully half of the values will be negative. If the additive model is correct, hardly any values will be negative. The mixture model suggests, perversely, that about 12% of the values will be negative. There is a huge difference between an interval [0%, 50%] and the point 12%. The former expresses the uncertainty we have, and latter does not. Combining the models as a mixture has erased the uncertainty.

If Bayesian model averaging were applied to the study of distributions (rather than estimating point values) in risk analysis, it could substantially underestimate the true tail risks in an assessment. Its outputs are distributions that no theories for any of the models would consider reasonable. Some probabilists maintain that one can use stochastic mixtures to represent model uncertainty and that this does *not* average alternative models so long as the results are presented properly. It is hard for us to see how this is a tenable position if we want to be able to interpret the probabilities in the output of a quantitative risk analysis as frequencies.

Note that robust Bayes methods (see page 38) provide an entirely different approach that can be used to address model uncertainty about the shape of the prior distribution or the likelihood function.

Structural omniscience

Bayesian methods, and to be fair, most all of statistics and scientific modeling, presume that the analyst has a profound understanding of the *structure* of the systems being modeled. The

Bayesian requirement for structural omniscience seems particularly severe in some respects. It presumes an analyst can examine his or her subjective preconceptions in minute detail, even though the analyst may never have encountered or even thought about the problem before. It presumes that the analyst can compose a likelihood function that depends on a sampling model for data, even though no data have yet been collected. Previous sections have reviewed the omniscience needed for the prior and the likelihood. In this section, we consider the Bayesian presumption that the analyst knows the structure of the sample space, even before studying the problem or gathering samples of any kind.

In the numerical example on page 16 involving the bag of marbles, we presumed that we knew how many colors of marbles there were in the bag. Although this might seem like a common and presumption that should be entirely innocuous, Walley (1996; Bernard 2003) showed that this value is critical to the entire analysis when sample data are sparse or nonexistent. What can you say about the probability of drawing a red marble if you don't even know how many colors k there are in the bag? For instance, if we let the sample space Ω = {red, not red}, we would get a different answer than we would if we let Ω = {red, blue, green}, and different still if we let Ω = {red, blue, green, yellow}. There are obviously *lots* of ways to form Ω, and therefore there are lots of different answers a Bayesian could come up with. We cannot look inside the bag, and we have only limited samples from it, or maybe even no samples yet. If we tried to form Ω using the available data, we'd run into logical trouble and the strategy would be incoherent. What should we do? It seems clear that attempting to use subjective judgment in a situation like this for something as fundamental as the structure of the problem would be hardly better than naked guessing. Walley (pers. comm.) noted that this criticism was leveled against Bayesians almost a century ago, and that interest in the Bayesian approach generally waned in the wake of the criticism. Bayesians have never adequately addressed the problem, and it seems there really is no solution to it within their framework. Indeed, the revival of Bayesianism over the last few decades seems dependent on *not bringing up the subject*.

Walley's (1996) solution to this problem uses robust Bayes methods (see page 38) and involves his imprecise Dirichlet model (IDM). In this model, the likelihood is multinomial, as in the ordinary Bayesian formulation. The sampling model is pretty clear, so this part seems completely reasonable. But instead of selecting a particular prior probability distribution, we can model the prior as the class of *all* Dirichlet distributions $\mathcal{M}_0^s = \{F : F = \text{Dirichlet}(s, t_j)\}$. For this model, the value of $s > 0$ has to be fixed to ensure coherence, but the value does not depend on Ω. If the prior is a collection of possible priors, the posterior would likewise be a collection of posteriors, each formed by multiplying the likelihood by one of the possible priors. The corresponding posterior class is $\mathcal{M}_N^s = \{F : F = \text{Dirichlet}(N+s, (n_j+st_j)/(N+s))\}$. Under this family of posteriors, the predictive probability that the next observation will be color ω_j is the posterior mean of θ_j which is $(n_j + st_j)/(N + s)$. Now, we don't know what the values t_j should be, because that knowledge depends on knowing what Ω is, but we can still bound this quotient by considering the largest and smallest possible values the t_j could take. Because they're means of probabilities, they can range anywhere on the interval between zero and one. If we let $t_j \to 0$, we'd get

$$\underline{\mathrm{P}} = n_j / (N + s)$$

As the lower bound on the predictive probability. And if we let $t_j \to 1$, we get

$$\overline{\mathrm{P}} = (n_j + s) / (N + s)$$

as its upper bound. The width of the interval decreases as sample size N increases, but it *doesn't depend* on k or the sample space Ω at all. In the limit where the number of marbles we've sampled becomes very large, this interval and the Bayesian and frequentist answers all converge to the observed frequency n_j/N of how many marbles were red of those we saw. This approach works even if there's no or little data. Before any observations, $\mathrm{P} = [\underline{\mathrm{P}}, \overline{\mathrm{P}}] = [0,1]$. The answer is vacuous, but—given there are *no data*—it seems it ought to be vacuous. It not only works in the sense of mechanically producing an answer, but it works in the sense of giving an answer that makes sense and is justifiable beyond subjective judgments and guessing.

The analysis above assumed s was some positive number. But what should s be? The value will determine the prior's influence on posterior. As s approaches zero, the interval will get narrower and narrower. We might expect such results to be overconfident. As s gets larger, the interval approaches the vacuous interval. The value used for s determines how educable we are with respect to the incoming data and how reliable we want the estimate to be. Setting it low lets you learn fast (at the cost of being wrong more often); setting it high makes your bounds more reliable (at the cost of needing more data to get tight intervals). Walley (1996) suggests that using $s = 1$ or 2 might be reasonable, but cautions that it would be incoherent to let it vary or let it depend on N. Letting s be 2 or larger will produce intervals that encompass both frequentist and standard objective Bayesian answers.

Notice that different values of s in Walley's IDM give *consistent* inferences that are nested. For instance, the interval for s=1 will always be inside the interval for s=2. In contrast, different Bayesian calculations conducted by different analysts with different priors and perhaps different conceptions of the sample space will lead to answers that are merely different and therefore *inconsistent*. By simultaneously requiring analysts to posit the structure of the problem in more detail than is available and producing results that depend rather sensitively on these decisions, Bayesian methods offer an empty promise of utility in real-world problems. Without the knowledge necessary to choose which assumptions to make, analysts will surely obtain different and therefore unreliable answers. What is the purpose of computing such answers if they reflect details about the analyst more than they reflect the reality of the problem analyzed?

Numerical example

Suppose the sample data are

blue, green, blue, blue, red, green, red,…, (35 red out of 100),…, (341 red out of 1000),

and that they arrived in this order. After seeing each datum, we can compute the interval bounds on the predictive probability that the next marble will be red as

$$[\, \underline{\mathrm{P}},\ \overline{\mathrm{P}}\,] = [\, n_{\mathrm{red}} / (N+s),\ \ (n_{\mathrm{red}}+s) / (N+s)\,].$$

Note that this interval always includes the observed frequency n_{red}/N. We can also compute the probability before the first marble is drawn. The results are listed in this table:

n_{red}	N	$\underline{\mathrm{P}}$	$\overline{\mathrm{P}}$
0	0	0	1
0	1	0	0.5
0	2	0	0.334
0	3	0	0.25
0	4	0	0.2

1	5	0.166	0.334
1	6	0.142	0.286
2	7	0.25	0.375
⋮	⋮		
35	100	0.346	0.357
⋮	⋮		
341	1000	0.340	0.342

The probabilities are graphed in Figure 6 as a function of sample size. The squares mark the upper bounds, and the diamonds mark the lower bounds. The bounds shown in the graph were computed using $s = 1$, but s could be any positive number. If s is a very small value, the interval gets very tight very fast around the observed frequency of red marbles. If s is a very large number, the interval stays wide, close to the vacuous interval [0,1], for a long time while marbles are drawn from the bag and observed.

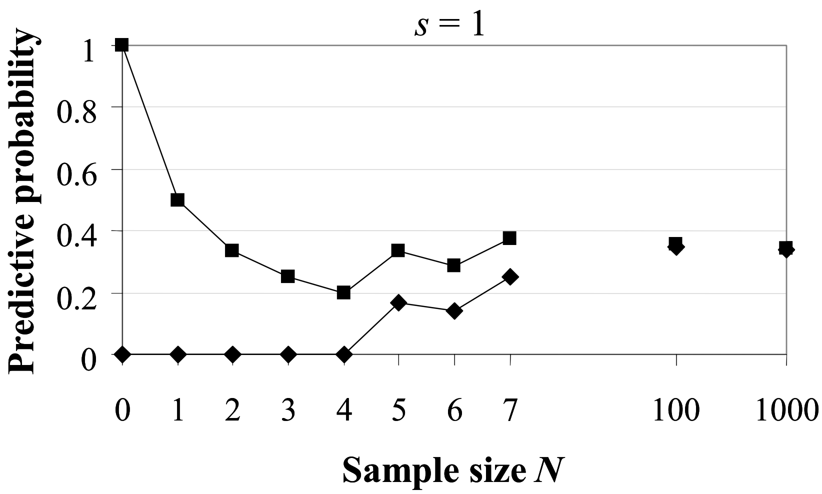


Figure 6. Surety of the predictive probability of drawing a red marble as sample size grows.

Bayes' rule says more than is justified

Several examples have already been given in which the standard Bayesian approach produces results that seem to say more, that is, have stronger inferences, than can really be justified given the limited information available to the analyst. For instance, in the paternity example, we saw that the standard approach (using a prior of ½) resulted in precise estimate of the posterior probability the man was the girl's father, but other priors could have produced posterior probabilities anywhere between zero and one. When the blood testing is not dispositive, the Bayesian calculation purports to tell us what the data imply about the probability of a hypothesis (that the man is the father), when all that is arguably germane in the case is what the evidence says about this hypothesis. We don't need Bayes' rule to tell us that. When we use Bayes' rule, we get more that is justified because the prior is specific even though it embodies no real information. In contrast, in the disease testing example the prior was based on evidence about the prevalence of the disease in the population.

