Reprinted from: P.F. Fougère (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 81–142. ©1990 Kluwer Academic Publishers.

FROM LAPLACE TO SUPERNOVA SN 1987A: BAYESIAN INFERENCE IN ASTROPHYSICS

T. J. LOREDO
Dept. of Astronomy and Astrophysics
University of Chicago
5640 South Ellis Ave.
Chicago, IL 60637

ABSTRACT. The Bayesian approach to probability theory is presented as an alternative to the currently used long-run relative frequency approach, which does not offer clear, compelling criteria for the design of statistical methods. Bayesian probability theory offers unique and demonstrably optimal solutions to well-posed statistical problems, and is historically the original approach to statistics. The reasons for earlier rejection of Bayesian methods are discussed, and it is noted that the work of Cox, Jaynes, and others answers earlier objections, giving Bayesian inference a firm logical and mathematical foundation as the correct mathematical language for quantifying uncertainty. The Bayesian approaches to parameter estimation and model comparison are outlined and illustrated by application to a simple problem based on the gaussian distribution. As further illustrations of the Bayesian paradigm, Bayesian solutions to two interesting astrophysical problems are outlined: the measurement of weak signals in a strong background, and the analysis of the neutrinos detected from supernova SN 1987A. A brief bibliography of astrophysically interesting applications of Bayesian inference is provided.

Contents

1. Introduction

2. What is Probability?
2.1 TWO DEFINITIONS OF PROBABILITY
2.2 SOME EARLY HISTORY: BERNOULLI, BAYES, AND LAPLACE
2.2.1 *Frequency from Probability*
2.2.2 *Probability from Frequency: Bayes' Theorem*
2.3 FREQUENTIST PROBABILITY
2.4 CRITICISM OF THE FREQUENTIST APPROACH
2.4.1 *Arbitrariness and Subjectivity*
2.4.2 *Comparison with Intuition*
2.4.3 *Randomness vs. Uncertainty*
2.4.4 *The Frequentist Failure*
2.5 SOME RECENT HISTORY

3. Bayesian Probability Theory: A Mathematical Language for Inference
3.1 THE DESIDERATA
3.2 THE GRAMMAR OF INFERENCE: THE PROBABILITY AXIOMS
3.2.1 *The Product Rule*
3.2.2 *The Sum Rule*

3.3 THE VOCABULARY OF INFERENCE: ASSIGNING PROBABILITIES
3.3.1 *Least Informative Probabilities*
3.3.2 *Informative Probabilities, Bayes' Theorem, and Maximum Entropy*
3.4 THE FREQUENCY CONNECTION

4. Some Well-Posed Problems
4.1 BAYESIAN PARAMETER ESTIMATION
4.1.1 *Parametrized Models*
4.1.2 *Summarizing Inferences*
4.2 BAYESIAN MODEL COMPARISON
4.3 PROBLEMS WITH FREQUENTIST MODEL ASSESSMENT
4.3.1 *Reliance on Many Hypothetical Data Sets*
4.3.2 *Reliance on a Single Hypothesis*
4.3.3 *Implicit Alternatives: The Myth of Alternative-Free Tests*
4.3.4 *Violation of Consistency and Rationality*

5. Bayesian and Frequentist Gaussian Inference
5.1 THE STATUS OF THE GAUSSIAN DISTRIBUTION
5.2 ESTIMATING THE SIGNAL
5.2.1 *The Frequentist Approach*
5.2.2 *The Bayesian Approach*
5.2.3 *Comparison of Approaches*
5.2.4 *Improper Priors*
5.2.5 *The Robustness of Estimation*
5.2.6 *Reference Priors*
5.3 MODEL COMPARISON

6. Case Study: Measuring a Weak Counting Signal

7. Case Study: Neutrinos from SN 1987A
7.1 A BEWILDERING VARIETY OF FREQUENTIST ANALYSES
7.2 THE BAYESIAN ANALYSIS

8. Where to Go From Here
8.1 ASTROPHYSICAL ESTIMATION PROBLEMS
8.2 BAYESIAN SPECTRUM ANALYSIS
8.3 INVERSE PROBLEMS
8.4 JAYNESIAN PROBABILITY THEORY

9. Acknowledgements

10. References

1. Introduction

Few astrophysicists have expertise in the use of advanced statistical methods. The reason for this is not difficult to find, for examination of the use of statistics in the astrophysical literature reveals the lack of a clear rationale for the choice and use of advanced methods.

Unfortunately, this problem is not intrinsic to astrophysics, but has been inherited from statistics itself. To an outsider, statistics can have the appearance of being merely an "industry" where statistical methods are invented without a clear design rationale, and then evaluated by mass-producing simulated data sets and analyzing the average, long-run behavior of the methods. As a result, there often are several methods available for addressing a particular statistical question, each giving a somewhat different answer from the others, with no compelling criteria for choosing among them. Further, the reliance on long-run behavior for the evaluation of statistical methods makes the connection between textbook statistical inferences and the real life problems of scientists seem rather tenuous. This problem can be particularly acute in astrophysics, where the notion of a statistical ensemble is often extremely contrived and can hence seem irrelevant. The gamma-ray astronomer does not want to know how an observation of a gamma-ray burst would compare with thousands of other observations of that burst; the burst is a unique event which can be observed only once, and the astronomer wants to know what confidence should be placed in conclusions drawn from the one data set that actually exists. Similarly, the cosmologist is not comforted to learn that his statement about the large scale structure of the Universe would be correct 95% of the time were he to make similar observations in each of thousands of universes "like" our own. He wants to know how much confidence should be placed in his statement about our particular Universe, the only one we know exists.

Given these difficulties, it is no wonder that many scientists are dubious about results obtained using any but the simplest statistical methods, and no wonder that some openly assert, "If it takes statistics to show it, I don't believe it." It is no wonder, but it is unfortunate. Among all scientists, it is perhaps most unfortunate for the astronomer, who studies objects inaccessible to direct manipulation in a laboratory, and whose inferences are thus fraught with uncertainty, uncertainty crying out for quantification.

It is the thesis of this paper that this situation is unnecessary, that there exists a simple mathematical language for the quantification of uncertainty, that this language produces *unique* answers to well-posed problems, and that its answers are demonstrably optimal by rather simple, compelling desiderata. This language is *Bayesian Probability Theory* (BPT), and far from being a new approach to statistics, it is the *original* approach to statistics, predating the current long-run performance approach by a century. Ironically, it was originally developed by an astrophysicist: Laplace used such methods to analyze astronomical observations for comparison with his famous calculations in celestial mechanics, and developed them at length in his *Théorie Analytique des Probabilités* (Laplace 1812). Heightening the irony, many later developments of Laplace's theory also came from mathematicians and physicists analyzing astronomical problems (see Feigelson 1989 for a brief review). More recently, a full development of Laplace's theory, including the solutions to dozens of practical statistical problems, was published by Sir Harold Jeffreys while a professor of astronomy at Cambridge University in the chair previously held by Eddington (Jeffreys 1939).*

The Bayesian approach to probable inference is remarkably straightforward and intuitive. In fact, it is most likely what the reader already believes probability theory is, since the

* This work remains little known among astronomers. A recent obituary of Jeffreys (Runcorn 1989) fails even to mention this important work, described by the prominent statistician I. J. Good as being "of greater importance for the philosophy of science, and obviously of greater immediate practical importance, than nearly all the books on probability written by professional philosophers *lumped together*" (Good 1980).

intuitive understanding physicists have of the more common statistical notions (such as 1σ error bars) is often identical to the Bayesian interpretation of the notion, and far from the rigorous "classical" or "orthodox" interpretation. But the precise quantification of such intuitive notions in Bayesian inference allows one to extend them into the realm where subtleties often leave our intuition—and classical statistics—at a loss. In such cases, the Bayesian solution often appears beautifully intuitive *a posteriori*, our intuition having been trained and sharpened by probability theory.

The plan of this paper is as follows. First, we will have to discuss exactly what one means by the word "probability." This may sound like a topic for philosophers, but the whole course of probability theory is set by what one decides the conceptual playground of the theory is, so the discussion is crucial. Next we will see that the Bayesian notion of probability, which appears at first to be too vague for quantitative analysis, in fact allows one to develop a complete mathematical language for dealing with uncertainty that is both simpler than standard statistics and more general than it, including much of it as a special case. Following this, we will learn how to use the theory to address two classes of problems of particular interest to scientists: the estimation of parameters in a model, and the assessment of competing models. The basic ideas will be illustrated by comparing the Bayesian approach to measuring a signal in Gaussian noise with the standard long term performance approach.

Once the general theory is set up, we will outline its application to two real astrophysical problems: the measurement of a weak photon counting signal in a (possibly strong) background, and the analysis of the neutrinos detected from the supernova SN 1987A. The failure of orthodox methods to guide astronomers to a single, optimal solution to a problem as simple and fundamental as the measurement of a weak signal is a powerful indication of the poverty of such methods. The Bayesian solution to this problem is so simple that it is reduced from a research problem (Hearn 1969; O'Mongain 1973; Cherry *et al.* 1980) to an undergraduate homework problem.

This is a lot of ground to cover in the pages of a single paper, and much of it will be covered unevenly and incompletely. Hopefully, the reader will be induced to study the cited references where the theory is developed both more eloquently and more fully. To this end, the concluding section not only summarizes the contents of this work, but also points the reader to Bayesian literature covering several topics of particular interest to astrophysicists, including Bayesian spectrum analysis and the Bayesian approach to inverse problems.

2. What is Probability?

2.1 TWO DEFINITIONS OF PROBABILITY

Traditionally, probability is identified with the *long-run relative frequency of occurrence of an event*, either in a sequence of repeated experiments or in an ensemble of "identically prepared" systems. We will refer to this view of probability as the "frequentist" view; it is also called the "classical," "orthodox," or "sampling theory" view. It is the basis for the statistical procedures currently in use in the physical sciences.

Bayesian probability theory is founded on a much more general definition of probability. In BPT, probability is regarded as a real-number-valued measure of the plausibility of a proposition when incomplete knowledge does not allow us to establish its truth or falsehood with certainty. The measure is taken on a scale where 1 represents certainty of the truth of the proposition, and 0 represents certainty of its falsehood. This definition has an obvious connection with the colloquial use of the word "probability." In fact, Laplace viewed probability theory as simply "common sense reduced to calculation" (Laplace 1812, 1951). For Bayesians, then, probability theory is a kind of "quantitative epistemology", a numerical encoding of one's state of knowledge.

Few works on statistics for the physical sciences bother to note that there is controversy over so fundamental a notion as the definition of probability. In fact, two of the most influential works introducing statistical methods to physical scientists neither define probability nor discuss the complicated frequentist derivation and interpretation of concepts as simple and as widely used as the 1σ confidence region (Bevington 1969; Press *et al.* 1986). Other texts, noting that there is some controversy over the definition, adopt the frequency definition, asserting that there is little practical difference between the approaches (Eadie *et al.* 1971; Martin 1971; Mendenhall *et al.* 1981).

Of course, it is futile to argue over which is the "correct" definition of probability. The different definitions merely reflect different choices for the types of problems the theory can address, and it seems possible that either definition could lead to a consistent mathematical theory. But though this is true, it leaves open the question of which approach is more useful or appropriate, or which approach addresses the types of problems actually encountered by scientists in the most straightforward manner.

In fact, it will not take much deep thought for us to see that the Bayesian approach to probability theory is both more general than the frequentist approach, and much more closely related to how we intuitively reason in the presence of uncertainty. We will also find that Bayesian solutions of many important statistical problems are significantly simpler to derive than their frequentist counterparts. But if this is true, and if, as noted earlier, the Bayesian approach is the historically older approach, why was the frequentist definition adopted, and why has it dominated statistics throughout this century? To address these questions, our discussion of the contrast between Bayesian and frequentist reasoning will be quasi-historical. More extensive discussions of the history of probability theory and the Bayesian/frequentist controversy are available in Rényi (1972, Appendices III and IV), Jaynes (1978, 1986a), and Grandy (1987, Ch. 2).