In the example involving model uncertainty, Bayesian model averaging masked the fact that the *frequency* of negative values of $f(A,B)$ could be much higher than the computed *probability* of 12%. This problem is mostly the consequence of the use of a mixture to aggregate the competing models which averages together the distributions. In general, averaging erases variation, and this can be a problem in risk analysis if it makes the analysis seem to say more than is actually justified by evidence or assumption. Whenever Bayesians reduce distributions down to their means, they inadvertently convey precision that may not be present. If a risk analysis reveals that a flood could rise water levels between 1 and 3 meters, it is certainly inappropriate—both scientifically and as a matter of protectiveness—to report that the flood will be 2 meters.

The example of the bag of marbles demonstrated that using Bayesian methods demands structural knowledge, without which no answer can even be computed. If the analyst does not posses that knowledge, the answers produced by the analysis cannot be justified except as hypotheses or perhaps wishful thinking.

In the numerical example on page 14 involving the height, width and area constraints, the Bayesian solution was a precise distribution for the three variables. It is perhaps intuitively clear that this precise distribution is really a reflection of the uniform prior that we used, and nothing in the statement of the problem justified this selection. If we had picked another prior distribution, subject to the constraints specified in the problem, namely, that its support lie inside the box given by the three intervals, we would have obtained a different distribution for the posterior. In fact, the precision of the Bayes results is entirely due to the prior we used. If we used all possible priors consistent with the constraints, the *posteriors would fill up the gray intervals* in Figure 4. Because nothing in the problem justified the use of the uniform prior and the answer depends on that choice, the Bayesian answer says more than can legitimately be concluded. From the perspective of a risk analyst, the Bayesian answers touted for their precision and detail are seen in this light to simply be wrong. They are not justified by the knowledge stated in the problem and express less uncertainty than is actually present.

Elementary constraint analysis can be used to infer that the limits on the variables are actually just

$$W = [23, 28.57],$$
$$H = [112, 139.13] \text{ and}$$
$$A = [2576, 3200].$$

All one can conclude from the information given in the problem is that the probability distributions for these three variables must have supports that are entirely within this cube. The formulas needed to obtain the updated constraints are trivial. For instance, the formula for the width is

$$W = W \cap (A \,/\, H)$$

where $\cap$ denotes intersection and the / denotes interval division (see page 38). In terms of real numbers, to get the largest possible W, for instance, let A be as large as possible and H as small as possible, and solve for $W = A/H = 3200/112 = 28.57$. The answers are easy to compute and they comprehensively characterize the uncertainty about the constrained variables. Constraint analysis can also be employed when more is known about the inputs than their ranges. For instance, straightforward methods are available for the case when the inputs are characterized by p-boxes (Ferson et al. 2003). These methods are also inexpensive to compute and likewise properly express the true uncertainty in the problem. In contrast, the Bayes solution to this updating problem gives the wrong answers, and they're difficult to compute. (One is reminded of Groucho Marx's restaurant review, "the food is bad and the portions are small".)

Tony O'Hagan (2002) has argued that examples like this are unfair because the problem itself is wrong in the sense that it suggests it is possible that we might know only bounds on the height, width and area variables. He argued that in real situations we would also know more about the variables than mere their ranges, and that a proper analysis would elicit this information and use it to construct justifiable priors. It does not seem tenable to claim, however, that just enough information always exists in a real-world setting to specify a prior precisely. Certainly we should use whatever information is available, but an analyst may or may have particular information in any given situation. Furthermore, it seems that a method proposed for use in risk analysis ought to be able to handle a problem like this well, even if it is rarely encountered in practice.

Bayesian rationality does not extend to group decisions

Bayesian methods are regarded by many analysts as the best (or even the only coherent) way to represent and manipulate uncertainty in inference and decision making (Jeffreys 1961; Lindley 1965; de Finetti 1970; Bernardo and Smith 1994; Berry 1996; Malakoff 1999; Jaffe 2003; Jaynes 2003). It seems clear, however, that the touted advantages of Bayesian methods collectively called "rationality" do not generally extend to the context of group decision making. Various paradoxes have been encountered when there is more than one person in the group whose beliefs and preferences must be taken into account. For instance, when Seidenfeld and his colleagues (1990; Seidenfeld et al. 1989) explored the properties of the Bayesian approach in the context of group decisions, they found "dilation", a highly counterintuitive situation in which the overall uncertainty can, in some situations, be guaranteed to increase after collecting empirical data *no matter what data values are observed.* They realized that the only way to sustain rationality in the group decision setting was to reduce the problem to an individual decision problem by requiring a dictator make all the decisions for the group. Seidenfeld suggests that we can choose between rationality and democracy, but we cannot have both. Mongin (1995) similarly showed that, under certain conditions, Bayesians cannot aggregate the preferences of multiple rational decision makers in a way that simultaneously satisfies the Bayesian axioms and the Pareto property (that is, whenever all decision makers agree about a preference, the aggregation result also has that preference).

Interestingly, when decision makers are told that the methods cannot ensure rationality, they don't seem to mind very much. Kathy Laskey (pers. comm.) has suggested that the dogmatic Bayesian who insists on coherence at all costs is simply bad decision theorist. She argues that it is not reasonable to care more about coherence than about getting a sensible answer to the problem at hand. She argues that applications of probability theory should be judged by how well they help us solve the problem they were designed to solve. Sagoff (2004) has suggested that the purpose of risk analysis and scientific modeling in general is not to discover the truth about the world anyway. Rather, they are simply helpful tools of mediation to settle disputes about what will be regulated and what will be allowed, and who will pay for what. Formal quantitative assessments are certainty not trusted to make the decisions for us; they only help us to make our assumptions clear. The suggestion in the introduction that the Bayesian approach could be used to frame and make decision seems naïve in view of the assumption pandering, horse trading, and general posing and politicking that characterize how risk assessments are actually used.

Fixing the Bayesian paradigm: robust Bayes

Robust Bayes analysis, also called Bayesian sensitivity analysis, investigates the robustness of answers from a Bayesian analysis to uncertainty about the precise details of the analysis (Berger 1985; 1994; Insua and Ruggeri 2000; Pericchi 2000). An answer is robust if it does not depend sensitively on the assumptions and calculation inputs on which it's based. Robust Bayes methods acknowledge that it is sometimes very difficult to come up with precise distributions to be used as priors. Likewise the appropriate likelihood function that should be used for a particular problem may also be in doubt. In a robust Bayes approach, a standard Bayesian analysis is applied to all possible combinations of prior distributions and likelihood functions selected from *classes* of priors and likelihoods considered empirically plausible by the analyst. In this approach, a class of priors and a class of likelihoods together imply a class of posteriors by pairwise combination through Bayes' rule. Robust Bayes also uses a similar strategy to combine a class of probability models with a class of utility functions to infer a class of decisions, any of which might be the answer given the uncertainty about best probability model and utility function. In both cases, the result is said to be robust if it's approximately the same for each such pair. If the answers differ substantially, then their range is taken as an expression of how much (or how little) can be confidently inferred from the analysis.

Although robust Bayes methods are clearly inconsistent with Bayesian idea that uncertainty should be measured by a single additive probability measure and that personal attitudes and values should always be measured by a precise utility function, they are often accepted as a matter of convenience (e.g., because the cost or schedule do not allow the more painstaking effort needed to get a precise measure and function). Some analysts also suggest that robust methods extend the traditional Bayesian approach by recognizing incertitude as of a different kind of uncertainty (Walley 1991; Pericchi 2000). Analysts in the latter category often quip that the set of distributions in the prior class is not a class of reasonable priors, but that it is reasonable class of priors. The idea behind this riddle is that no single distribution is reasonable as a model of ignorance, but considered as a whole, the class is a reasonable model for ignorance.

Robust Bayes methods are related to important and seminal ideas in other areas of statistics such as robust and resistance estimators (Huber 1981; Agostinelli 2000). The arguments in favor of a robust approach from those areas can often be applied without modification to Bayesian analyses. For example, Huber (1981) criticized methods that must assume the analyst is "omniscient" about certain facts such as model structure, distribution shapes and parameters. Because such facts are themselves potentially in doubt, an approach that does not rely too sensitively on the analysts getting the details exactly right would be preferred.

There are several ways to design and conduct a robust Bayes analysis, including the use of (i) parametric conjugate families, (ii) parametric but non-conjugate families, (iii) density-ratio (bounded density distributions), (iv) ε-contamination, mixture, quantile classes, etc. and (v) bounds on cumulative distributions. Although the problem of computing the solutions to robust Bayes problems can, in some cases, be computationally intensive, there are several special cases in which the requisite calculations are (or can be made) straightforward. The following subsections review a few of these cases.

Intervalizing Bayes' rule for events

The section starting on page 7 considered the application of Bayes' rule for events and offered a numerical example of inferring the chance one has a disease from a positive test for the disease.

This section shows how this application can be extended to the case in which the various probabilities used as inputs in the calculation are generalized to intervals rather than point estimates. The formulation requires us to define addition, multiplication, subtraction and division for intervals. When the operands represent probabilities, these rules are

$$x + y = [x_1, x_2] + [y_1, y_2] = [x_1 + y_1,\ x_2 + y_2],$$
$$x - y = [x_1, x_2] - [y_1, y_2] = [x_1 - y_2,\ x_2 - y_1],$$
$$x \times y = [x_1, x_2] \times [y_1, y_2] = [x_1 \times y_1,\ x_2 \times y_2], \text{ and}$$
$$x \div y = [x_1, x_2] \div [y_1, y_2] = [x_1 \div y_2,\ x_2 \div y_1].$$

For example, if $x = [0.3, 0.5]$ and $y = [0.1, 0.2]$, then $x + y = [0.4, 0.7]$ and $x - y = [0.1, 0.4]$. Notice that in subtraction and division, the opposite endpoints are combined together. If one of the operands is a precise real value s, it can be converted into the corresponding degenerate interval $[s, s]$ for use in these formulas. One should keep in mind that the rules for multiplying and dividing interval probabilities are much simpler than they are for general intervals. The reason is that probabilities are constrained to range over [0,1], which simplifies things a bit.