2.2 SOME EARLY HISTORY: BERNOULLI, BAYES, AND LAPLACE

2.2.1. *Frequency from Probability.* Though statistical problems, particularly those related to gambling and games of chance, have entertained the minds of thinkers since ancient times, the first formal account of the calculation of probabilities is Bernoulli's *Ars Conjectandi* ("The Art of Conjecture", Bernoulli 1713). Bernoulli was what we would today term a Bayesian, holding that probability is "the degree of certainty, which is to the certainty as the part to the whole." He clearly recognized the distinction between probability and frequency, deriving the relationship between probability of occurrence in a single trial and frequency of occurrence in a large number of independent trials now known as Bernoulli's theorem, or the law of large numbers.

Bernoulli's theorem tells us that, if the probability of obtaining a particular outcome in a single trial is known to be p, the relative frequency of occurrence of that outcome in a large number of trials converges to p.

Also of interest to Bernoulli was the inverted version of this problem: supposing the probability of occurrence in a single trial is unknown, what does the observation of the outcome n times in N repeated, independent trials tell us about the value of the probability? Bernoulli never solved this problem, but his interest in it further emphasizes the distinction made by him and his contemporaries between probability ("degree of certainty") and frequency.

2.2.2. *Probability from Frequency: Bayes' Theorem.* A solution to Bernoulli's problem was published posthumously by the Rev. Thomas Bayes (1763). It was soon rediscovered by Laplace, in a much more general form, and this general form is known as *Bayes' Theorem* (BT). It can be derived very simply as follows.

The mathematical content of the probability theory of Bernoulli, Bayes, and Laplace was specified by taking as *axioms* the familiar sum rule,

$$p(A \mid C) + p(\overline{A} \mid C) = 1, \tag{1}$$

and product rule,

$$p(AB \mid C) = p(A \mid BC)p(B \mid C). \tag{2}$$

Here the symbols, A, B, C, represent propositions, $\overline{A}$ represents the denial of A (read "not A"), and AB means "A *and* B," a proposition that is true only if A and B are both true. The vertical bar is the conditional symbol, indicating what information is assumed for the assignment of a probability. We must always assume something about the phenomenon in question, and it is good practice to put these assumptions out in the open, to the right of the bar. Failure to do this can lead to apparent paradoxes when two problems with different background assumptions are compared; see Jaynes (1980a) for an educational example.

All legitimate relationships between probabilities can be derived from equations (1) and (2). For example, we may want to know the probability that either or both of two propositions is true. Denoting this by $p(A + B \mid C)$, it can be easily shown (Jaynes 1958, 1990b; Grandy 1987) that the axioms imply

$$p(A + B \mid C) = p(A \mid C) + p(B \mid C) - p(AB \mid C). \tag{3}$$

In fact, we can take this in place of (1) as one of our axioms if we wish. If A and B are exclusive propositions, so that only one of them may be true, $p(AB \mid C) = 0$, and equation (3) becomes the familiar sum rule for exclusive propositions: $p(A+B \mid C) = p(A \mid C) + p(B \mid C)$.

It is important to keep in mind that the arguments for a probability symbol are propositions, not numbers, and that the operations inside the parentheses are logical operations. The symbols for logical operations are here chosen to make the axioms mnemonic. Thus logical "and," represented by juxtaposition in the argument list, leads to multiplication of probabilities. Similarly, logical "or," indicated by a "+" in the argument list, leads to sums of probabilities. But the meanings of juxtaposition and the "+" sign differ inside and outside of the probability symbols.

The propositions AB and BA are obviously identical: the ordering of the logical "and" operation is irrelevant. Thus equation (2) implies that $p(A \mid BC)p(B \mid C) = p(B \mid AC)p(A \mid C)$. Solving for $p(A \mid BC)$, we find

$$p(A \mid BC) = p(A \mid C)\frac{p(B \mid AC)}{p(B \mid C)}. \tag{4}$$

This is Bayes' theorem; it is a trivial consequence of axiom (2).

Bayesian probability theory is so-called because of its wide use of BT to assess hypotheses, though of course Bayesians use all of probability theory, not just BT. To see how BT can help us assess an hypothesis, make the following choices for the propositions A, B, and C. Let $A = H$, an hypothesis we want to assess. Let $B = D$, some data we have that is relevant to the hypothesis. Let $C = I$, some background information we have indicating the way in which H and D are related, and also specifying any alternatives we may have to H.* With these propositions, BT reads

$$p(H \mid DI) = p(H \mid I)\frac{p(D \mid HI)}{p(D \mid I)}. \tag{5}$$

Thinking about this a little, we see that BT represents *learning*. Specifically, it tells us how to adjust our plausibility assessments when our state of knowledge regarding an hypothesis changes through the acquisition of data. It tells us that our "after data" or *posterior probability* of H is obtained by multiplying our "before data" or *prior probability* $p(H \mid I)$ by the probability of the data assuming the truth of the hypothesis, $p(D \mid HI)$, and dividing it by the probability that we would have seen the data anyway, $p(D \mid I)$. The factor $p(D \mid HI)$ is called the *sampling distribution* when considered as a function of the data, or the *likelihood function*, $\mathcal{L}(H)$, when considered as a function of the hypothesis. For reasons that will become clear below, $p(D \mid I)$ is sometimes called the *global likelihood*, and usually plays the role of an ignorable normalization constant.

Two points are worth emphasizing immediately about BT. First, there is nothing about the passage of time built into probability theory. Thus, our use of the terms "after data," "before data," "prior probability," and "posterior probability" do not refer to times before or after data is available. They refer to *logical* connections, not temporal ones. Thus, to be precise, a prior probability is the probability assigned before *consideration* of the data, and similarly for the other terms.

Second, for those who may have been exposed to BT before and heard some ill-informed criticisms of it, the I that is always to the right of the bar in equation (5) is not some major premise about nature that must be true to make our calculation valid. Nor is it some strange, vague proposition defining some universal state of ignorance. It simply is the background information that defines the problem we wish to address at the moment. It may specify information about H that we are content to assume true, or it may simply specify some alternative hypotheses we wish to compare with H. We will have the chance to elaborate on this point below, when we see how to use (5) to solve concrete problems.

To solve Bernoulli's problem, Bayes used a special case of BT to evaluate different propositions about the the value of the single trial probability of an outcome, given its relative frequency of occurrence in some finite number of trials (Bayes 1763; Jaynes 1978). Later, independently, Laplace greatly developed probability theory, with BT playing a key role. He used it to address many concrete problems in astrophysics. For example, he used BT to estimate the masses of the planets from astronomical data, and to quantify the uncertainty of the masses due to observational errors. Such calculations helped him choose which problems in celestial mechanics to study by allowing him to identify significant perturbations and to make predictions that would be testable by observers.

---

* To be precise, H is a proposition asserting the truth of the hypothesis in question ("The plasma temperature is T."), D is a proposition asserting the values of the data ("The observed photon energy is ϵ."), etc., but we will usually be a bit free with our language in this regard.

2.3 FREQUENTIST PROBABILITY

Despite the success of Laplace's development of probability theory, his approach was soon rejected by mathematicians seeking to further develop the theory. This rejection was due to a lack of a compelling rationale for some of the practices of Bernoulli, Bayes, Laplace, and their contemporaries.

First, the idea that probability should represent a degree of plausibility seemed too vague to be the foundation for a mathematical theory. The mathematical aspect of the theory followed from the axioms (1) and (2), but it was certainly not obvious that calculations with degrees of plausibility had to be governed by those axioms and no others. The axioms seemed arbitrary.

Second, there were problems associated with how prior probabilities should be assigned. The probability axioms described how to manipulate probabilities, but they did not specify how to assign the probabilities that were being manipulated. In most problems, it seemed clear how to assign the sampling probability, given some model for the phenomenon being studied. But finding compelling assignments of prior probabilities proved more difficult. In a certain class of problems, Bernoulli and his successors found an intuitively reasonable principle for such an assignment that we will call the *Principle of Indifference* (PI; it is also known as the Principle of Insufficient Reason). It is a rule for assignment of probabilities to a finite, discrete set of propositions that are mutually exclusive and exhaustive (*i.e.*, one proposition, and only one, must be true). The PI asserts that if the available evidence does not provide any reason for considering proposition A_1 to be more or less likely than proposition A_2, then this state of knowledge should be described by assigning the propositions equal probabilities. It follows that in a problem with N mutually exclusive and exhaustive propositions, and no evidence distinguishing them, each proposition should be assigned probability $1/N$.

While the PI seemed compelling for dealing with probability assignments on discrete finite sets of propositions, it was not clear how to extend it to cases where there were infinitely many propositions of interest. Such cases arise frequently in science, whenever one wants to estimate the value of a continuous parameter, θ. In this case, θ is a label for a continuous infinity of propositions about the true value of the parameter, and we need to assign a prior probability (density) to all values of θ in order to use BT. We might specify indifference about the value of θ by assigning a flat probability density, with each value of θ having the same prior probability as any other. Unfortunately, it seems that we could make the same statement about prior probabilities for the value of $\theta' \equiv \theta^2$. But a flat density for θ' does not correspond to a flat density for θ. For this reason, inferences about continuous parameters seem to have a disturbing subjectivity, since different investigators choosing to label hypotheses differently by using different parameters could come to different conclusions.

The mathematicians of the late nineteenth and early twentieth centuries dealt with these legitimate problems by surgical removal. To eliminate the arbitrariness of the probability axioms, they drastically restricted the domain of the theory by asserting that probability had to be interpreted as relative frequency of occurrence in an ensemble or in repeated random experiments. The algebra of relative frequencies obviously satisfied the axioms, so their arbitrariness was removed.

As a byproduct, the second problem with the Laplace theory disappeared, because the frequency definition of probability made the concept of the probability of an hypothesis illegitimate. This is because the frequency definition can only describe the probability of a *random variable:* a quantity that can meaningfully be considered to take on various values throughout an ensemble or a series of repeated experiments. An hypothesis, being either true or false for every element of an ensemble or every repetition of an experiment, is not a random variable; its "relative frequency of occurrence" throughout the ensemble or sequence of experiments is either 0 or 1. For example, were we to attempt to measure the radius of a planet by repeated observation, the observed radius would vary from repetition to repetition, but the actual radius of the planet would be constant, and hence not amenable to frequentist description. Put another way, were we to analyze the observations with BT, we would be attempting to find a posterior distribution for the radius; but if this posterior distribution is a frequency distribution, there is an obvious problem: how can the frequency distribution of a parameter become known from data that were taken with only one value of the parameter actually present?

For these reasons, the concept of the probability of an hypothesis is held as meaningless in frequentist theory. A consequence is that scientists are denied the ability to use BT to assess hypotheses, so the problem of assigning prior probabilities disappears. The resulting theory was originally deemed superior to BPT, especially because it seemed more objective. The apparent subjectivity of prior probability assignments was avoided, and the frequency definition of probability, by its reference to observation of repeated experiments, seemed to make probability an objective property of "random" phenomena, and not a subjective description of the state of knowledge of a statistician.

2.4 CRITICISM OF THE FREQUENTIST APPROACH

2.4.1. *Arbitrariness and Subjectivity.* Unfortunately, assessing hypotheses was one of the principle aims of probability theory. Denied the use of BT for this task, frequentist theory had to develop ways to accomplish it without actually calculating probabilities of hypotheses. The frequentist solution to this problem was the creation of the discipline of *statistics.* Basically, one constructs some function of observable random variables that is somehow related to what one wishes to measure; such a function is called a *statistic.* Familiar statistics include the sample mean and variance, the χ^2 statistic, and the F statistic. Since a statistic is a function of random variables, its probability distribution, assuming the truth of the hypothesis of interest, can be calculated. A hypothesis is assessed by comparing the observed value of the statistic with the long-run frequency distribution of the values of the statistic in hypothetical repetitions of the experiment.

Intuition was a clear guide for the construction of statistics for simple problems (the familiar statistics mentioned above refer to the rather simple gaussian distribution). But for complicated problems, there is seldom a compelling "natural" choice for a statistic. Several statistical procedures may be available to address a particular problem, each giving a different answer. For example, to estimate the value of a parameter, one can use the method of moments, the maximum likelihood method, or a more specialized *ad hoc* method. Or, to compare unbinned data with an hypothesized continuous distribution, one could use one of the three Kolmogorov-Smirnov tests, the Smirnov-Cramer-von Mises test, or any of a number of obvious generalizations of them.