Moore (1966) showed that interval expressions involving only single uses of each uncertain variable will yield best possible results when computed bu applying these rules. Notice, however, that there are repeated instances of intervals in the expression

$$\frac{\text{prevalence} \times \text{sensitivity}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}.$$

Because both prevalence and sensitivity appear multiple times in this expression, ordinary interval analysis cannot reliably be used to compute with this formulation.

If the inputs to Bayes' rule have uncertainty, then it would be preferable to use the formulation simplified to have no repeated parameters

$$1 / (1 + ((1/\text{prevalence} - 1) \times (1 - \text{specificity})) / \text{sensitivity}).$$

This formulation is algebraically equivalent to the formulas used on page 8, and it yields best possible (i.e., tightest) answers when its inputs are intervals.

Numerical example: disease testing revisited

Suppose that the test's sensitivity is 99.9% and that its specificity is 99.99%, but the prevalence of the disease in the population is imperfectly known and the best estimates say only that it is some value within the interval [0.005, 0.015]%. Putting these values into the expression given on page 8,

$$\frac{[0.005, 0.015]\% \times 99.9\%}{([0.005, 0.015]\% \times 99.9\% + (1 - [0.005, 0.015]\%) \times (1 - 99.99\%))},$$

and using interval analysis to compute the result, we would obtain [19.99, 99.95]%. This answer is correct in the sense that it encloses the probability that a person testing positive has the disease, but the interval is much too wide given the stated uncertainty about the disease's prevalence. The best possible answer comes from using the formulation

$$1 / (1 + ((1/\text{prevalence} - 1) \times (1 - \text{specificity})) / \text{sensitivity})$$

without repeated parameters which yields

$$1 / (1 + ((1/[0.005, 0.015]\% - 1) \times (1 - 99.99\%)) / 99.9\%)$$

and gives the tighter interval [33.31, 59.98]% as the bounds on the chance the positive testing person has the disease. This interval is as tight as is possible given the stated uncertainty about the prevalence (prior probability).

Trivial case of Bayes' rule on p-boxes

When robust Bayes methods are extended from events to random variables and distributions, a trivial but interesting case immediately arises. The example is instructive because it hints at a limitation of robust Bayes methods in general.

The term "p-box" denotes a pair of upper and lower bounds on cumulative distribution functions, and we identify a p-box with the class of distributions whose cumulative distributions lie entirely within these bounds (Ferson et al. 2003). P-boxes arise naturally in many situations, including from constraint specification in terms of moments or order statistics and as empirical distributions of interval data sets, and it is possible that one might wish to characterize the uncertainty about a prior distribution with a p-box. If the uncertainty about the likelihood function is also characterized by upper and lower bounds on its integral, these bounds can be normalized so that the likelihood also has the shape of a p-box. When two such p-boxes are combined in a robust Bayes analysis, the best-possible bounds on the cumulative distribution functions that might be the posterior is a rectangular p-box whose range is the intersection of the ranges of the prior and the likelihood p-boxes. This is illustrated in Figure 7. The reason that the posterior class is degenerate is easy to understand. As discussed in the section on zero preservation on page 30, whenever either the prior or the likelihood is zero, the posterior will be zero too. The p-box on the prior admits possible prior distributions like that depicted in Figure 8 where the cumulative distribution is flat for certain ranges of values of the abscissa. Of course such a shoulder in the distribution function corresponds to probability density that is identically zero. Similar shouldering can also occur in the likelihoods. Within a p-box there is a great deal of flexibility about where these shoulders might occur. To account for all possible members of the classes of priors and likelihoods, we have to account for pairs that effectively "miss" each other almost everywhere along the abscissa. Because any posterior that emerges from the analysis gets renormalized to have total probability of unity, if the prior and the likelihood miss each other everywhere except for one spot, then all the mass of the posterior will get concentrated at that spot. Where can such spots arise? Anywhere at all where the prior and the likelihood might overlap (this range is also determined by zero preservation). This means that we can get posteriors that are essentially Dirac delta functions anywhere within the interval where the supports of the class of priors and the class of likelihoods intersect. This is why the bounds on the posteriors form a degenerate p-box in the shape of a rectangular interval in Figure 7.

The conclusion of these considerations is that when the prior distribution and the likelihood function are characterized only by bounds on their cumulatives, then we cannot infer anything about the class of posteriors except bounds on its support, that is, the range over which it may have positive density. But we already knew this range; it is implied by elementary considerations about zero preservation. We haven't learned anything at all about the posteriors from the robust Bayes analysis in this case. The shapes of the two p-boxes don't matter at all.

The only thing that does matter is their intersection. This conclusion seems very unsatisfying. It is the result, of course, of saying very little about the prior and the likelihood. Although the graphs of p-boxes that characterize them would seem to suggest that we are saying a lot, the graphs are misleading in this situation because Bayes' rule involves probability *densities*, not cumulative probabilities.

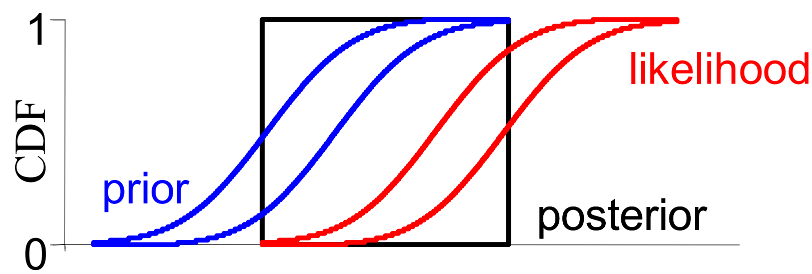


Figure 7. Robust Bayes analysis of a prior p-box (blue), a likelihood p-box (red), and the resulting posterior p-box (black).

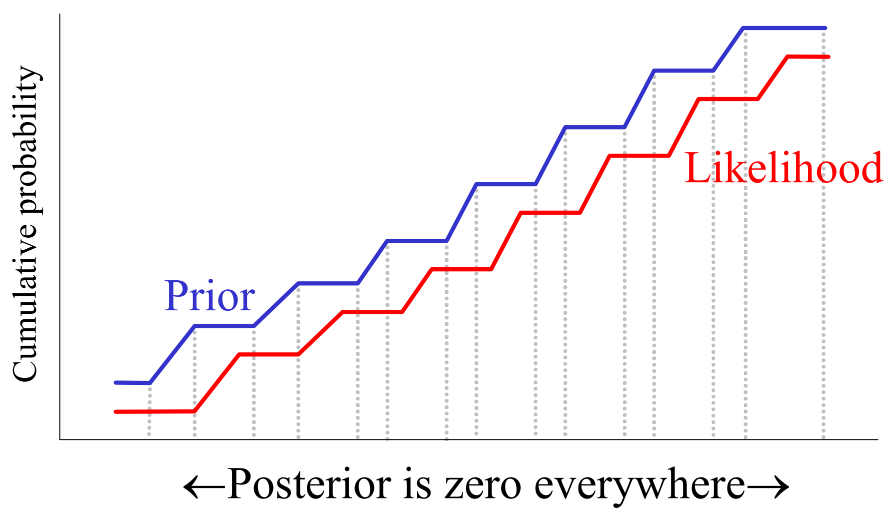


Figure 8. Shouldering, by which posterior distributions can be zero anywhere.

Parametric classes

Another way to define the class of distributions for a robust Bayes analysis to specify it parametrically. Figure 9 depicts caricatures of two such classes. The first set of green curves on the left of the graph represents a class of distributions all having the same shape but differing in central tendency. The other set of blue curves on the right side of the graph similarly represents a class of distributions that are all from the same distribution family and have the same mean, but different variances. One can easily imagine (although it would be a lot harder to draw) a set of distributions that varied in both mean and variance. The family of distributions indexed by the parameters might be normals or another shape from a conjugate pair, or it might some non-conjugate family. The class is intended to represent the analyst's uncertainty about the prior, so how the parameters vary will vary to define the class should reflect this uncertainty. Limits on the parameters can obviously be chosen to make the class arbitrarily *wide*, but the class can often still be too *sparse* in the sense that the distributions it admits are too similar to reflect the diversity of distribution functions the analyst might think are actually possible. For instance, I could set the limits on the mean and the variance to be very wide so that the prior is very uncertain and yet still be dissatisfied because the elements of the class are all vanilla normals. If I think that skewed, multimodal, truncated or non-continuous distributions should be in the class, I would need to change the parametric family or consider some non-parametric scheme to define the class. In a sense, the parametric approach to robust Bayes analysis has the opposite problem as that of the trivial example involving p-boxes in the previous section. In that case, the classes were too large to be able to conclude anything about important about the posterior. Here, the classes are too small to be really representative of the kind of uncertainty we might have about the prior and likelihood.

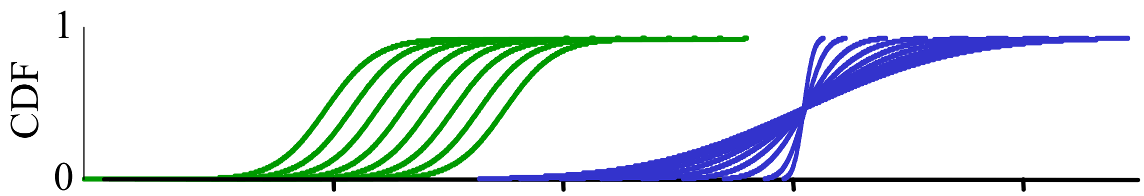


Figure 9. Parametric classes of distributions, varying the mean (green) or the variance (blue).