To provide a rationale for statistic selection, many principles and criteria have been added to frequentist theory, including unbiasedness, efficiency, consistency, coherence, the conditionality principle, sufficiency, and the likelihood principle. Unfortunately, there is an arbitrariness to these principles, and none of them have been proved to be of universal validity (for example, there is currently a growing literature endorsing the use of biased statistics in some situations; see Efron 1975 and Zellner 1986). Further, with the exception of the concept of sufficiency (which applies to only a limited family of distributions), none of these criteria alone leads to a unique choice for a statistic. Thus in practice more than one criterion must be invoked; but there are no principles specifying the relative importance of the criteria.

Once a statistic is selected, it must be decided how its frequency distribution will be used to assess an hypothesis. To replace the Bayesian notion of the probability of an hypothesis, other real number measures of the plausibility of an hypothesis are introduced, including confidence regions, significance levels, type I and II error probabilities, test size and power, and so on. These all require the consideration of hypothetical data for their definitions.

The resulting frequentist theory is far from unified, and the proliferation of principles and criteria in the theory and the availability of a plurality of methods for answering a single question place the objectivity of the theory in question. This situation is ironic. The frequency definition was introduced to eliminate apparent arbitrariness and subjectivity in the Laplace theory. Yet a large degree of arbitrariness must enter the frequency theory to allow it to address the problems Laplace could address directly.

2.4.2. *Comparison with Intuition.* Once a statistical procedure is chosen in frequentist theory, it is used to assess an hypothesis by calculating its long-term behavior, imagining that the hypothesis is true and that the procedure is applied to each of many hypothetical data sets. But this is strongly at variance with how we intuitively reason in the presence of uncertainty. We do not want a rule that will give good long term behavior; rather, we want to make the best inference possible given the one set of evidence actually available.

Consider the following three examples of everyday plausible inference. When we come to an intersection and must decide whether to cross, or wait for oncoming traffic to pass, we consider whether we will make it across safely or be hit, given the current traffic situation at the intersection. When a doctor diagnoses an illness, he or she considers the plausibility of each of a variety of diseases in the light of the current symptoms of the patient. When a juror attempts to decide the guilt or innocence of a defendant, the juror considers the plausibility of guilt or innocence in light of the evidence actually presented at the trial.

These three examples have a common structure: in the presence of uncertainty, we assess a *variety of hypotheses* (safe crossing or a collision; cold or flu or bronchitis; guilty or innocent) in the light of the *single set of evidence* actually presented to us. In addition, we may have strong, rational prior prejudices in favor of one or more hypotheses. The doctor may know that there is a flu epidemic in progress, or that the patient has had a recurrent viral infection in the past.

Bayes' theorem has just this structure. A variety of hypotheses, specified in I, are each assessed by calculating their posterior probability, which depends both on the prior probability of the hypothesis, and on the probability of the one data set actually observed.

In contrast, the roles of hypothesis and data are reversed in frequentist reasoning. Forbidden the concept of the probability of an hypothesis, the frequentist must assume the truth of a *single* hypothesis, and then invent ways to assess this decision. The assessment is made considering not only the data actually observed, but also *many hypothetical data sets* predicted by the hypothesis but not seen. It is as if the juror tried to decide guilt or innocence by taking into consideration a mass of evidence that might possibly have been presented at the trial but which was not.

In a word, frequentist reasoning assesses decisions to assume the truth of an hypothesis by considering hypothetical data, while the Bayesian approach assesses hypotheses directly by calculating their probabilities using only the data actually observed, the only hypothetical elements of the calculation being the hypotheses themselves.

2.4.3. *Randomness vs. Uncertainty.* Frequentist theory is forced to base inferences on hypothetical data because data, and not hypotheses, are considered to be "random variables." The concept of randomness is at the heart of the theory. But a close inspection of the notion of randomness reveals further difficulties with the frequentist viewpoint.

In frequentist theory, a quantity is random if it unpredictably takes on different values in otherwise identical repetitions of an experiment or among identically prepared members of an ensemble. To explore this concept, we will consider as an example the prototypical random experiment: the flip of a coin. Imagine an experiment specified by the statement, "A fair (*i.e.*, symetrical) coin is flipped." Since either heads or tails can come up in a flip, and since we cannot predict with certainty which will come up, the outcome of a flip is considered random. The probability of a particular outcome—heads, say—is defined as the limiting frequency with which heads comes up in an indefinitely large number of flips. This definition seems to refer to an observable property of the coin. For this reason, frequentist probability appears more objective than Bayesian probability; the latter describes a state of knowledge, while the former seems to describe an observable property of nature.

But certainly the motion of a coin is adequately described by classical mechanics; if we knew the physical properties of the coin (mass, inertia tensor, etc.), the initial conditions of the flip, and exactly how it was flipped, we could predict the outcome with certainty. If the same coin was flipped under precisely the same conditions, the outcome would be the same for each flip. What, then, gives rise to the "randomness" of the outcomes of repeated flips?

If "identical" repetitions of a coin flip experiment produce different outcomes, something must have changed from experiment to experiment. The experiments could not have been precisely identical. Hidden in the adjective, "identical", describing repetitions of an experiment or elements of an ensemble in frequentist theory is the true source of "randomness": the repeated experiments must be identical only in the sense that in each of them we are in the same state of knowledge in regard to the detailed conditions of the flip. Put another way, the description of the experiment is incomplete, so that repetitions of the experiment that agree with our description vary in details which, though not specified in our description, nevertheless affect the outcome. In the coin example, we have specified only that the same (*i.e.*, physically identical) coin be flipped in repeated experiments. But this leaves the initial conditions of the flips, and the precise manner of flipping, completely unspecified. Since the outcome of a flip depends as much on these unspecified details as on the physical properties of the coin, it is unpredictable.

There is variability in the outcome of "random" experiments only because our incomplete knowledge of the details of the experiment permit variations that can alter the outcome. In some cases, our knowledge may not constrain the outcome at all. This could be the case in a coin flipping experiment, where merely specifying that the same coin be flipped leaves so much room for variation that the outcome is totally uncertain, heads and tails being equally probable outcomes for a particular flip. But often our knowledge, though incomplete, sufficiently constrains the experiment so that some general features of the outcome can be predicted, if not with certainty, than at least with high plausibility. The best example is statistical mechanics. There, measurement of the temperature of an equillibrium system provides us with knowledge about its total energy. Though many, many microstates are compatible with the measurement, our limited knowledge of the precise microstate of the system still permits us to make very accurate predictions of, say, its pressure. This is because the vast majority of microstates compatible with our limited knowledge have very nearly identical pressures.

Thus even in frequency theory, situations are described with probability, not because they are intrinsically random or unpredictable, but because we want to make the most precise statements or predictions possible given the variations permitted by the uncertainty and incompleteness of our state of knowledge (Jaynes 1985d). "Randomness", far from being an objective property of an object or phenomenon, is the result of uncertainty and incompleteness in one's state of knowledge. Once this is realized, the frequentist distinction between the uncertainty one may have about the value of a "random variable" and the uncertainty one may have about the truth of an hypothesis appears highly contrived. Randomness, like any uncertainty, is seen to be "subjective" in the sense of resulting from an incomplete state of knowledge.*

Two operational difficulties with frequentist theory clearly indicate that it is as subjective as BPT, and in some contexts even more subjective. First, though probability is defined as long-term frequency, frequency data is seldom available for assignment of probabilities in real problems. In fact, the infinite amount of frequency data required to satisfy the frequentist definition of probability is *never* available. As a result, the frequentist must appeal to an imaginary infinite set of repeated experiments or an imaginary infinite ensemble. Often, which imaginary reference set to choose will not be obvious, as the single data set we wish to analyze can often be considered to be a member of many reasonable reference sets. This subjectivity of frequentist theory has led to statistical paradoxes where simple, apparently well-posed problems have no obvious solution. In the Bayesian approach, where probability assignments describe the state of knowledge defined by the problem statement, such paradoxes disappear (see Jaynes 1973 for an instructive example).

The second operational difficulty arises in the analysis of data consisting of multiple samples of a random quantity. Since frequentist theory requires the consideration of hypothetical data to assess an hypothesis, analysis requires the specification, not only of the phenomenon being sampled and the results of the sample, but also the specification of what other samples might have been seen. These hypothetical samples are needed to specify the reference set for the observed sample, but unfortunately their specification can depend on the thoughts of the experimenter in disturbing ways. This complicated phenomenon is best described by an example (Berger and Berry 1988).

Consider again the flip of a coin, and imagine that a coin has been flipped $N = 17$ times, giving $n_H = 13$ heads and $n_T = 4$ tails. Is this evidence that the coin is biased? Strangely, a frequentist cannot even begin to address this question with the data provided, because it is not clear from these data what the reference set for the data is. If the frequentist is told

* A reader may object at this point, arguing that the success of quantum theory "proves" that phenomena can be intrinsically random. But the successes of quantum theory no more prove the randomness of nature than the success of statistical description of coin flipping proves that coin flipping is intrinsically random, or the fact that a random number algorithm passes statistical tests proves that the numbers it produces (in a purely deterministic fashion!) are random. Indeed, BPT offers much hope in helping us to unravel inference from physics in quantum theory; see Jaynes (1989a,b) for preliminary analyses.

that the experimenter planned beforehand on flipping the coin 17 times, then analysis can proceed, with probabilities determined by embedding the data in a reference set consisting of many sets of 17 flips. But this is not the only way the data could have been obtained. For example, the experimenter may have planned to flip the coin until he saw 4 tails. In that case, the reference set will be many sets of flips differing in their total number, but with each set containing 4 tails.

In the first case, the number of heads (or tails) is the random quantity, and in the second, the total number of flips is the random quantity. Depending on which quantity is identified as random, a different reference set will be used, and different probabilities will result. The results of the analysis thus depend on the stopping rule used by the experimenter. Experiments must therefore be carefully planned beforehand to be amenable to frequentist analysis, and if the plan is altered during execution for any reason (for example, if the experimenter runs out of funds or subjects), the data is worthless and cannot be analyzed. An example is worked out in Section 4.3.1, where it is shown that this so-called *optional stopping problem* can lead to dispute over whether or not an hypothesis is rejected by a given data set.

Intuition rebells against this strange behavior. Surely my conclusions, given the one data set observed, should not depend on what I or anyone else might have done if different data were obtained. And surely, if my plan for an experiment has to be altered (as is often the case in astronomy, where observations can be cut short due to bad weather or fickle satellite electronics), I should still be able to analyze the resulting data. In Bayesian probability theory, the stopping rule plays no role in the analysis, and this has been an important factor in bringing many statisticians over to the Bayesian school of thought (Berger and Berry 1988). There is no ambiquity over which quantity is to be considered a "random variable", because the notion of a random variable and the consequent need for a reference set of hypothetical data is absent from the theory. All that is required is a specification of the state of knowledge that makes the outcome of each element of the data set uncertain.

2.4.4. *The Frequentist Failure.* The frequentist approach to probability theory was motivated by important deficiencies in the Bayesian theory that it replaced. Unfortunately, frequentist theory addressed these deficiencies only by burying them under a superficially more objective facade. When examined more deeply, we see that frequentist theory only exacerbates the ambiguity and subjectivity of the Bayesian theory.

One motivation for frequentist theory was the apparent arbitrariness of the probability axioms. To make the axioms compelling, the frequency definition of probability was introduced. But this definition forbade the use of Bayes' Theorem for the analysis of hypotheses, and the resulting frequentist theory cannot by itself produce unique solutions to well-posed problems. A wide variety of principles and criteria must be added to the theory, each at least as arbitrary as the probability axioms seemed to be.

Another important motivation for frequentist theory was the subjective nature of Bayesian probability assignments, particularly in regard to prior probabilities for hypotheses. Frequentist theory replaces the subjective probability assignments of BPT with relative frequencies of occurence of random variables. But the notion of randomness is itself subjective, dependent on one's state of knowledge in a manner very similar to that of Bayesian probability. In many problems, it is substantially *more* subjective, since the identification of random variables and their probability assignments can depend in disturbing ways on the thoughts of the experimenter. Such is the case in the optional stopping problems just described.

Finally, frequentist theory is badly at odds with the manner in which we intuitively reason in the presence of uncertainty. Rather than evaluate a variety of hypotheses in the light of the available evidence, the theory attempts to evaluate a single hypothesis by considering a variety of hypothetical data. It also ignores any prior information one may have regarding the possible hypotheses.