Numerical example

Suppose that the prior distribution is within the class of all normal distributions having a mean within [1, 6] and a variance within [2.5, 6]. Suppose further that the likelihood function is also characterized by a normal shape, with a mean in the interval [5, 13] and variance in the interval [4, 5]. In Figure 10, a few prior distributions (shown as blue curves) and likelihood functions (shown as red curves) from their respective classes are drawn on the θ-axis in terms of density (the vertical axis not shown). The purpose of the robust Bayes analysis is to compute the corresponding class of posterior distributions that are obtained by applying Bayes' rule to every possible pair of prior distribution and likelihood function. Because the priors and likelihoods in this example are conjugate pairs, it is easy in this example to compute the posteriors. When a prior is normal with a known variance v_1 and a likelihood is normal with a known variance v_2, the posterior distribution also is normal with variance $1/(1/v_1 + 1/v_2)$. In this case, the variance must lie within the range $1/(1/[2.5,6] + 1/[4,5]) = [1.53, 2.73]$. The posterior's mean is $(m_1/v_1 + m_2/v_2)/(1/v_1 + 1/v_2)$, where m_1 and m_2 are the means of the prior and likelihood respectively. This formulation has multiple instances of v_1 and v_2, so naïve interval analysis does not produce the best possible answer. A brute-force calculation strategy involving reconstitution of subintervals (Moore 1966) reveals the mean to be within [2.32, 10.22]. The posterior is thus a normal distribution whose mean is within [2.32, 10.22] and whose variance is within [1.53, 2.73]. A few representatives from this class of posteriors are also shown depicted as gray curves in Figure 10.

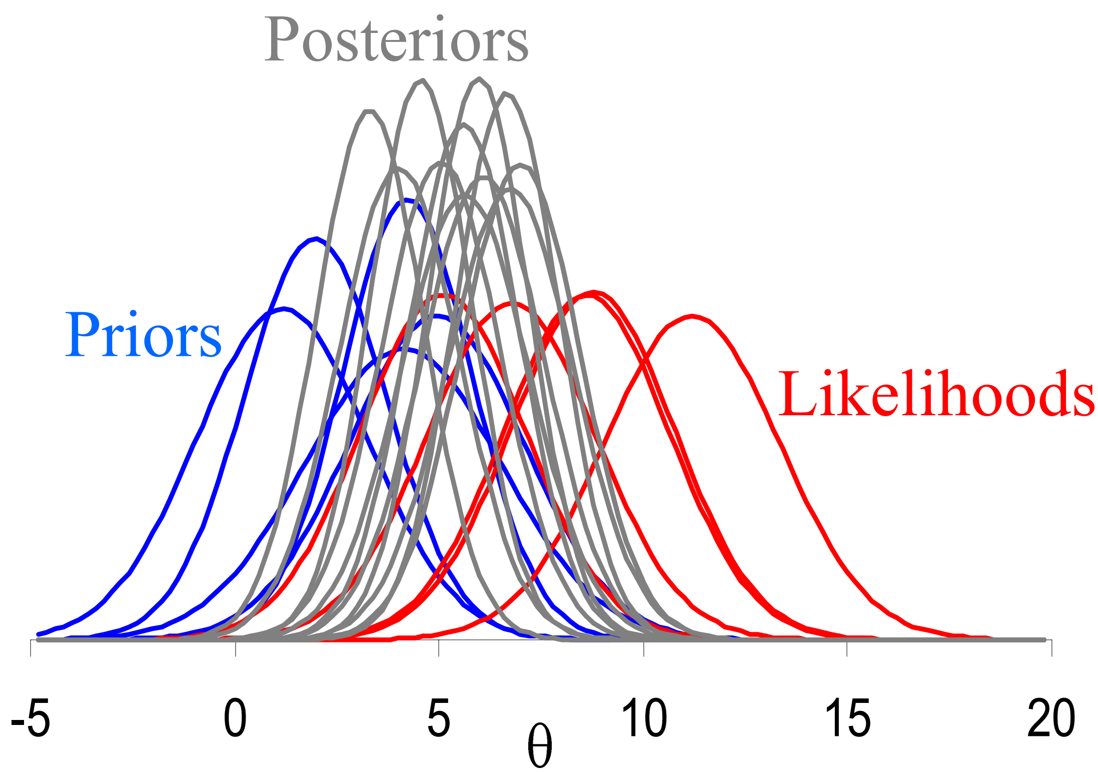


Figure 10. Example of robust Bayes combination of several prior distributions and likelihood functions to obtain many possible posterior distributions.

Bounded-density classes

It is also possible to bound the prior and likelihood in terms of probability density rather than the cumulative probability (Lavine 1991; cf. DeRobertis and Hartigan 1981). An example of this idea is illustrated in Figure 11 and Figure 12. The class of priors consists of all distributions whose curves lie entirely within the blue bounds. The class of likelihoods consists of any function whose graph is likewise bounded by the red curves. These bounds can easily be combined to obtain the implied bounds on the products that could result from any pairing of a prior and likelihood from the two classes. For any value of θ, there is an interval bounding the prior and another interval bounding the likelihood. These two intervals are multiplied together using interval arithmetic (page 38). The answer is an interval product for that value of θ. Such multiplications for all the possible values of θ produce a pair of upper and lower bounds for the class of unnormalized posteriors. These bounds are shown in Figure 11 as gray curves.

Every posterior must be normalized to have area one so that it is a true probability distribution. How can the bounds be transformed to reflect bounds on the normalized posteriors corresponding to any pair of prior and likelihood from their respective classes? It turns out the transformation is easy: the upper gray bound is transformed by dividing it by the area of the lower gray bound, and the lower gray bound is transformed by dividing it by the area of the upper gray bound. The result is depicted, now in black curves, in Figure 12. (The blue bounds on the prior and the red bounds on the likelihood are the same as in Figure 11.) Any function that lies entirely within these black bounds is a possible posterior given the bounds on the priors and the likelihoods.

It is possible to compute interval bounds on the mean of the posterior distribution from these bounds, although they will often be rather wide. The lower bound on the mean will be the mean associated with the left-most distribution that can be inscribed within the black bounds. This distribution will have part of its mass as characterized by the lower black bound. The remainder of its mass will concur with the left tail of the upper black bound. The upper bound on the posterior mean is associated with an analogous right-most distribution compatible with the black bounds. It is also possible, using a simple graphical algorithm, to compute interval bounds on the mode of the posterior distribution. In this case, the mean lies inside the interval [11.1, 19.6], and the mode is inside [5.9, 24.4]. These statistical bounds are best possible in the absence of other information besides the original bounds on the prior density and likelihood function. The bounded-density approach is invariant under any linear transformations of θ, although it is not so under arbitrary reparameterizations.

There are a few logical restrictions on the specification of the prior and likelihood bounded-density classes. For instance, the area under the prior's lower bound must be no larger than one, and the area under the upper bound must be no smaller than one. If either bound has area equal to one, then the prior is thereby perfectly determined. There are no restrictions on the likelihood bounds, except positivity.

If the classes formed by p-boxes (page 40) are too loose and large, and parametric classes (page 41) are too restrictive and small, then perhaps bounded-density classes will be just right for practical robust Bayes analysis. The analyst is the judge of whether the set of prior distributions and the set of likelihood functions defined by such bounds are reasonable characterizations of the uncertainty about them, and how wide the bounds need to be to reflect the uncertainty honestly. Pericchi and Walley (1991) suggested letting the prior's lower bound be obtained by elicitation and letting its upper bound be an uninformative distribution that touches the top of the lower bound.

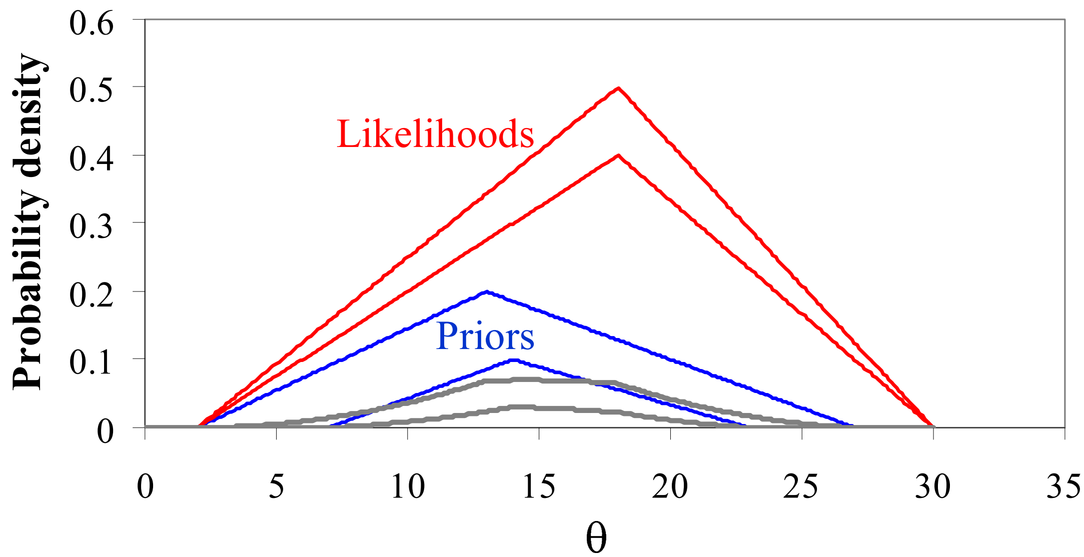


Figure 11. Bounded-density classes of priors (blue), likelihoods (red) and unnormalized posteriors (gray).

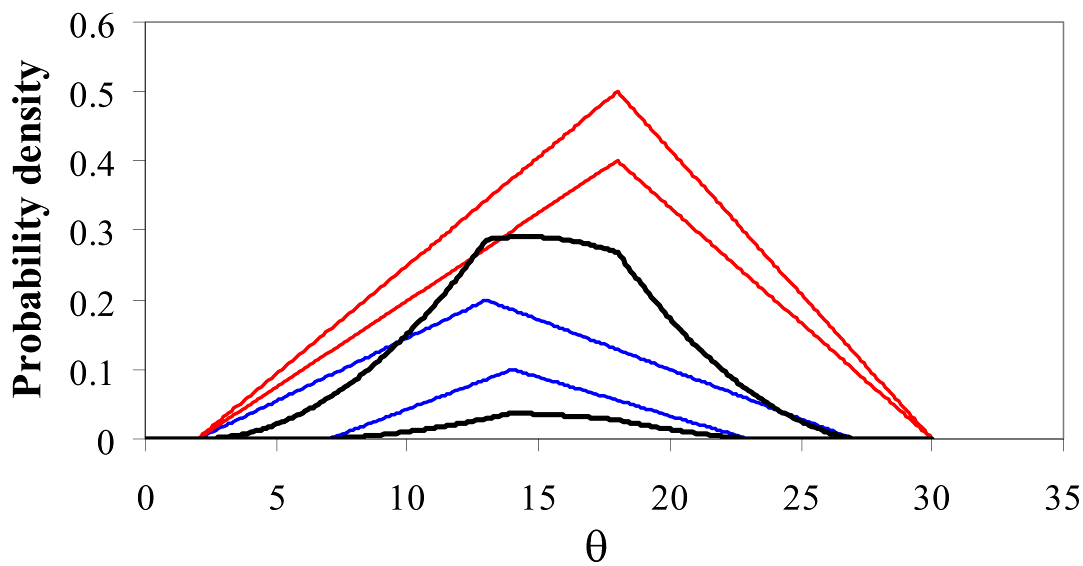


Figure 12. Bounds on the posteriors (black) arising from bounds on priors (blue) and likelihoods (red).

Other distribution classes

Pericchi (2000) outlined several other schemes for expressing uncertainty about prior distributions in robust Bayes analyses. Berger (1994) reviewed the desirable features of such distribution classes. He suggested they should be easy to understand and elicit, easy to compute with, sufficiently big to reflect one's uncertainty, and generalize easily to higher dimensions. The problem is to navigate between the devil of overspecifying the class (such as in the parametric class example on page 41) and the deep blue sea of underspecifying the class (such as the trivial example on page 40). In the former case, the analysis depends on fragile assumptions and the results may not really be robust, and in the latter case the analysis relaxed all the assumptions and could make no conclusions. Researchers are studying the problem of how to express situations of near-ignorance where the analyst's knowledge is vacuous in some but not all ways.

Caveats about robust Bayes

The concerns that attend the use of Bayesian methods about specifying a single prior and likelihood are relaxed by the use of classes of priors and likelihoods in robust Bayes, but this approach is still subject to some of the reservations and concerns associated with Bayesian methods in general. For instance, robust Bayes methods can be computationally difficult. Although the calculations in the numerical examples above could be done almost by inspection, this is not generally possible. In fact, the computational difficulties associated with robust Bayes methods can be severe. Depending on how the class of priors and the class of likelihoods are specified, the actual calculations required may be rather burdensome, and, in the general case, may be qualitatively more difficult than the calculations for an ordinary Bayesian calculations. Like ordinary Bayesian methods, robust Bayes continues to have the zero preservation problem. Zero preservation is the fact that any values for which the prior distribution is zero will also have zero probability in the posterior, no matter what the likelihood is and no matter what new data may arrive. In the case of robust Bayesian analysis, the zeros are those regions of the real line where *all* the prior distributions in the class are identically zero. Analysts must be cognizant of zero preservation and take care to avoid inadvertently or unthinkingly excluding possibilities at the outset. Finally, robust Bayes methods do not obey what Walley (1991) calls the Bayesian dogma of ideal precision. That is, they decline to condense all of the analyst's uncertainty into single precise probability distributions. This makes the approach controversial with some strict Bayesians.

Two-dimensional Monte Carlo simulation

Given the deficiencies of Bayesian methods outlined in this report, a risk analyst might elect to forego the approach entirely and rely instead on two-dimensional Monte Carlo simulation to account for both variability and incertitude in a risk assessment. This section outlines the advantages and disadvantages of this option.

A two-dimensional Monte Carlo (2MC) simulation is a nesting of two ordinary Monte Carlo simulations. By nesting one Monte Carlo simulation within another, analysts hope to discover how variability and incertitude interact to create risk. Typically, the inner simulation represents natural variability of the underlying physical and biological processes, while the outer simulation represents the analyst's incertitude about the particular parameters that should be used to specify inputs to the inner simulation. This structure means that each replicate in the outer simulation entails an entire Monte Carlo simulation which, obviously, can lead to a very large computational burden. The approach is similar to a sensitivity study, except that the combinations to be studied are determined not directly by the analyst, but as the result of statistical distributions on the values of the parameters. Two-dimensional Monte Carlo methods were championed for use in risk analysis by Hoffman and Hammonds (1994) and Helton (1994) among many others. Cullen and Frey (1999) give a good introduction to the techniques of the method, which has roots in Fisher (1957) and Good (1965). The risk analysis literature on 2MC has grown incredibly over the last decade, and the method is considered by many analysts to be the state of the art.

Although the mechanics and algorithms employed in 2MC simulations can be used in both Bayesian and frequentist analyses, the concepts behind 2MC simulation as it is typically employed in risk assessment are very different from those underpinning the Bayesian approach. Whereas Bayesian methods model the uncertainty about the parameters of the probabilistic model automatically by virtue of using distributions to represent them, frequentists had to invent 2MC simulation to do the analogous thing in a frequentist context. In this approach, *variability* is rigorously distinguished from *incertitude*. Variability is the form of uncertainty that arises because of natural stochasticity, variation or fluctuation in variables across space or through time, heterogeneity from or manufacturing errors or variation among components or genetic differences among organisms. Variability is sometimes called "aleatory uncertainty" (from *alea*, the Latin word for dice) to emphasize that it is the same kind of uncertainty as that exhibited in games of chance. Incertitude, on the other hand, is the lack or incompleteness of knowledge. Incertitude arises from limited sample sizes, mensurational limits ('measurement error'), data censoring, missing values, and the use of surrogate data such as using toxicity values for chickens to parameterize a model for hawks which are harder to study. Incertitude is sometimes called "epistemic uncertainty" to emphasize that it is a paucity of knowledge. In principle, incertitude can be reduced by additional empirical study. Although variability could be better characterized, it cannot be reduced by study, but would require some kind of management intervention to reduce it. Distinguishing variability and uncertainty from one another may not always be easy to do, and it may sometimes be like trying to divide ice from snow. But the focal intent of 2MC is nevertheless to make this distinction and to explore the consequence of each type of uncertainty on the resulting risk. In a strict Bayesian approach, in contrast, variability and incertitude may not be distinguished from each other. Both are forms of uncertainty, after all, and Bayesians believe that probability theory is a universal calculus for all manner of uncertainty.

In a 2MC simulation, parameters of input distributions can themselves be distributions too. Analysts can model uncertainty about any constants in the inner simulation with distributions in the outer simulation. Consult Hoffman and Hammonds (1994) or Cullen and Frey (1999) for advice on how to conduct the 2MC simulation. The results it produces are a collection of possible distributions for each output variable of interest (Figure 13). Typically, the collection consists of many distributions, indeed, it is a distribution whose elements are themselves distributions. This object can be called a "metadistribution" for short. Strict probabilists might suggest condensing the metadistribution into a single precise probability distribution, but most risk analysts prefer not to do this because it would confound variability with incertitude.

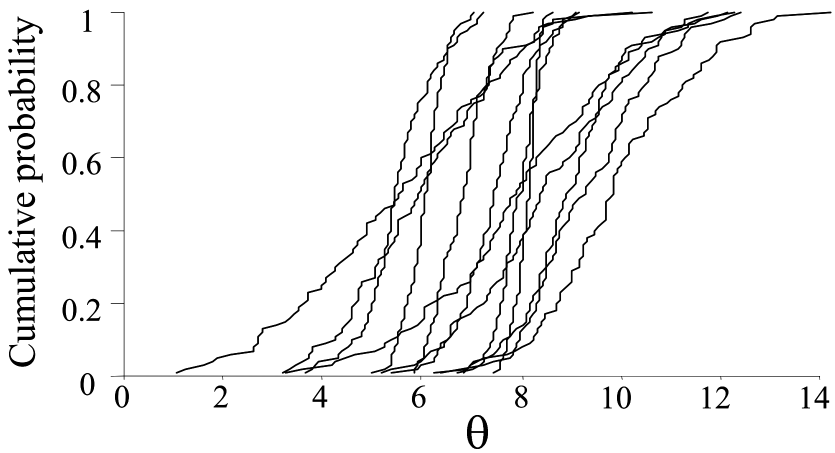


Figure 13. Hypothetical distribution of distributions (metadistribution).

The advantages of 2MC are that it acknowledges the uncertainty about an assessment more fully and incorporates this uncertainty into the mathematical calculation of the risks. It allows analysts to be forthright about what they know and don't know and lessens their need to make assumptions purely for the sake of mathematical convenience. The outputs from a 2MC simulation are useful for expressing confidence about exposure or risk predictions. They should be helpful in directing future data gathering by identifying important variables with high incertitude.

Disadvantages

This section reviews the disadvantages of 2MC. It covers the issues in increasing order of significance.

Computational cost. The first disadvantage of 2MC simulation is its computational complexity. Because the Monte Carlo simulations are nested, 2MC requires squared computational effort compared to the base simulation. This disadvantage is often the easiest to overcome. With computing power ever more accessible and inexpensive, 2MC simulations that used to require two weeks to compute on a desktop microcomputer now can be completed in several hours. Despite the improvements, the computational cost of 2MC can still be considerable in complex simulations.

Backcalculations difficult. Like ordinary Monte Carlo simulation, 2MC has no direct strategy for solving backcalculations. Backcalculation problems are those in which $C = f(A_1, A_2, \ldots, A_k, B)$ and we have empirical estimates for the A_i and want to find the B that will guarantee certain constraints on C. For example, a risk analyst might be asked to find the cleanup target for the concentration of some environmental contaminant such that the resulting exposures experienced by human or animal receptors are below tolerable limits. Answering such engineering questions is one of the most common reasons for undertaking risk assessments in the first place. Although backcalculations have been attempted in the context of 2MC simulations, they are at best rather cumbersome because they must be done via trial and error, with trials that are computationally expensive to start with.