Frequentist theory has thus failed to address the problems that motivated it, and in fact has exacerbated them. Though it has been used with great success for the analysis of many problems, it is far from unified, and can give anti-intuitive and paradoxical results. These problems signal a deep flaw in the theory, and indicate the need to find a better theory. This new theory should duplicate the successes of frequentist theory, and eliminate its defects.

In the remainder of this paper, we will see that such a better theory exists, and is in fact identical to the original probability theory of Bernoulli, Bayes, Laplace, and their contemporaries, though with a sounder rationale.

2.5 SOME RECENT HISTORY

Sir Harold Jeffreys was one of the earliest critics of the frequentist statistics of his day. But he did more than criticize; he offered an alternative. In his book (Jeffreys 1939) he presented Bayesian solutions to dozens of practical statistical problems. He also tried to provide a compelling rationale for Bayesian probability theory, and although he was not completely successful in this, his mass of intuitively appealing results, many of them inaccessible to frequentists, should have been a clear indication that "something is *right* here." But his work was rejected on philosophical grounds, and has remained largely unnoticed until recently.

In the 1940's and 1950's, R. T. Cox, E. T. Jaynes, and others began to provide the missing rationale for Bayesian probability theory. Their work was little appreciated at first, but others rediscovered some of this rationale, and over the past few decades there has been a slow but steady "Bayesian revolution" in statistics. Astrophysicists have been slow to reap the benefits of this revolution. But in the last 15 years Gull, Skilling, Bretthorst and others have begun working out astrophysical applications of BPT. In the remainder of this paper, we will examine the rationale and foundations of BPT, learn how it is used to address well-posed statistical problems, and then briefly review some of the recent astrophysical applications of BPT.

3. Bayesian Probability Theory: A Mathematical Language for Inference

The difficulties with frequentist theory, particularly its clash with common sense reasoning, lead us to conclude that it is not generally appropriate for the analysis of scientific data. The intuitive appeal of BPT and the mass of successful results from it lead us to suspect that it may be the correct theory. But can a compelling, rigorous mathematical theory be erected upon a concept as apparently vague as the notion that probability is a measure of degree of plausibility?

Happily, the answer is yes. In this section we will see that a small set of compelling qualitative desiderata for a measure of plausibility will be sufficient to completely specify a quantitative theory for inference that is identical to the probability theory used by Laplace and Jeffreys. Specifically, these qualitative desiderata will allow us to *derive* the "axioms" of probability theory, giving them an unassailable status as the correct rules for the manipulation of real number valued degrees of plausibility. We will also recognize that these rules for combination and manipulation—a "grammar" for plausible inference—are only half of the required theory. The other half of the theory is the problem of assigning initial probabilities to be manipulated—the "vocabulary" of the mathematical language—and the desiderata will provide us with rules for unambiguous assignment of probabilities in well-posed problems.

The desiderata make no reference to frequencies, random variables, ensembles, or imaginary experiments. They refer only to the plausibility of propositions. Deductive reasoning, by which we reason from true propositions to other true propositions, will be a limiting case of the theory, and will guide its development. Thus, the theory can be viewed as *the extension of deductive logic to cases where there is uncertainty* (Jaynes 1990a,b). Of course, we are free to use the resulting theory to consider propositions about frequencies in repeated experiments. In this way, connections between probability and frequency, including Bernoulli's theorem and its generalizations, will be derived consequences of the theory, and all the useful results of frequentist theory will be included in the new theory as special cases.

The missing rationale for BPT was first provided by Cox (1946, 1961) and Jaynes (1957, 1958). Similar results were soon found in other forms by other statisticians (see Lindley 1972 for a terse review). We will only be able to describe briefly this profound and beautiful aspect of BPT here. More detailed, highly readable developments of these ideas may be found in Jaynes (1957, 1958), Tribus (1969), Grandy (1987), and Smith and Erickson (1989). A particularly eloquent and complete development will be available in the upcoming book by Jaynes (1990b).

3.1 THE DESIDERATA

Our first desideratum for a theory of plausibility is simple:

(I) Degrees of plausibility are represented by real numbers.

Perhaps there are useful generalizations of the theory to different number systems. But if our theory is to represent something similar to the way we reason, or if we wish to consider it possible to design a computer or robot that follows our quantitative rules, at some point we will have to associate plausibility with some physical quantity, meaning we will have to associate it with real numbers.

Not yet identifying degrees of plausibility with probability, we will indicate the plausibility of proposition A given the truth of proposition C by the symbol $A \mid C$. We will take it as a convention that greater plausibility will correspond to a greater number.

Our second desideratum will be

(II) Qualitative consistency with common sense.

There are several specific ways we will use this; they are noted below. For example, if the plausibility of A increases as we update our background information from C to C' (that is, $A \mid C' > A \mid C$), but our plausibility of B is unaffected ($B \mid C' = B \mid C$), then we expect that the new information can only increase the plausibility that A and B are both true, and never decrease it ($AB \mid C' > AB \mid C$). Effectively, this desideratum will ensure that the resulting theory is consistent with deductive logic in the limit that propositions are certainly true or certainly false.

Our final desideratum is

(III) Consistency.

More explicitly, we want our theory to be consistent in 3 ways.

(IIIa) *Internal Consistency:* If a conclusion can be reasoned out in more than one way, every possible way must lead to the same result.

(IIIb) *Propriety:* We demand that the theory take into account all information provided that is relevant to a question.

(IIIc) *Jaynes Consistency:* Equivalent states of knowledge must be represented by equivalent plausibility assignments. This desideratum, a generalization of the Principle of Indifference, is the key to the problem of assigning prior probabilities. Though it seems obvious once stated, its importance has only been appreciated beginning with the work of Jaynes (1968).

Amazingly, these few compelling desiderata will be sufficient to completely specify the form of Bayesian probability theory.

3.2 THE GRAMMAR OF INFERENCE: THE PROBABILITY AXIOMS

Given two or more propositions, we can build other, more complicated propositions out of them by considering them together. We would like to have rules to tell us how the plausibilities of these new, compound propositions can be calculated from the plausibilities of the original propositions. We will assume for the moment that the original plausibilities are given. The rules we seek will play the role of a "grammar" for our theory.

Some of the ways we can build new propositions out of a set of propositions $\{A, B, C\ldots\}$ include logical negation ($\overline{A}$, "not A"), logical conjunction (AB, "A and B"), and logical disjunction ($A+B$, "A or B"), mentioned above. An example of another important operation is implication: $A \Rightarrow B$ is the proposition, "If A is true, then B follows." The symbolic system governing the combination of propositions like this is *Boolean Algebra.* We want our plausibility calculus to enable us to calculate the plausibility of any proposition built from other propositions using Boolean algebra.

It will come as no surprise to students of computer science that only a subset of the logical operations we have listed is needed to generate all possible propositions. For example, the proposition $A + B$ is identical to the proposition $\overline{\overline{A}\,\overline{B}}$; that is, $A + B$ is true unless both A and B are false. One adequate subset of Boolean operations that will be convenient for us to consider is conjunction and negation. If we can determine how to calculate the plausibility of the negation of a proposition, given the plausibility of the original proposition, and if we can determine how to calculate the plausibility of the conjunction of two propositions from their separate plausibilities, then we will be able to calculate the plausibilities of all possible propositions that can be built from one or more "elementary" propositions.

Our desiderata are sufficient to specify the desired rules for calculation of the plausibility of a negated proposition and of the conjunction of two propositions; not surprisingly, they are the sum rule and product rule, equations (1) and (2) above. We do not have the space to discuss the derivation of these rules fully here. But since the resulting rules are the foundation for probability theory, and since the kind of reasoning by which such quantitative rules are derived from qualitative desiderata is prevalent in Bayesian probability theory, it is important to have an understanding of the derivation. Therefore, we will outline here the derivation of the product rule, and only present the results of the similar derivation of the sum rule; the above mentioned references may be consulted for further details.

3.2.1 *The Product Rule.* We will first look for a rule relating the plausibility of AB to the plausibilities of A and B separately. That is, we want to find $AB \mid C$ given information about the plausibilities of A and B. The separate plausibilities of A and B that may be known to us include the four quantities $u \equiv (A \mid C)$, $x \equiv (B \mid C)$, $y \equiv (A \mid BC)$, and $v \equiv (B \mid AC)$. By desideratum (IIIb), we should use all of these, if they are relevant.

Now we invoke desideratum (II) to try to determine if only a subset of these four quantities is actually relevant. Common sense tells us right away, for example, that $(AB \mid C)$ cannot depend on only one of x, y, u, or v. This leaves eleven combinations of two or more of these plausibilities. A little deeper thought reveals that most of these combinations are at variance with common sense. For example, if $(AB \mid C)$ depended only on u and x we would have no way of taking into account the possibility that A and B are exclusive.

Tribus (1969) goes through all eleven possibilities, and shows that all but two of them exhibit qualitative violations of common sense. The only possible relevant combinations are x and y, or u and v. We can understand this by noting that there are two ways a decision about the truth of AB can be broken down into decisions about A and B. Either we first decide that A is true, and then, accepting the truth of A, decide that B is true. Or, we first decide that B is true, and then make our decision about A given the truth of B. Finally, we note that since the proposition AB is the same as the proposition BA, we can exchange A and B in all the quantities. Doing so, we see that the different pairs, x, y and u, v, merely reflect the ordering of A and B, so we may focus on one pair, the other being taken care of by the commutativity of logical conjunction.

Denoting $(AB \mid C)$ by z, we can summarize our progress so far by stating that we seek a function F such that

$$z = F(x, y). \tag{6}$$

Now we impose desideratum (IIIa) requiring internal consistency. We set up a problem that can be solved two different ways, and demand that both solutions be identical. One such problem is finding the plausibility that *three* propositions, A, B, C, are simultaneously true. The joint proposition ABC can be built two different ways: $ABC = (AB)C = A(BC)$. The first of these equations and equation (6) tell us that $(ABC \mid D) = F[(BC \mid D), (A \mid BCD)]$, where we treat the proposition AB as a single proposition. A similar equation follows from the second equality. Internal consistency then requires that the function F obey the equation,

$$F[F(x, y), z] = F[x, F(y, z)], \tag{7}$$

for all real values of x, y, and z. Crudely, the function F is "associative." The general solution of this functional equation is $F(x, y) = w^{-1}[w(x)w(y)]$, where $w(x)$ is any positive, continuous, monotonic function of plausibility. Thus equation (7) does not uniquely specify F, but only constrains its form. Using this solution in equation (6), our consistency requirement tells us that

$$w(AB \mid C) = w(A \mid BC)w(B \mid C). \tag{8}$$

This looks like the product rule of probability theory. But at this point we cannot identify probability with plausibility, because equation (8) involves the arbitrary function w.

3.2.2 *The Sum Rule.* We may apply similar reasoning to determine how to calculate $w(\overline{A} \mid B)$ from $w(A \mid B)$. Consistency with common sense and internal consistency again lead to a functional equation whose solution implies that

$$w^m(A \mid B) + w^m(\overline{A} \mid B) = 1. \tag{9}$$

Here $w(x)$ is the same function as in (8), and m is an arbitrary positive number. Along the way, it is found that certainty of the truth of a propositition must be represented by $w = 1$, and by convention, $w = 0$ is chosen to represent impossibility.

A new arbitrary element—the number m—has appeared; but since the function w is itself arbitrary, we are free to make a simple change of variables from $w(x)$ to the different monotonic function $p(x) \equiv w^m(x)$, so that we may always write

$$p(A\ B) + p(\overline{A} \mid B) = 1, \tag{10}$$

and

$$p(AB \mid C) = p(A \mid BC)p(B \mid C), \tag{11}$$

in place of equations (8) and (9). Thus the choice of m is irrelevant, and does not offer us any degree of freedom we did not already have in our choice of $w(x)$.

The arbitrary function $p(x)$ indicates that our desiderata do not lead to unique rules for the manipulation of plausibilities. There are thus an infinite number of ways to use real numbers to represent plausibility. But what we *have* shown is that for any such plausibility theory that is consistent with our desiderata, there must be a function $p(x)$ such that the theory can be cast in the form of equations (10) and (11). These equations thus contain the content of all allowed plausibility theories.

Equations (10) and (11) are the "axioms" of probability theory, so we identify the quantity $p(A \mid B)$ as the probability of A given B. That is, probability is here taken to be a technical term referring to a monotonic function of plausibility obeying equations (10) and (11). We have shown that *every allowed plausibility theory is isomorphic to probability theory.* The various allowed plausibility theories may differ in form from probability theory, but not in content. Put another way, since $p(x)$ is a monotonic function of the plausibility x, x is a monotonic function of p. Therefore all allowed plausibility theories can be created by considering all possible functions $x(p)$ and the corresponding transformations of (10) and (11). Of all these theories, differing in form but not in content, we are choosing to use the one specified by $x(p) = p$, since this leads to the simplest rules of combination, equations (10) and (11).

An analogy can be made with the concept of temperature in thermodynamics, a real number encoding of the qualitative notion of hot and cold (Jaynes 1957, 1990b). Different temperature scales can be consistently adopted, each monotonically related to the others, but the Kelvin scale is chosen for the formulation of thermodynamics, because it leads to the simplest expression of physical laws.

3.3 THE VOCABULARY OF INFERENCE: ASSIGNING PROBABILITIES

We have found the rules for combining probabilities, a kind of "grammar" for inference. Now we ask how to assign numerical values to the probabilities to be so combined: we want to define a "vocabulary" for inference. Probabilities that are assigned directly, rather than derived from other probabilities using equations (10) and (11), are called *direct probabilities.* We seek rules for converting information about propositions into numerical assignments of direct probabilities. Such rules will play a role in probability theory analogous to deciding the truth of a proposition in deductive logic. Deductive logic tells us that certain propositions will be true or false *given* the truth or falseness of other assumed propositions, but the rules of deductive logic do not determine the truth of the assumed propositions; their truth must be decided in some other manner, and provided as input to the theory. Direct probabilities are the analogous "input" for probability theory.

It is worth emphasizing that probabilities are *assigned*, not *measured.* This is because probabilities are measures of the plausibilities of propositions; they thus reflect whatever information one may have bearing on the truth of propositions, and are not properties of the propositions themselves. This is reflected in our nomenclature, in that all probability symbols have a vertical bar and a conditioning proposition indicating exactly what was assumed in the assignment of a probability. In this sense, BPT is "subjective," it describes states of knowledge, not states of nature. But it is "objective" in that we insist that equivalent states of knowledge be represented by equal probabilities, and that problems be well-posed: enough information must be provided to allow unique, unambiguous probability assignments.

We thus seek rules for assigning a numerical value to $p(A \mid B)$ that expresses the plausibility of A given the information B. Of course, there are many different kinds of information one may have regarding a proposition, so we do not expect there to be a universal method of assignment. In fact, only recently has it been recognized that finding rules for converting information B into a probability assignment $p(A \mid B)$ is fully half of probability theory. Finding such rules is a subject of much current research.

Rules currently exist for several common types of information; we will outline some of the most useful here. The simplest kind of information we can have about some proposition A_1 is a specification of alternatives to it. That is, we can only be uncertain of A_1 if there are alternatives $A_2, A_3 \ldots$ that may be true instead of A_1; and the nature of the alternatives will have a bearing on the plausibility of A_1. Probability assignments that make use of only this minimal amount of information are important in BPT as objective representations of initial ignorance, and they deserve a special name. We will refer to them as *least informative probabilities* (LIPs).* Probability assignments that make use of information beyond the specification of alternatives we will call *informative probabilities.*

3.3.1 *Least Informative Probabilities.* For many problems, our desiderata are sufficient to specify assignment of a LIP. Consider a problem where probabilities must be asigned to two propositions, A_1 and A_2. Suppose we know from the very nature of the alternatives that they form an exclusive, exhaustive set (one of them, and only one, must be true), but that this is all we know. We might indicate this symbolically by writing our conditioning information as $B = A_1 + A_2$. Since the propositions are exclusive, $p(A_1A_2 \mid B) = 0$, so the sum rule (3) implies that $p(A_2 \mid B) = 1 - p(A_1 \mid B)$. But this does not specify numbers for the probabilities.

Now imagine someone else addressing this problem, but labeling the propositions differently, writing $A_1' = A_2$ and $A_2' = A_1$. This person's conditioning information is $B' = A_1' + A_2' = A_1 + A_2 = B$. Obviously, $p(A_1' \mid B) = p(A_2 \mid B)$, and $p(A_2' \mid B) = p(A_1 \mid B)$. But now note that since B is indifferent to A_1 and A_2, the state of knowledge of this second

* Such probabilities are also referred to as *uninformative probabilities* in the literature.

person regarding A_1' and A_2', including their labeling, is the same as that in the original problem. By desideratum (IIIc), equivalent states of knowledge must be represented by equivalent probability assignments, so $p(A_1' \mid B) = p(A_1 \mid B)$. But this means that $p(A_2 \mid B) = p(A_1 \mid B)$ which, through the sum rule, implies $p(A_1 \mid B) = p(A_2 \mid B) = 1/2$. We finally have a numerical assignment!

This line of thought can ge generalized to a set of N exclusive, exhaustive propositions A_i ($i = 1$ to N), leading to the LIP assignments $p(A_i \mid B) = 1/N$ (Jaynes 1957, 1990b). This is just Bernoulli's principle of indifference mentioned earlier, now seen to be a consequence of consistency when all the information we have is an enumeration of an exclusive exhaustive set of possibilities, with no information leading us to prefer some possibilities over the others.

Note that other information could lead to the same assignment. For example, if we are tossing a coin, and we know only that it has head and tail sides, we would assign least informative probabilities of 1/2 to the possibilities that heads or tails would come up on a single toss. Alternatively, we may have made careful measurements of the shape and inertia tensor of the coin, compelling us to conclude that both outcomes are equally likely and hence to assign *informative* probabilities of 1/2 to both heads and tails. The difference between these assignments would show up once we flipped the coin a few times and then reassessed our probabilities. If three flips gave three heads, in the first state of knowledge this would constitute evidence that the coin was biased and lead us to alter our probability assignment for the next toss, but in the informative state of knowledge it would not, since our information leads us to believe very strongly that the two sides are equally probable.

When the set of possibilities is infinite, as when we want to assign probabilities to the possible values of continuous parameters, the analysis becomes more complicated. This is because it may not be obvious how to transform the original problem to an equivalent one that will help us determine the probability assignment. In the finite discrete case, the only transformation that preserves the identity of the possibilities is permutation, leading to the PI. But in the continuous case, there is an infinite number of possible reparametrizations.

The key to resolving this dilemma is to realize that specifying the possibilities not only provides labels for them, but tells you about their nature. For example, the finite discrete problem we solved assumed that the nature of the possibilities indicated they formed an exhaustive, exclusive set (this implied $p(A_1 A_2 \mid B) = 0$, which we used in the sum rule). In problems with continuous parameters, transformations that lead to equivalent problems that can help one assign a LIP can often be identified by the nature of the parameters themselves. Information unspecified in the problem statement can be as important for this identification as the specified information itself, for problems that differ with respect to unspecified details are equivalent.

For example, suppose we want to find the probability that a marble dropped at random (*e.g.*, by a blindfolded person) will land in a particular region of a small target on the floor. Intuition tells us that the probability is proportional to the area of the region. How could we have established this by logical analysis? Draw an (x, y) coordinate system on the target, so the possibilities are specified by intervals in x and y. Write the probability that the ball will fall in the small area $dxdy$ about (x, y) as $p(x, y, dxdy \mid I) = f(x, y)dxdy$; here I specifies the target region. But nothing in the problem specified an origin for the coordinate system, so our assignments to $p(x, y, dxdy \mid I)$ and $p(x' = x + a, y' = y + b, dx'dy' \mid I)$ must be the same for any choice of a or b. It follows that $f(x, y) = f(x + a, y + b)$ for any (a, b), so $f(x, y) = const$ (the constant is determined by normalization to be 1/[target area]), and the probability is proportional to the area $dx dy$.* Such arguments can produce LIPs for many interesting and useful problems (Jaynes 1968, 1973, 1980; Rosenkrantz 1977; Bretthorst 1989). This tells us that mere specification of the possibilities we are considering, *including their physical meaning*, is a well-defined state of knowledge that can be associated with an unambiguous probability assignment.

3.3.2 *Informative Probabilities, Bayes' Theorem, and Maximum Entropy.* Besides the specification of possibilities, I, we may have some additional information I_A that should lead us to probability assignments different from least informative assignments. Rather than $p(A_i \mid I)$, we seek $p(A_i \mid II_A)$, an informative probability assignment.

One way to find $p(A_i \mid II_A)$ is to use Bayes' Theorem, equation (5), to update our assignments for each of the A_i one at a time. To do this, the additional information $D \equiv I_A$ must be able to play the role of data, that is, it must be meaningful to consider for each A_i the "sampling probability" $p(D \mid IA_i)$ that occurs on the right hand side of BT. Specifically, D has to be a possible consequence of one or more of the A_i considered individually, since each application of BT will require us to assume that one of the A_i is true to calculate the likelihood of the additional information. If the information D is of this type, we do not need any new rules for probability assignment; our rules of combination tell us how to account for the additional information by using BT.**

But data—observation of one of the possible consequences of the A_i—is not the only kind of information we may have about the various possibilities. Our information may refer directly to the possibilities themselves, rather than to their consequences. In our coin example above, the evidence E provided by the measurements took the form of the proposition, "the probability of heads is the same as that of tails." Information like this cannot be used in Bayes' theorem because it does not refer to a consequence of one of the possibilities being considered. For example, here our possibilities are A_1 = heads on the next toss, A_2 = tails. To use BT, we need $p(E \mid IA_1)$, which in words is "the probability that heads and tails are equally probable if heads comes up on the next toss." But since either heads or tails *must* come up on the next toss, asserting that one or the other will come up tells us nothing about their relative probabilities. Put another way, a statement about the relative probabilities of A_1 or A_2 is not a possible logical implication of the truth of either of them, so it cannot be used in BT. Yet such information is clearly relevant for assessing the plausibility of the propositions. We must therefore find rules that will allow us to use information of this kind to make probability assignments.

Such a rule exists for converting certain types of information called *testable information* to a probability assignment. The information E is testable if, given a probability distribution over the A_i, we can determine unambiguously if the distribution is consistent with the information E. In the example above, this was trivially true; E asserted all probabilities were equal, and only one distribution is consistent with this. But in general, there may

* We expect the result to also be invariant with respect to rotations and scale changes. Since the area element is already invariant to these operations, considering them does not alter the result.

** Of course, we must now address the problem of assigning $p(D \mid IA_i)$. This is no different in principle than assigning $p(A_i \mid I)$, and is treated analogously. We start by specifying what other consequences of A_i are possible, assign a LIP, and then account for any other information we have about the possible consequences. In this sense, the distinction between prior probabilities and sampling probabilities is somewhat artificial; both are direct probabilities, and the same rules are used for their assignment.

be many distributions consistent with testable information E. For example, we may know that the mean value of many roles of a die was 4.5 (rather than 3.5 expected for a fair die), and want to use this knowledge to assign probabilities to the six possible outcomes of the next role of the die. This information is testable—we can calculate the mean value of any probability distribution for the six possible outcomes of a roll and see if it is 4.5 or not—but it does not single out one distribution. But despite the multiplicity of distributions consistent with this information, our common sense seems to tell us something about the distribution which represents knowledge of the mean value, *and nothing else*, beyond the fact that it must be one of the distributions with the indicated mean value. For example, we would reject the assignment $\{p_1 = 0.3,\ p_6 = 0.7,$ all other $p_i = 0\}$ as unreasonable, despite the fact that it agrees with the mean value constraint. This is because this particular distribution, by excluding several of the possibilities that the evidence does not compel us to exclude, violates our propriety desideratum (IIIb).

Denote the operation of altering a LIP distribution to reflect testable information E by $\mathcal{O}$, writing $p(H \mid IE) = \mathcal{O}[p(H \mid I); E)]$. Shore and Johnson (1980) have shown that our desiderata are sufficient to specify the operation $\mathcal{O}$. They consider three general types of transformations of a problem into an equivalent one, and show that the requirement that the solutions of these equivalent problems be consistent uniquely specifies $\mathcal{O}$: It selects from among all the possible normalized distributions satisfying the constraints imposed by E, the one with maximum entropy, where the entropy of a finite discrete distribution over exclusive, exhaustive alternatives is defined by

$$H = -\sum_{i=1}^{N} p_i \log p_i, \tag{12}$$

and that of a continuous distribution is defined analogously by

$$H = -\int p(\theta) \log\left(\frac{p(\theta)}{m(\theta)}\right) d\theta, \tag{13}$$

with $m(\theta)$ the LIP assignment for the parameter θ. (Actually any monotonic function of entropy will suffice.) This rule is of enormous practical and theoretical importance; it is called *the maximum entropy principle* (MAXENT).