Parameterizations cumbersome. Because 2MC simulation requires the analyst to specify a distribution for each uncertain parameter of the primary distributions and perhaps the ancillary inter-parameter dependencies as well, some analysts who were already heavily taxed in simple simulations can find it overwhelming to parameterize the full two-dimensional effort. Risk analysts sometimes feel the task of parameterizing a full two-dimensional simulation to be daunting at best, and at worst dizzying. In fact, the method depends on deploying two entirely distinct definitions of probability. In the inner simulations, the distributions represent variability and the normal aleatory interpretation of the distribution is used. But in the outer simulation, the distributions are intended to represent incertitude, and they must use an epistemic interpretation of probability. Some analysts find it difficult to express their incertitude numerically at all, much less as particular probability distributions. In principle, one could define even higher-dimensional* models to represent the uncertainty. Some analysts have attempted three- and even four-dimensional Monte Carlo simulations (Jaworska and Aldenberg 2000), but this hardly seems a workable solution when the complexity of the analysis is the primary problem. Some analysts object to assignment of any precise probability distribution to epistemically uncertain values. They argue that an interval, for instance, such as might be obtained as a measurement with finite accuracy is not equivalent to a uniform or any other distribution (Ferson and Ginzburg 1996; Walley 1991).

Ugly outputs. Because the result of 2MC simulation is, even for a single variable, a distribution of distributions, visualizing outputs can be difficult and the implications of results challenging to explain to managers and decision makers. The displays of metadistributions are sometimes called by the evocative names of “spaghetti plots” or “horsetail plots” or even “bad-hair-day plots”. These outputs can be confusing to interpret even for professionals, and often engender bewilderment among non-technical decision makers to whom they are shown. The first reaction of risk managers to such displays is often to laugh out loud—rarely an encouraging omen. Analysts often try to mask the complexity of the metadistributions by greatly simplifying the displays. For instance, risk analysts often replace a metadistribution with a three curves representing the “average” or “median” distribution and the 2.5th and 97.5th percentiles of the distributions in the metadistribution, considered pointwise. Even neglecting the enormous loss of information this entails, a three-curve display may often be misinterpreted by decision makers. They might believe, for instance, that 95% of the possible distributions represented by the metadistribution lie entirely within the outer curves. (This is the interpretation of Kolmogorov-Smirnov confidence limits for distributions.) This interpretation would be totally incorrect because the outer curves are pointwise percentiles, whatever *that* means. The average curve itself can be misleading as well. One might think that the average curve represents the most likely or, in some sense, the best-estimate distribution from the metadistribution. Such a thought would, however, be completely erroneous. Averaging distributions in this way can create Frankenstein distributions that are fundamentally unlike any of the distributions of the metadistribution. For instance, the pointwise average of normal distributions is not a normal distribution. Some analysts have suggested using gray scales or color coding to denote the complexity of a metadistribution, but this strategy could also be misunderstood in similar ways. Precious little research effort has been devoted to the question of how results from complex analyses should be communicated to the risk managers and decision makers who most need them. No studies have yet been published on the interpretive error rates associated with the visualizations from 2MC simulations, but it seems clear that new display techniques are needed.

Parameter dependence. There are also some serious technical difficulties that can beset 2MC simulations. For instance, if the distributions for the minimum and maximum of a uniform

*Good (1965) considered this idea. Kyburg (1989) pointed that if concern about the uncertainty in parameters can in principle drive the analysis to a higher level, one could fall into an insoluble infinite cascade.

distribution overlap at all, then there is a possibility that the minimum selected in the outer simulation exceeds the selected maximum. If that happens, the uniform distribution in the inner simulation is undefined. It is unclear what analysts should do in such a situation. The problem is actually rather general, because, for many statistical distributions, there are logical constraints that govern the relations between the parameters (see Frey and Rhodes 1998). For other distributions such as the beta or Weibull, the nature of the dependence between the parameters can be much subtler than it is for the uniform. The assumption of independence between parameters of distributions is not justifiable in general. Analysts who follow standard practice in the field and make independence assumptions may be creating impossible mathematical fictions akin to specifying a correlation matrix that is not positive semi-definite. Although it is easy to demonstrate with numerical examples that such fictions could in principle lead to serious miscalculations of risks, it is not clear whether this problem commonly occurs in real risk assessments, or what numerical consequences it may generate.

Cannot account for model uncertainty. Model uncertainty is the doubt about the correct form of the model used to compute the risk (Apostolakis 1995; Morgan and Henrion 1990). The 2MC strategy to express model uncertainty and incorporate it into calculations begins by defining a parameter, say *m*, whose possible values represent different possible models. Because model uncertainty is incertitude, the value of *m* is set in the outer loop of the 2MC simulation. In an inner simulation, the value of *m* is a constant, and the model it denotes is assumed to be the correct and incorrigible model, at least for the purposes within the single Monte Carlo simulation. If there are two possible models, then *m* would have a binary distribution, with the relative probabilities of the two models reflected by the probabilities of the binary states. If there are many distinct models, *m* would be some discrete distribution, with the probability associated with each possible value of *m* set to reflect the analyst's beliefs about how likely the each model is. The distribution could even be continuous if there are infinitely many possible distributions. If this probabilities of the various models are unknown, the traditional approach is to assume all possible models are equiprobable. This approach obviously translates the problem of model uncertainty into a problem of parametric uncertainty. This approach requires that the analyst know, and be able to explicitly enumerate or at least continuously parameterize, all the possible models. In practice of course, there might be many competing models, and analysts might not find it easy to identify each of them. It is easy to disprove the comfortable myth that some risk analysts seem to have that it is sufficient to "try a few models". If the resulting metadistribution is condensed into a single, precise probability distribution, then the approach represents model uncertainty as a stochastic mixture that averages together incompatible models (Finkel 1995), and therefore erases the uncertainty rather than truly capturing and propagating the uncertainty. For this reason, it is much preferable to keep the results as a metadistribution.

Confounding different kinds of probabilities. Cooke (2004) ridiculed fuzzy theorists for not having an operational definition of 'possibility'. Yet it is hard to seriously argue that probability theory as it has been practiced in risk analysis has the high ground with respect to having operational definitions. Many if not most applications of probability theory in risk analysis today confound different interpretations of probability. Two-dimensional simulations are arguably the worst offenders on this score as they commonly mix and match inputs from multiple, incomparable sources. When pressed for their definition of probability, analysts will generally offer something along frequentist lines involving an ensemble of like people or similar sites or analogous situations. But this interpretation is affixed at the end of an analysis that does not itself justify or give warrant to the interpretation. In practice, analyses are dubious admixtures of assumptions and distributions obtained from subjective judgments and opinions from multiple expert informants. In many cases, analysts cannot even specify the ensemble that the distribution proffered for a quantity is supposed to characterize. They literally do not know what they are modeling. Needless to say, they often do not distinguish, much less observe care in combining, distributions representing spatial variability, temporal fluctuation, inter-individual heterogeneity, or social diversity of opinions.

Inadequate model of incertitude. The most serious problem with 2MC simulation as a general tool for risk analysis is it does not handle incertitude correctly. It shares this problem with ordinary Monte Carlo simulation, Bayesian methods and probability theory in general. At least for the purposes of risk analysis, all of these approaches have an inadequate model of incertitude. There are two ramifications of this inadequacy. The first is that uniform distributions should not be used to represent the analyst's incertitude about a variable or parameter. The same holds true for any single uninformative or reference distribution (Irony and Singpurwalla 1997; Syversveen 1998). Although a two-dimensional Monte Carlo simulation does conscientiously separate variability from incertitude, it still applies the method of Bayes (and Laplace, Jaynes, etc.) to incertitude and treats ignorance the *same way* it treats variability. Picking a single distribution necessarily implies a great deal about a variable. As Bernardo acknowledged, any prior reflects some knowledge, and the uniform and its cousins are very informative of non-location parameters (Irony and Singpurwalla 1997). As a result, it can produce estimates that may be inappropriate or unusable in risk analysis (Ferson and Ginzburg 1996). This problem was reviewed by Ferson (1996). The second ramification of probability's inadequate model of incertitude concerns uncertainty about dependence. Ferson et al. (2004) showed that Monte Carlo methods cannot account for the case in which the dependence between parameters is unknown, even by varying the correlation coefficient among possible values in a 2MC simulation (cf. Cullen and Frey 1999). Essentially, the problem is that the simulation techniques available in Monte Carlo simulation cannot reproduce the infinite-dimensional nonlinearity that dependence functions might take.

Whenever 2MC simulations employ uniform distributions to model incertitude about parameters (in either the inner or outer loop), or model inter-variable relationships with precise dependence functions for the sake of mathematical convenience, the results of the simulation will likely fail to capture the extant uncertainty. Both of these mistakes will tend to artificially narrow the uncertainty, so that the analysis would tend to underestimate the possible probabilities of extreme events. Attempts to represent model uncertainty in a 2MC simulation are also likely to severely underestimate tail risks if the enumeration of possible models is depauperate. On the other hand, it is hard to anticipate whether counterfactually assuming the parameters of a distribution to be independent would tend to under- or overestimate the tail risks.

Conclusions

Ferson (1996) reviewed the limitations and disadvantages of (first-order) Monte Carlo simulation as a method for risk assessment and uncertainty analysis. This report considers the deficiencies of Bayesian methods and, secondarily, two-dimensional Monte Carlo simulation.

It is fair to say that the stories, so far, of Bayesian methods and two-dimensional Monte Carlo simulation in risk analysis are ones of promises yet to be fulfilled. If their respective assumptions hold, then both approaches work reasonably well and, clearly, they have their places in risk analysis and probabilistic uncertainty propagation. However, these methods are not robust in those situations when their assumptions don't hold. They presume too much information can be provided by the analyst than can commonly be provided in risk assessments where data and scientific understanding are constitutionally rare. The standard assumptions made in such cases are not appropriately protective because they can substantially underestimate tail risks. In the jargon of risk analysts, the methods are not "fail-safe" because they produce answers that can be misleading without warning. It might be entirely reasonable to be a Bayesian in making one's own inferences and personal decisions, but maintain that such methods ought not generally be employed, or at least not in their full strength, for group decisions, public policies or scientific purposes where there is a higher standard of accountability.

No data. The greatest of the promises tendered to risk analysts is that the methods can function and produce answers when there are no data at all. When there are no data, no likelihood can be constructed, so the unmodified initial prior serves as the zeroth posterior. It is guessing to be sure, but it is principled guessing. The central problem is that these answers will not tend to be empirically correct (Mayo 1996). One could not expect, for instance, that Bayesian 95% credibility intervals would enclose the parameters they represent 95% of the time. If they cannot be defended on empirical grounds, it is hard to discern the point in computing them at all. As Jim Dukelow has quipped, "it's hard to see the advantage in this; it has always been easy to get *wrong* answers."

Subjective judgments. Another seminal advantage touted for Bayesian methods and 2MC simulation is that they provide an architecture for handling subjective information that an analyst collates from personal experience or the advice of experts. This advantage is crucial because sometimes beliefs are all we have from which to construct inferences, synthesize conclusions and make decisions. The Bayesian approach acknowledges the inescapability of subjectivity in scientific endeavors and promises to provide a formal framework in which analysts can make use of relevant knowledge not expressed as empirical data. It promises to legitimize the use of subjective information in science. But many people maintain profound philosophical and scientific objections to such use of subjective information. It seems likely these objections are not going to fade away any time soon.

Generality. Bayesian methods promise to regularize the cacophony of inputs variously representing spatial or temporal variability, differences among components or individuals, analyst judgment or expert opinion into a coherent analysis based on a universal definition of probability. This promise would be very important if it could be sustained in a way that allowed analysts to generate results of relevance to risk analysis problems in the real world. The problem is that the Bayesian solution to the disparity among these inputs is to melt them all into a common denominator of subjective judgment.

Flexibility. Another often claimed advantage of Bayesian methods is their flexibility. They provide an adaptable general framework for risk analysis problems at multiple levels from the overarching to the purely technical. They permit arbitrary stopping rules in data collection. But flexibility, like any virtue, can be both a good and bad thing. The criticism of the Bayesian solution in the paternity example suggests a sense in which the method is actually too flexible. If different analysts can and regularly do get substantially different answers, observers might not consider the analysis scientific in any important sense or deserving of public confidence as a relevant characterization of the problem, as opposed to a diversion about the analysts themselves. At the same time, the claims of flexibility belie certain rigidities in the Bayesian approach. For instance, the zero preservation problem, which ruthlessly propagates an analyst's erroneous preconceptions, is one way in which Bayes' rule is not flexible enough. And, in keeping the prior model forever, some critics suggest that Bayesian updating is unscientifically rigid in principle.

Other purported advantages of Bayesian methods such as rationality, formality and connections with decision theory fall away in the context of group decision making or seem irrelevant in the rough-and-tumble of actual risk assessments in politically charged environments. One certainly does not need to be Bayesian to use decision theory, although some Bayesian like to pretend so. The permission that the Bayesian approach gives for data mining is welcome, but will rarely be important in risk analysis where data are so rare. Claims of the naturalness of Bayesian interpretations seem strained given the pedagogical experience, and they are diminished by the difficulty and computational burden of obtaining answers in practice.

Probability theory has been so successful in the marketplace of ideas because it has exploited polysemy in the word ‘probability’. It means different things to different people, and everyone hopes that probability theory will solve their particular problems. Contradicting this pluralism with a catholicity of their own, some Bayesians have claimed all of probability for themselves. For instance, Jaynes (2003) argued strenuously that probability theory is not the mathematical science of frequencies, but rather a calculus that tells us what our beliefs about some things should imply about our beliefs about other things. But risk analysis, it seems, should not be about *beliefs* in the first place. For the purposes of risk analysis, we do seem to require a mathematical science of frequencies, a calculus that allows us to compute with distributions, interpreted simply as collections of possible numbers weighted by how often they are realized in some actual or hypothetical ensemble. At the same time, we need this calculus to be able to make use of subjective information but in a way that respects incertitude as a distinct from variability and makes calculations and inferences that do not confound them. Although its promise is darkened by the specter of still greater computational difficulty, the robust Bayesian approach holds great promise to fix the most egregious problems of Bayesian methods and perhaps manifest the enriched mathematical science of frequencies needed in risk analysis.

For a long time, statistics was polarized by fierce debates between frequentists and Bayesians. In recent years, however, the profession has matured, and this controversy has calmed considerably. Discussions are more tolerant and inclusive, and both sides recognize the intrinsic advantages and disadvantages of the other. But outside of statistics, in the provinces of risk analysis and other mathematical and quantitative disciplines, the debates are only now heating up. At risk assessment review meetings, the arguments border on the vituperative, and discussion, if it can be called that, is characterized by fundamental disputes that seem beyond any possible reconciliation. We seem to still be a long way even from Sagoff’s (2004) peaceful coexistence, much less the elucidation of the truth by the application of our quantitative methods.

There is nevertheless hope for the future, and we should take comfort in the fact that the controversy in risk analysis means that it is young and growing field. If the debate has been intense, it is because the issues are important and the stakes are high. Some of the underlying points of divergence are true dilemmas, forks in the roads of thinking, that lead to really different conceptions. These will never be straightened out because they represent the fundamental complexity of human decision making in an incompletely understood world. But as other confusions are swept aside and misunderstandings are repaired by discourse and review, risk analysis will be richer for its growth. The different conceptions will lead to complementary perspectives that enlarge and deepen our understanding, readying us to answer the hard questions we face.

Acknowledgments

Many thanks are owed to Dominique Guyonnet for initiating and funding the work that led to this report. We also thank Tony Cox of Cox Associates, Dima Burmistrov of Menzie-Cura & Associates, Peter Walley, Teddy Seidenfeld of Carnegie Mellon University, Willem Roelofs of Central Sciences Laboratory in the United Kingdom, Cédric Baudrit of Institut de Recherche en Informatique de Toulouse, Scott Bartell of Emory University, Brendan Wintle of The University of Melbourne, Alyson Wilson of Los Alamos National Laboratory, Laura Wolf of Brigham Young University, Tom Aldenberg of Rijksinstituut voor Volksgezondheid & Milieu (RIVM) in the Netherlands, and W. Troy Tucker and David S. Myers of Applied Biomathematics for discussions on various topics Bayesian. These discussants are not responsible, of course, for any errors or omissions that may mar this report.

References

Abbas, A.E. (2003). Entropy methods for univariate distributions in decision analysis. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 22nd International Workshop*, edited by C.J. Williams, American Institute of Physics.

Agostinelli, C. (2000). References on Robust Statistics. University of Padua. http://homes.stat.unipd.it/claudio/red_rob.html.

Aven, T. (1997). Alternative Bayesian approaches to risk analysis. Stavanger University College and University of Oslo. http://www.nr.no/TilfeldigGang/sept97/taven/taven.html.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53: 370-418. Reprinted in 1958 in *Biometrika* 45: 293-315.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Bernard, J.-M. (2003). An introduction to the imprecise Dirichlet model for multinomial data. Tutorial for the Third International Symposium on Imprecise Probabilities and Their Applications (ISIPTA '03), Lugano, Switzerland. http://www.sipta.org/~isipta03/jean-marc.pdf.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, Series B* 41: 113-147 (with discussion).

Bernardo, J.M. and A.F.M. Smith (1994). *Bayesian Theory*. John Wiley & Sons, New York.

Berry, D.A. (1996). *Statistics: A Bayesian Perspective.* Duxbury Press, Belmont, California.

Berry, D.A. (1997). Bayesian Statistics. Institute of Statistics and Decisions Sciences, Duke University, http://www.pnl.gov/Bayesian/Berry/.

Berry, D.A. and D. Stangl (1996). *Bayesian Biostatistics*, Marcel Dekker Publishers.

Bertrand, J. (1889). *Calcul des Probabiltés.* Gauthier-Villars, Paris.

Boole, G. (1854). *An Investigation of the Laws of Thought, On Which Are Founded the Mathematical Theories of Logic and Probability*. Walton and Maberly, London. Reprinted in 1958 by Dover Publications, New York.

Casscells, W., A. Schoenberger and T.B. Graboys (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine* 299: 999-1001.

Clemen, R.T. and R.L. Winkler (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis* 19(2): 187-203.

Clyde, M. and E.I. George (2003). Model uncertainty. Technical report #2003-16, Statistical and Applied Mathematical Sciences Institute, www.samsi.info.

Cooke, R. (2004). The anatomy of the squizzel: the role of operational definitions in representing uncertainty. *Reliability Engineering and System Safety* 85: 313-319.

Cosmides, L. and J. Tooby (1996). Are humans good statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* 58: 1-73.

Cournot, A.A. (1843). *Exposition de la théorie des chances et des probabilités*. Hachette, Paris.

Cullen, A.C., and H.C. Frey (1999). *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Plenum Press: New York.

Dawid, A.P. (1982). Intersubjective statistical models, In *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino, North-Holland, Amsterdam.

de Finetti, B. (1970). *Teoria delle Probabilita*. Appeared in English translation in 1974 as *Theory of Probability*, volume 1. Wiley, London.

DeGroot, M.H. (1975). *Probability and Statistics*. Addison-Wesley, Reading, Massachusetts.