MAXENT assignments have a number of intuitively appealing interpretations, and were in fact introduced long before the work of Shore and Johnson, based on just such interpretations (Jaynes 1957a,b, 1958; Tribus 1969). For example, we can seek a measure of the amount of uncertainty expressed in a distribution. Arguments originating with Shannon (1948) show that a few compelling desiderata lead to entropy as the appropriate measure of uncertainty. It then seems reasonable to choose from among all distributions satisfying the constraints imposed by E that which is otherwise the most uncertain (*i.e.*, assuming the least in addition to E); this leads to MAXENT (Jaynes 1957, 1958). An example of the use of MAXENT to assign a direct probability distribution will be mentioned in Section 5.1 below; instructive worked examples can be found in Jaynes (1958, 1963, 1978), Tribus (1962, 1969), Fougere (1988, 1989), and Bretthorst (1990).

3.4 THE FREQUENCY CONNECTION

Since there is presumably no end to the types of information one may want to incorporate into a probability assignment, BPT will never be a finished theory. Yet the existing rules are already sufficient for the analysis of uncertainty in many problems in the physical sciences, and in this sense the theory is complete.

Note that the entire theory has been developed without ever even mentioning relative frequencies or random variables. Yet the success of some frequentist methods indicates that there must be a connection between frequency and probability. There is, and such connections arise naturally in the theory, as *derived consequences* of the rules, when one calculates the probabilities of propositions referring to frequencies.

For example, given that the probability of a particular outcome in a single trial of an experiment is p, we can calculate the probability that N repetitions of the experiment will give this outcome n times. This calculation is just what Bernoulli did to prove his large number theorem—the equality of long-term relative frequency and probability in a single trial—mentioned above. Note, however, that the theorem is restricted to the case where each trial is independent of all the others; BPT is not so restricted.

Bernoulli's theorem is an example of reasoning from probability to frequency. But BPT, through Bayes' Theorem, also allows us to reason from observed frequency to probability. The observed frequency constitutes data which we can use to estimate the value of the single trial probability. Such a calculation can be done for any number of trials; it is not restricted to the infinite case. This is of immense importance. In frequentist theory, there is no way to reason from an observed frequency in a finite number of trials to the value of the probability (identified as long-term frequency). This is an awkward situation, because the theory by definition deals with long-term frequencies, but has no way of inferring their values from actual data.

Other connections between frequency and probability can also be derived within BPT by considering other propositions about frequencies. One connection of particular interest is a kind of consistency relationship between Bayes' Theorem and MAXENT. Testable information—information that refers directly to the relative probabilities of the events or hypotheses under consideration—cannot be used in Bayes' Theorem because such information does not refer to possible consequences in a single trial. But if we consider many repeated trials, and reinterpret the testable information as referring to relationships between relative frequencies in many trials rather than probabilities in single trials, we *can* use the information in BT to infer probabilities. For any finite number of trials, precise values of the probabilities will not be specified; rather, BT will provide a distribution for the values. But as the number of trials becomes infinite, the assignment from BT converges to the MAXENT assignment for a single trial (Jaynes 1978, 1982, 1988a; van Campenout and Cover 1981). This result has been used as a justification for MAXENT when the notion of repeated independent trials is meaningful. But MAXENT is not restricted to such cases.

4. Some Well-Posed Problems

Our theory so far is rather abstract; now we take a step toward concreteness by illustrating how two common types of statistical problems are addressed using BPT. We begin by noting that any problem we wish to address with BPT must be *well-posed*, in the sense that enough information must be provided to allow unambiguous assignment of all probabilities required in a calculation. As a bare minimum, this means that an exhaustive set of possibilities must be specified at the start of every problem.* We will call this set the *sample space* if it

* Readers familiar with Kolmogorov's measure theory approach to probability theory will

refers to possible outcomes of an experiment, or the *hypothesis space* if it specifies possible hypotheses we wish to assess.

Using experimental data to analyze parametrized models is an important task for physical scientists. The two classes of well-posed problems we will focus on here are designed for such analysis. They are called *estimation* and *model comparison.*** Estimation explores the consequences of assuming the truth of a particular model, and model comparison assesses a model by comparing it to one or more alternatives. These problems thus differ in regard to the specification of an hypothesis space. We discuss them in turn. Further details may be found in the excellent review of Bretthorst (1990).

4.1 BAYESIAN PARAMETER ESTIMATION

4.1.1 *Parametrized Models.* A parametrized model is just a set of exclusive hypotheses, each labeled by the value of one or more parameters. The parameters may be either continuous or discrete. For simplicity, we will focus attention on a model with a single parameter, θ.

In an estimation problem one assumes that the model is true for *some* (unknown) value of its parameter, and explores the constraints imposed on the parameter by the data using BT. The hypothesis space for an estimation problem is thus the set of possible values of the parameter, $\mathcal{H} = \{\theta_i\}$. The data consist of one or more samples; to make the problem well-posed, the space of possible samples, $\mathcal{S} = \{s_i\}$, must also be specified. The hypothesis space, the sample space, or both can be either discrete or continuous.

Writing the unknown true value of the parameter as Θ, we can use BT to address an estimation problem by calculating the probability that each of the possible parameter values is the true value. To do this, make the following identifications in equation (5). Let D represent a proposition asserting the values of the data actually observed. Let H be the proposition $\Theta = \theta$ asserting that one of the possible parameter values, θ, is the true value (we will abbreviate this by just using the proposed value, θ, as H in BT). The background information I will define our problem by specifying the hypothesis space, the sample space, how the hypotheses (parameter values) and sample values are related, and any additional information we may have about the hypotheses or the possible data. Symbolically, we might write I as the proposition asserting (1) that the true value of the parameter is in $\mathcal{H}$; (2) that the observed data consisting of N samples is in the space $\mathcal{S}^N$; (3) the manner in which the parameter value relates to the data, I_r; and (4) any additional information I_A; that is, $I = (\Theta \in \mathcal{H})(D \in \mathcal{S}^N)I_r I_A$. Of course, the physical nature of the model parameters and the data is implicit in the specification of $\mathcal{H}$, $\mathcal{S}$, and I_r.

Bayes' Theorem now reads*

$$p(\theta \mid DI) = p(\theta \mid I)\frac{p(D \mid \theta I)}{p(D \mid I)}. \tag{14}$$

---

recognize this as similar to the requirement that probabilities refer to elements of a σ-field. The close connection of BPT with Kolmogorov's theory is elaborated on in Jaynes (1990b).

** Some model comparison problems are also called *significance tests* in the literature.

* Bayes' Theorem refers to probabilities, not probability densities. Thus when considering continuous parameters, we technically should write $p(\theta \mid DI)d\theta = p(\theta \mid I)d\theta p(D \mid \theta I)dD/p(D \mid I)dD$, where the p's are here understood to be densities. But the differentials cancel, so equation (14) is correct for densities as well as probabilities.

To use it, we need to know the three probabilities on the right hand side. The prior $p(\theta \mid I)$ and the likelihood $p(D \mid \theta I)$ are both direct probabilities and must be assigned *a priori* using the methods described previously; concrete examples are given below. The term in the denominator is independent of θ. Given the prior and the likelihood, its value can be calculated using the probability axioms as follows.

First, recall that we are assuming the model to be true for *some* value of its parameter(s). Thus the proposition, "$\Theta = \theta_1$ or $\Theta = \theta_2$ or ..." is true, and so has a probability of 1, given I. Writing this proposition symbolically as $(\theta_1 + \theta_2 + \ldots)$, we thus have from axiom (2),

$$\begin{aligned} p(D[\theta_1 + \theta_2 + \ldots] \mid I) &= p(D \mid I)p(\theta_1 + \theta_2 + \ldots \mid I) \\ &= p(D \mid I). \end{aligned} \tag{15}$$

But by expanding the logical product on the left, and again using (2), we also have

$$\begin{aligned} p(D[\theta_1 + \theta_2 + \ldots] \mid I) &= p(D\theta_1 \mid I) + p(D\theta_2 \mid I) + \ldots \\ &= \sum_i p(D\theta_i \mid I) \\ &= \sum_i p(\theta_i \mid I)p(D \mid \theta_i I). \end{aligned} \tag{16}$$

Equations (14) and (15) together imply that

$$p(D \mid I) = \sum_i p(\theta_i \mid I)p(D \mid \theta_i I). \tag{17}$$

This expresses $p(D \mid I)$ in terms of the prior and the likelihood, as promised. Each term in the sum is just the numerator of the posterior probability for each θ_i. Thus in an estimation problem, the denominator of BT is just the normalization constant for the posterior. The probability, $p(D \mid I)$, is sometimes called the *prior predictive distribution*, since it is the probability with which one would predict the data, given only the prior information about the model. Though here it is just a normalization constant, it plays an important role in model comparison, as will be shown below.

The trick we just used to calculate $p(D \mid I)$—inserting a true compound proposition and expanding—arises frequently in BPT. It is just like expanding a function in a complete orthogonal basis; here we are expanding a *proposition* in a complete "orthogonal" basis. This trick is important enough to deserve a name: it is called *marginalization.** The quantity $p(D \mid I)$ is sometimes called the *marginal likelihood* or the *global likelihood.* Of course, when dealing with continuous parameters, the sum becomes an integral, and (17) reads

$$p(D \mid I) = \int p(\theta \mid I)p(D \mid \theta I)d\theta. \tag{18}$$

Inserting the various probabilities into BT, we can calculate the posterior probabilities for all values of the parameters. The resulting probability distribution represents our inference about the parameters completely. As an important matter of interpretation, note that in this and any Bayesian distribution, it is the *probability* that is distributed, not the parameter

* This name is historical, and refers to the practice of listing joint probabilities of two discrete variables in a table, and listing in the *margins* the sums across the rows and columns.

(Jaynes 1986a). Stating an inference by saying something like "the parameter is distributed as a gaussian..." is misleading. The parameter had a single value during the experiment, and we want to infer something about this single value. We do not know it precisely, but the data tell us something about it. We express this incomplete knowledge by spreading our belief regarding the true value among the possible values according to the posterior distribution.

4.1.2 *Summarizing Inferences.* We can present the full posterior graphically or in a table. But usually we will want to summarize it with a few numbers. This will be especially true for multiparameter problems, where graphical or tabular display of the full posterior may be impossible because of the dimension of the parameter space. There are various ways to summarize a distribution, depending on what is going to be done with the information.

One summarizing item is a "best fit" value for the parameter. Which value to choose will depend on what is meant by "best". One obvious choice is the most probable parameter value, the *mode.* It is the best in the sense of being the single value one has greatest confidence in. But its selection does not reflect how our confidence is spread among other values at all. For example, if a distribution is very broad and flat with a small "bump" to one side, the mode will not be a good summarizing item, since most of the probability will be to one side of it. In this case, the *mean* of the distribution would be a better "best" value. On the other hand, if the distribution has two narrow peaks, the mean could lie between them at a place where the probability is small or even zero. So some common sense has to be used in choosing a best fit value.

There is a formal theory for making decisions about best fit values; it is called *decision theory* (Eadie *et al.* 1971; Berger 1985). Decision theory is very important in business and economics where one frequently must make a decision about a best value and then act as if it were true. But in the physical sciences, best values are usually just a convenient way to summarize a distribution. For this, common sense is usually a good enough guide, and a formal decision theory is not needed.

Besides a best value, it is useful to have a simple measure of how certain one is of this value. Again, decision theory can be brought to bear on this problem, but the traditional practice of quoting either the standard deviation (second moment about the mean) or the size of intervals containing specified fractions of the posterior probability is usually adequate. Of course, since probability is a measure of the plausibility of the parameter values, when we quote an interval, we should choose its boundaries so that all values inside it have higher probability than those outside. Such an interval is called a *credible region*, or a *highest posterior density interval* (HPD interval) when it is used to summarize the posterior distribution of one or more continuous parameters.*

In multiparameter problems, we may be interested only in certain subsets of the parameters. Depending on how many parameters are of interest, the distribution may be summarized in different ways. If the values of all of the parameters are of interest, a best fit point can be found straightforwardly by locating the mean or mode in the full parameter space. To quantify the uncertainty in the best fit point, all of the second moments can be calculated and presented as an $N \times N$ matrix; but off-diagonal moments are not an intuitively appealing measure of the width of the distribution. Alternatively, one can calculate

* Sometimes the name *confidence interval* is given to credible intervals, and indeed it reflects well the intuitive meaning of a credible interval. But "confidence interval" has a technical meaning in frequentist theory that is different from its meaning here, and so we avoid this term.

an HPD region in the full parameter space, and present it by plotting its *projection* onto one, two, or three dimensional subspaces of the full parameter space. Some information is lost in such projections—the HPD region cannot be uniquely reconstructed from them — but they conservatively summarize the HPD region in the sense that they will show the full range of parameter values permitted in the region. They will also probably indicate the nature of any correlations among parameters, though two dimensional *cross sections* of the HPD better reveal correlations.

If only a subset of the parameters is of interest, the other parameters are called *nuisance parameters* and can be eliminated from consideration by marginalization. For example, if a problem has two parameters, θ and ϕ, but we are interested only in θ, then we can calculate $p(\theta \mid DI)$ from the full posterior $p(\theta\phi \mid DI)$ by using the trick we used to calculate $p(D \mid I)$. The result is $p(\theta \mid DI) = \int d\phi\, p(\theta\phi \mid DI)$; this is called the marginal distribution for θ. Using BT and the product rule, the marginal distribution can be written

$$p(\theta \mid DI) = \frac{1}{p(D \mid I)} \int p(\phi \mid I) p(\theta \mid \phi I) p(D \mid \theta\phi I) d\phi. \tag{19}$$

Marginalization is of great practical and theoretical importance, because it can often be used to significantly reduce the dimensionality of a problem by eliminating nuisance parameters, making numerical calculations and graphical presentation much more tractable. Denied the concept of the probability of a parameter value, frequentist theory is unable to deal with nuisance parameters, except in special cases where intuition has led to results equivalent to marginalization (Lampton, Margon, and Bowyer 1976; Dawid 1980). Marginalization is thus an important technical advantage of BPT. It is a quantitative way of saying, in regard to the uninteresting parameters, "I don't know, and I don't care."

As useful and necessary as summaries of distributions are, we must always remember that the entire distribution is the full inference, not the summary.

4.2 BAYESIAN MODEL COMPARISON

Estimation problems assume the truth of the model under consideration. We often would like to test this assumption, calling into question the adequacy of a model. If the model is inadequate, then some alternative model must be better, and so BPT assesses a model by comparing it to one or more alternatives. This is done by assuming that some member of a set of competing models is true, and calculating the probability of each model, given the observed data, with BT. As we will see in Section 5, the Bayesian solution to this problem provides a beautiful quantification of Ockham's razor: simpler models are automatically preferred unless a more complicated model provides a significantly better fit to the data.

To use BT for model comparison, I asserts that one of a set of models is true. This means that I will have all the information needed to address an estimation problem for each model, plus any additional information I_0 that may lead us to prefer certain models over others *a priori*. Denote the information needed to address an estimation problem with model number k as I_k ($k = 1$ to M). Then symbolically we may write $I = (I_1 + I_2 + \ldots + I_M)I_0$. Let D stand for the data, and let k stand for the hypothesis, "Model number k is true." BT can now be used to calculate the *probability of a model:*

$$p(k \mid DI) = p(k \mid I) \frac{p(D \mid kI)}{p(D \mid I)} \tag{20}$$

To use this, we must calculate the various probabilities. Here we will consider the case where we have no prior information preferring some models over the other, so the prior is $p(k \mid I) = 1/M$.

To calculate $p(D \mid kI)$, note that since k asserts the truth of model number k, only the information I_k in I is relevant: $kI = k(I_1 + I_2 + \ldots)I_0 = I_k$. Thus, $p(D \mid kI) = p(D \mid I_k)$, the marginal likelihood for model k, described above. Labeling the parameters of model k by θ_k, this can be calculated from

$$p(D \mid kI) = \int d\theta_k p(\theta_k \mid I_k) p(D \mid \theta_k I_k). \tag{21}$$

To calculate $p(D \mid I)$, we marginalize by inserting the true proposition ($k = 1 + k = 2 + \ldots$). This gives

$$p(D \mid I) = \sum_k p(k \mid I) p(D \mid kI). \tag{22}$$

As in an estimation problem, $p(D \mid I)$ is simply a normalization constant. In model comparison problems, we can avoid having to calculate it by focusing attention on the ratios of the probabilities of the models, rather than the probabilities themselves. Such ratios are called *odds*, and the odds in favor of model k over model j we will write as $O_{kj} \equiv p(k \mid DI)/p(j \mid DI)$. From the above equations, the odds can be calculated from

$$\begin{aligned} O_{kj} &= \left[\frac{p(k \mid I)}{p(j \mid I)}\right] \frac{\int d\theta_k p(\theta_k \mid I_k) p(D \mid \theta_k I_k)}{\int d\theta_j p(\theta_j \mid I_j) p(D \mid \theta_j I_j)} \\ &\equiv \left[\frac{p(k \mid I)}{p(j \mid I)}\right] B_{kj}, \end{aligned} \tag{23}$$

where the factor in brackets is called the *prior odds* (and is here equal to 1), and B_{kj} is called the *Bayes factor*. The Bayes' factor is just the ratio of the prior predictive probabilities, $B_{kj} = p(D \mid I_k)/p(D \mid I_j)$.

Equation (22) is the solution to the model comparison problem. In principle, such problems are little different from estimation problems; Bayes' theorem is used similarly, with an enlarged hypothesis space. In practice, more care must be exercised in calculating probabilities for models than for model parameters when there is little prior knowledge of the values of the parameters of the models under consideration. This is illustrated by way of an example in Section 5 below.

4.3 PROBLEMS WITH FREQUENTIST MODEL ASSESSMENT

As a basic principle for the design of well-posed problems, we have demanded that an exhaustive set of possibilities be specified at the beginning of any problem. In an estimation problem, this is accomplished by asserting the truth of a model, so that the hypotheses labeled by values of model parameters form an exhaustive set of alternatives. In model comparison, we satisfied this principle by explicitly specifying a set of competing models. How does this compare with frequentist methods for estimation and model assessment?

One of the most important frequentist statistics in the physical sciences is the χ^2 statistic. It is used both for parameter estimation and for assessing the adequacy of a model (see, *e.g.*, Lampton, Margon, and Bowyer 1976). The use of χ^2 for obtaining best fit parameters and confidence regions is mathematically identical to Bayesian parameter estimation for models with gaussian "noise" probabilities and with flat priors for the parameters. This is because χ^2 is proportional to the log of the likelihood when there is gaussian noise, and BT tells us that the posterior is proportional to the likelihood when the priors are flat.

Besides being used for estimation, frequentist theory also uses the χ^2 statistic to assess an hypothesis by calculating the tail area above the minimum χ^2 value in the χ^2 distribution—the probability of seeing a χ^2 value as large or larger than the best fit value if the model is true with its best fit parameters. This is very different in character from the Bayesian approach to model assessment. In particular, in this χ^2 goodness-of-fit (GOF) test and other GOF tests (*e.g.*, the Kolmogorov-Smirnov test, the Smirnov-Cramer-von Mises test, etc.) *no explicit alternatives are specified.* At first sight, this seems to be an important advantage of frequentist theory, because it may be difficult to specify concrete alternatives to a model, and because it appears restrictive and subjective to have to specify an explicit set of alternatives to assess a model.

Deeper thought reveals this apparent advantage of frequentist GOF tests to be a defect, a defect that can be all the more insidious because its manifestations can be subtle and hidden. The resulting problems with GOF tests and other frequentist procedures that rely on tail areas began to be discussed openly in the statistics literature at least as early as the late 1930s (Jeffreys 1939), and continue to be expounded today (see Berger and Berry 1988 and references therein). Disturbingly, they are seldom mentioned in even the most recent frequentist texts. We will briefly note some of these important problems here.

4.3.1. *Reliance on Many Hypothetical Data Sets.* The χ^2 GOF test is based on the calculation of the probability P that χ^2 values equal to or larger than that actually observed would be seen. If P is too small (the critical value is usually 5%), the model is rejected. The earliest objections to the use of tests like χ^2 focused on the reliance of such tests, not only on the probability of the observed value of the statistic, but on the probability of values that have not been observed as well. Jeffreys (1939) raised the issue with particular eloquence:

> *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it.

Indeed, many students of statistics find that the unusual logic of P-value reasoning takes some time to "get used to."

Later critics strengthened and quantified Jeffreys' criticism by showing how P-value reasoning can lead to surprising and anti-intuitive results. This is because the reliance of P-values on unobserved data makes them dependent on what one believes such data might have been. The intent of the experimenter can thus influence statistical inferences in disturbing ways, a phenomenon alluded to in Section 2.4.3 above, and known in the literature under the name *optional stopping.* Here is a simple example (after Iverson 1984, and Berger and Berry 1988).

Suppose a theorist predicts that the number of A stars in an open cluster should be a fraction $a = 0.1$ times the total number of stars in that cluster. An observer who wants to test this hypothesis studies the cluster and reports that his observations of 5 A stars out of 96 stars observed rejects the hypothesis at the 95% level, giving a χ^2 P-value of 0.03. To check the observer's claim, the theorist calculates χ^2 from the reported data, only to find that his hypothesis is acceptable, giving a P-value of 0.12. The observer checks his result, and insists he is correct. What is going on?

The theorist calculated χ^2 as follows. If the total number of stars is $N = 96$, his prediction is $n_A = 9.6$ A stars and $n_X = 86.4$ other stars. Pearson invented the χ^2 test for just such a problem; χ^2 is calculated by squaring the difference between the observed and expected numbers for each group, dividing by the expected numbers, and summing (Eadie *et al.* 1971). From the predictions and the observations, the theorist calculates $\chi^2 = 2.45$, which has a P-value of 0.12, using the χ^2 distribution for one degree of freedom (given N, n_X is determined by n_A, so there is only one degree of freedom).

Unknown to the theorist, the observer planned his observations by deciding beforehand that he would observe until he found 5 A stars, and then stop. So instead of the number of A and non-A stars being random variables, with the sample size N being fixed, the observer considers $n_{A,obs} = 5$ to be fixed, and the sample size as being the random variable. From the negative binomial distribution, the expected value of N is $5/a = 50$, and the variance of the distribution for N is $5(1-a)/a^2 = 450$. Using the observed $N = 96$ and the asymptotic normality of the negative binomial distribution, these give $\chi^2 = 4.70$ with one degree of freedom, giving a P-value of 0.03 as claimed.

The reason for the difference between the two analyses is due to different ideas of what other data sets might have been observed, resulting in different conclusions regarding what observed quantities should be treated as "random." But why should the plans of the observer regarding when to stop observing affect the inference made from the data? If, because of poor weather, his observing run had been cut short before he observed 5 A stars, how then should his analysis proceed? Should he include the probability of poor weather shortening the observations? If so, shouldn't he then include the probability of poor weather in the calculation when he *is* able to complete the observations?

Because of problems like this, some statisticians have adopted the *conditionality principle* as a guide for the design of statistical procedures. This principle asserts that only the data actually observed should be considered in a statistical procedure. Birnbaum (1962) gave this principle an intuitively compelling rationale through a *reductio ad absurdum* as follows. Suppose there are two experiments that may be performed to assess an hypothesis, but that only one can be performed with existing resources. A coin is flipped to determine which experiment to perform, and the data is obtained. If the data are analyzed with any method relying on P-values, we have to consider what other data might have been observed. But in doing so, should we consider the possibility that the coin could have landed with its other face up, and therefore consider all the data that might have come from the *other* experiment in our analysis? Most people's intuition compels them to assert that only data from the experiment actually performed should be relevant. Birnbaum argued that if this is accepted, the conditionality principle follows, and only the one data set actually obtained should be considered. Of course, BT obeys the conditionality principle, since it uses only the probability of the actually observed data in the likelihood and the marginal likelihood.