DeRobertis, L. and J.A. Hartigan (1981). Bayesian inference using intervals of measures. *The Annals of Statistics* 9: 235-244.
Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B* 57: 45–97.
Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press.
Ferson, S. (1996). What Monte Carlo methods cannot do. *Human and Ecological Risk Assessment* 2: 990–1007.
Ferson, S. (2002). *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. Lewis Publishers, Boca Raton, Florida.
Ferson, S. and L.R. Ginzburg (1996). Different methods are needed to propagate ignorance and variability. *Reliability Engineering and Systems Safety* 54: 133–144.
Ferson, S., V. Kreinovich, L. Ginzburg, K. Sentz and D.S. Myers. (2003). *Constructing Probability Boxes and Dempster-Shafer Structures*. Sandia National Laboratories, SAND2002-4015, Albuquerque, New Mexico.
Ferson, S., R.B. Nelsen, J. Hajagos, D.J. Berleant, J. Zhang, W.T. Tucker, L. Ginzburg and W.L. Oberkampf. (2004). *Dependence in Probabilistic Modeling, Dempster-Shafer Theory, and Probability Bounds Analysis*. Sandia National Laboratories, SAND2004-3072, Albuquerque.
Finkel, A.M. (1995). A second opinion on an environmental misdiagnosis: the risky prescriptions of *Breaking the Vicious Circle*. *New York University Environmental Law Journal* 3: 295–381.
Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.
Fisher, R.A. (1973). *Statistical Methods and Scientific Inference,* 3rd ed. Hafner Press, New York.
Fisher, R.A. (1957). The underworld of probability. *Sankhya* 18 201-210.
Frank, M.J., R.B. Nelsen, and B. Schweizer (1987). Best-possible bounds for the distribution of a sum—a problem of Kolmogorov. *Probability Theory and Related Fields* 74, 199-211.
Fréchet, M. (1935). Généralisations du théorème des probabilités totales. *Fundamenta Mathematica* 25: 379-387.
Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, Sect. A* 9: 53-77.
Gelman, A., J.B. Carlin, H.S. Stern, D.B. Rubin (1995). *Bayesian Data Analysis*. CRC Press, Boca Raton.
Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond heuristics and biases. *European Review of Social Psychology* 2: 83-115.
Gigerenzer, G. (2002). *Calculated Risks: How to Know When Numbers Deceive You.* Simon and Schuster, New York.
Glymour, C. (1980). *Theory and Evidence.* [see the essay "Why I am not a Bayesian" on pages 63–93]. Princeton University Press, Princeton. Reprinted in *Philosophy of Science: The Central Issues*, edited by M. Curd and J.A. Cover, W.W. Norton and Company. Also available at http://www.philosophy.ubc.ca/faculty/savitt/glymour.htm
Good, I.J. (1965). *The Estimation of Probabilities*. MIT Press, Cambridge, Massachusetts.
Grandy, W.T., Jr. and L.H. Schick (1991). *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, Dordrecht.
Hájek, A. (2003). Conditional probability is the very guide of life. Pages 183-203 in *Probability Is the Very Guide of Life*, edited by H.E. Kyburg, Jr., and M. Thalos. Open Court, Chicago.
Hammitt, J.K. (1995). Can more information increase uncertainty? *Chance* 15–17,36.
Helton, J.C. (1994). Treatment of uncertainty in performance assessments for complex systems. *Risk Analysis* 14: 483-511.
Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science* 14: 382-401. Correction: *Statistical Science* 15: 193-195. http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Hoffman, F. O. and J.S. Hammonds (1994). Propagation of uncertainty in risk assessments: The need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Analysis* 14(5):707-712.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Insua, D.R. and F. Ruggeri (eds.) (2000). *Robust Bayesian Analysis*. Lecture Notes in Statistics, Volume 152. Springer-Verlag, New York.

Irony, T.Z. and N.D. Singpurwalla (1997). Noninformative priors do not exist: a discussion with Jose M. Bernardo. *Journal of Statistical Planning and Inference* 65: 159-189. Available at http://www.unine.ch/statistics/postgrad/images/Dialog.pdf.

Jaffe, A.H. (2003). A polemic in probability. *Science* 301: 1329-1330.

Jaffray, J.-Y. (1992). Bayesian updating and belief functions. *IEEE Transactions on Systems, Man, and Cybernetics* 22: 1144–1152.

Jaffray, J.-Y. (1994). Dynamic decision making with belief functions. *Advances in the Dempster-Shafer Theory of Evidence*, R.R. Yager, J. Kacprzyk and M. Fedrizzi (eds.). John Wiley and Sons, Inc.

Jaworska, J.S. and T. Aldenberg (2000). Estimation of HC5 taking into account uncertainties of individual dose response curves and species sensitivity distribution. Presentation at the 2000 Society for Risk Analysis annual meeting, Arlingotn, Virginia, http://www.riskworld.com/Abstract/2000/SRAam00/ab0ac164.htm.

Jaynes, E.T. (1957). Information theory and statistical mechanics. *Physical Review* 106: 620–630.

Jaynes, E.T. (2003). *Probability Theory*, edited by L. Bretthorst. Cambridge University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A* 186: 453-461.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford.

Kline, M. (1980). *Mathematics: The Loss of Certainty*. Oxford University Press, Oxford.

Kyburg, H. (1989). Higher order probabilities. Pages pages 15-22 in *Uncertainty in Artificial Intelligence*, volume 3, edited by L.N. Kanal, T.S. Levitt and J.F. Lemmer, Elsevier Science Publishers, Amsterdam.

Kyburg, H.E., Jr. (1999). "hume [sic] and induction", email to the Association for Uncertainty in Artificial Intelligence listserver, 8 July 1999, from kyburg@redsuspenders.com.

Laplace, Marquis de, P.S. (1820). *Théorie analytique de probabilités* (edition troisième). Courcier, Paris. The introduction (*Essai philosophique sur les probabilités*) is available in an English translation in *A Philosophical Essay on Probabilities* (1951), Dover Publications, New York.

Lavine, M. (1991). Sensitivity in Bayesian statistics, the prior and the likelihood. *Journal of the American Statistical Association* 86 (414): 396-399.

Lee, P.M. (1997). *Bayesian Statistics: An Introduction*. Edward Arnold (Hodder Arnold) London.

Lee, R.C. and W.E. Wright (1994). Development of human exposure-factor distributions using maximum-entropy inference. *Journal of Exposure Analysis and Environmental Epidemiology* 4: 329-341.

Levine, R.D. and M. Tribus (1976). *The Maximum Entropy Formalism*. MIT Press, Cambridge.

Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Volume 2: Statistics*. Cambridge University Press, Cambridge.

Malakoff, D. (1999). Bayes offers a 'new' way to make sense of numbers. *Science* 286: 1460-4.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. The University of Chicago Press, Chicago.

Mongin, P. (1995). Consistent Bayesian aggregation. *Journal of EconomicTheory* 6: 313-351.

Neapolitan, R.E. (1992). A survey of uncertain and approximate inference. Pages 55-82 in *Fuzzy Logic for the Management of Uncerainty,* edited by L.A.Zadeh and J. Kacprzyk. John Wiley & Sons, New York.

O'Hagan, T. (2002). "Some comments on the challenge problems", and "A rejoinder". http://www.sandia.gov/epistemic [follow the Submitted Comments link].
Owen, A.B. (2001). *Empirical Likelihood*. Chapman & Hall / CRC Press, Boca Raton, Florida.
Peirce, C.S. (1878). The probability of induction. *Popular Science Monthly* 12: 705-718.
Peirce, C.S. (1930-1935). *Collected Papers of Charles Sanders Peirce*. Volumes 1-6, edited by C. Hartshorne and P. Weiss. Harvard University Press, Cambridge.
Peirce, C.S. (1958). *Collected Papers of Charles Sanders Peirce*. Volumes 7-8, edited by A. Burks. Harvard University Press, Cambridge.
Pericchi, L.R. (2000). Sets of prior probabilities and Bayesian robustness. Imprecise Probability Project. http://ippserv.rug.ac.be/documentation/robust/robust.html. http://www.sipta.org/documentation/robust/pericchi.pdf.
Pericchi, L.R. and P. Walley (1991). Robust Bayesian credible intervals and prior ignorance. *International Statistical Review* 58: 1-23.
Raftery, A.E., D. Madigan, and J.A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92: 179–191.
Rice, A. (2003). "Everybody makes errors": The intersection of De Morgan's logic and probability, 1837-1847. *History and Philosophy of Logic* 24: 289-305.
Sagoff, M. (2004). *Price, Principle, and the Environment*. Cambridge University Press.
Sander, P. and R. Badoux (eds.) (1991). *Bayesian Methods in Reliability*. Kluwer, Dordrecht.
Savage, L.J. (1954). *Foundations of Statistics*. John Wiley & Sons, New York.
Seidenfeld et al. (1989). On the shared preferences of two Bayesian decision makers. *Journal of Philosophy* 86: 225-244.
Seidenfeld, T. (1990). Two perspectives on consensus for Bayesian inference and decisions. In *Knowledge Representation and Defensible Reasoning*, edited by Kyburg et al. Kluwer, Dordrecht.
Simon, J.L. (2003). *The Philosophy and Practice of Resampling Statistics*. http://www.resample.com/content/teaching/philosophy.
Solana, V. and N.C. Lind (1990). Two principles for data based on probabilistic system analysis,'' Proceedings of ICOSSAR '89, 5th International Conferences on Structural Safety and Reliability. American Society of Civil Engineers, New York.
Thompson, K. (2000). "[riskanal] Re: Many thanks: re. respectable Bayesias [sic]", email to the Riskanal listserver, 20 January 2000, from kimt@hsph.harvard.edu.
Uffink, J. (1996). The constraint rule of the maximum entropy principle. *Studies in History and Philosophy of Modern Physics* 27: 47-79.
Venn, J. (1866). The Logic of Chance. Macmillan, Cambridge.
Vose, D. (2000). *Risk Analysis: A Quantitative Guide*. John Wiley & Sons, Chichester.
Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.
Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society* 58: 3-37.
Yang, R. and J.O. Berger (1998). A catalog of noninformative priors. Parexel International and Duke University, http://www.stat.missouri.edu/~bayes/catalog.ps