In the same work, Birnbaum shows that another technical criterion (sufficiency) already widely employed by statisticians implies with the conditionality principle that all the evidence of the data is contained in the likelihood function. This *likelihood principle* is also adhered to in BPT. Though widely discussed in the literature (see Berger and Wolpert 1984, and references therein), the likelihood principle has so far had little effect on statistics in the physical sciences.

As Jeffreys himself noted (Jeffreys 1939), the fundamental idea behind the use of P-values—that the observation of data that depart from the predictions of a model call the model into question—is natural. It is the expression of this principle in terms of P that is unacceptable. The reason we would want to reject a model with large χ^2 is not that χ^2 is large, but that large values of χ^2 are less probable than values near the expected value. But very small values, with P near 1, are similarly unexpected, a fact not expressed by P-values.*

We have argued that only the probability of the actually observed χ^2 value is relevant. But this probability is usually negligible even for the expected value of χ^2 or any other GOF statistic. P-values adjust for this by considering hypothetical data. Bayes' Theorem remedies the problem by dividing this small probability by another small probability, the marginal likelihood. But the use of BT requires the specification of alternative hypotheses. The apparent absence of such alternatives in frequentist tests is the basis for the next two criticisms of such tests.

4.3.2. *Reliance on a Single Hypothesis.* GOF tests require one to assume the truth of a single hypothesis, without reference to any alternatives. But this is clearly a weakness when such tests are used to evaluate parameterized models, because they require one to assume, not only that the model under consideration is true, but also that the best fit parameter values are the true values. This raises two questions regarding the logic of GOF tests.

First, if we decide to reject the hypothesis, then certainly we must reject probabilities calculated conditional on the truth of the hypothesis. But the P-value itself is such a probability! Thus when an hypothesis is rejected, tail area reasoning seems to invalidate itself. Bayes' theorem avoids this problem because rather than calculating probabilities of hypothetical data conditional on a single hypothesis, it calculates the probabilities of various hypotheses conditional on the observed data (Jaynes 1985c, 1986a).

Second, even if the model is true or adequate, it is almost certain that the best fit parameter values are not the true values. This again seems to put the logical status of the test in question, since its probabilities must always be calculated conditional on an hypothesis we are virtually certain is false. One might appeal to intuition and argue that if the model is rejected with its best fit parameter values, then surely the model as a whole must be rejected. But if the best fit model is acceptable, the acceptability of the model as a whole does not necessarily follow. For example, we feel that a model that produces a good fit over a wide range of its parameter space is to be preferred to a model with the same number of parameters but which requires parameter values to be carefully "fine-tuned" to explain the data; the data are a more natural consequence of the former model. Frequentist GOF tests have no way to account for such characteristics of a model, since they consider only the best fit parameter values.

Bayesian methods account for our uncertainty regarding the model parameters naturally and easily through marginalization. The probability of model k is proportional to its marginal likelihood $p(D \mid I_k)$, which takes into account all possible parameter values. Bayesian methods also take this uncertainty into account when making predictions about future data. The probability of seeing data D', given the observation of data D and the

* Many astronomers seem to consider a fit with, say, $\chi^2 = 16$ with 25 degrees of freedom ($P = 0.915$) to be better than one with, say, $\chi^2 = 27$ ($P = 0.356$); in fact, the former value of χ^2 is 40% *less* probable than the latter, despite the fact that its P-value is over 2.5 times greater. To account for this, Lindley (1965) has advocated a 2-tailed χ^2 test, in which P is calculated by integrating over all less probable values of χ^2, not just greater values.

truth of model k, is easily shown to be

$$p(D' \mid DI_k) = \int d\theta_k p(\theta_k \mid DI_k) p(D' \mid \theta_k I_k). \tag{24}$$

This is called the *posterior predictive distribution*, and it is derived by marginalizing with respect to θ_k. It says that the probability of D' is just its average likelihood, taking the average over the posterior distribution for θ_k based on the observed data, D. Surely any model assessment based on how unexpected the data are should rely on the marginal likelihood or the predictive distribution, and not on distributions assuming the truth of particular parameter values.

4.3.3. *Implicit Alternatives: The Myth of Alternative-Free Tests.* Though GOF tests appear to make no assumptions about alternatives, in fact the selection of a test statistic corresponds to an implicit selection of a class of alternatives. For example, the χ^2 statistic is the sum of the squares of the residuals, and thus contains none of the information present in the order of the data points. The χ^2 test is thus insensitive to patterns in the residuals, and will not account for small but statistically significant trends or features in the residuals in its assessment of an hypothesis. Thus the χ^2 test implicitly assumes a class of alternatives for which the data are exchangeable, so that their order is irrelevant (Jaynes 1985c).

This characteristic of test statistics has long been recognized, beginning with the work of Neyman and Pearson (the inventor of χ^2) in 1938. It has led to the characterization of statistical tests, not only by P-values, but also by their *power*, the probability that they correctly identify a true model against a particular alternative. But though modern statistical theory insists that tests be characterized both by P-values and by their power, few statistics texts for the physical sciences even mention the concept of power (Eadie *et al.* 1971 is a notable exception), and as a rule, the power of a test is never considered by physical scientists.

It is a far from trivial asset of Bayesian probability theory that by its very structure it forces us to specify a set of alternative hypotheses explicitly, in I, rather than implicitly in the choice of a test statistic.

4.3.4. *Violation of Consistency and Rationality.* The many problems of alternative free GOF tests and tail area reasoning should come as no surprise in the light of the Cox-Jaynes derivation of the probability axioms. This is because the P-value is an attempt to find a real number measure of the plausibility of an hypothesis. But in Section 3 we saw that any such measure that is consistent with common sense and is internally consistent must be a monotonic function of the probability of the hypothesis. In general, a P-value will not be a monotonic function of the probability of the hypothesis under consideration, and so anti-intuitive and paradoxical behavior of P-value tests should be expected.

It also comes as no surprise that some of the most useful tail area tests have been shown to lead to P-values that *are* monotonic functions of Bayesian posterior probabilities with specific classes of alternatives, thus explaining the historical success of these tests (see, *e.g.*, Zellner and Siow 1980; Bernardo 1980). Of course, the Bayesian counterparts to these tests are superior to the originals because they reveal the assumed alternatives explicitly, showing how the test can be generalized; and because they produce a probability for the hypothesis being assessed, a more direct measure of the plausibility of the the hypothesis than the P-value.

5. Bayesian and Frequentist Gaussian Inference

We will now apply BPT to a common and useful statistical problem: infering the amplitude of a signal in the presence of gaussian noise of known standard deviation σ, given the values x_i of N independent measurements. We will solve this problem with both frequentist and Bayesian methods. The Bayesian result is mathematically identical to the familiar frequentist result, but it is derived very differently and has a different interpretation. Bayesian and frequentist results will *not* be identical in general; our study will tell us about the conditions when identity may be expected.

5.1 THE STATUS OF THE GAUSSIAN DISTRIBUTION

We begin by first discussing the model: in what situations is a "gaussian noise" model appropriate?

In frequentist theory, the noise model should be the frequency distribution of the noise in an infinitely large number of repetitions of the experiment. But there is seldom even a moderate finite number of repetitions available to provide us with frequency data, so some other justification for the gaussian distribution must be offered. Sometimes it is used simply because it has convenient analytical properties. Often it is justified by appealing to the central limit theorem (CLT), which states that if the noise in a single sample is the the result of a number of independent random effects, the gaussian distribution will be a good approximation to the actual frequency distribution of the noise in many trials regardless of the distributions for each of the effects, if the number of independent effects is large. But in general noise is not the result of a large number of independent effects; and even when it is, there is no way to be sure that the gaussian distribution is an adequate approximation for a finite number of effects without knowing the distributions describing each effect.

Bayesians interpret a noise distribution as an expression of our state of knowledge about the size of the noise contribution in the single data set actually being considered. Of course, if frequency data from many independent repetitions of an experiment are available, they will be relevant for assigning a noise distribution. But such data is typically not available, and the methods for assigning direct probabilities described in Section 3 must be used to find the quantitative expression of our state of knowledge about the noise.

Usually by noise we mean effects from unknown causes that we expect would "average out": positive and negative values are equally likely. Thus we expect the mean of the noise distribution to be zero. Additionally, we usually expect there to be a "typical scale" to the noise; we do not expect very large noise contributions to be as probable as smaller ones. Thus we expect the noise distribution to have some finite standard deviation, though we may not have a good idea what its value should be.

The information that a distribution have zero mean and standard deviation σ is testable: given any distribution, we can see if its mean vanishes and if its second moment is σ^2. Thus we can use MAXENT to assign the noise distribution, using the zero mean and σ as constraints. The resulting distribution is the gaussian distribution! Thus in BPT the gaussian distribution is appropriate whenever we know or consider it reasonable to assume that the noise has zero mean and finite standard deviation, but we do not have further details about it (Jaynes 1985c, 1987; Bretthorst 1988b, 1990). Additionally, we often need not specify the actual value of σ if it is not known. We can consider it a parameter of our model, and estimate it from the data or marginalize it away.

The status of the gaussian distribution in BPT is thus very different from its status in frequentist theory. In BPT it simply represents the most conservative distribution consistent with minimal information about the noise phenomena, and it will be appropriate whenever such information is all we know about the noise, regardless of whether or not the CLT applies. This accounts for the great practical success of models assuming gaussian noise.

The reasoning used in BPT to assign the gaussian distribution can be easily generalized to other situations. For example, there is no single distribution for directional data on a circle or on a sphere that has all of the properties of the gaussian distribution on a line, and so there is some controversy over what distributions are the counterparts of the gaussian distribution for directional data (Mardia 1972). But if our knowledge is restricted to specification of a mean direction and an expected angular scale for deviations, then MAXENT identifies the correct distributions as the von Mises distribution for circular data and the Fisher distribution for spherical data (these distributions are discussed in Mardia 1972).

Having justified our model, we now describe the development of frequentist and Bayesian procedures for estimating the amplitude μ of a signal for which there are N measurements x_i contaminated with noise with standard deviation σ.

5.2 ESTIMATING THE SIGNAL

5.2.1. *The Frequentist Approach.* In frequentist theory, since the signal strength μ is not a random variable taking on values according to a distribution, we are forbidden to speak of a probability distribution for μ. But the x_i *are* considered random variables, and their distribution is just gaussian,

$$p(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{x_i - \mu}{\sigma}\right]^2. \tag{25}$$

To estimate μ, the frequentist must choose a statistic—a function of the random variables x_i—and calculate its distribution, connecting it with μ. A few of the many possible statistics for estimating μ include x_3 (the value of the third sample); $(x_1 + x_N)/2$ (the mean of the first and last samples); the median of the observations; or their mean, $\bar{x} = \sum_i x_i/N$.

To choose from among these or other statistics, some criteria defining a "best" statistic must be invoked. For example, it is often required that a statistic be *unbiased*, that is, that the average value of the statistic in many repeated measurements converges to the true value of μ. But the distributions for all of the above mentioned statistics can be calculated and reveal them *all* to be unbiased, so additional criteria must be specified. Unfortunately, all such criteria have a certain arbitrariness to them. For example, the criterion of unbiasedness focuses on the long-term mean value of the statistic. But the long-term median or most probable value would also reflect the intuitive notion behind the idea of bias, and in general would lead to a different choice of "best" statistic.

Of course, intuition suggests that to estimate the mean of a distribution, one should take the mean of the sample.* Various criteria of frequentist theory are chosen with this in mind, and eventually identify the mean, $\bar{x}$, as the "best" estimate of μ.

Now we would like to know how certain we are that $\bar{x}$ is near the unknown true value of μ. Interestingly, frequentist theory treats this problem as logically distinct from estimating

* Such intuitive reasoning does not always lead to good statistics; see Section 8.2.

best values, and in general completely different statistics and procedures can be used for these problems. In this simple gaussian problem, intuition again compels us to focus our attention on $\bar{x}$, and a *confidence region* for μ is found from $\bar{x}$ as follows.

Suppose μ were known. Then the distribution for $\bar{x}$ can be calculated from equation (25); a somewhat tedious calculation gives

$$p(\bar{x} \mid \mu) = \left[\frac{N}{2\pi\sigma^2}\right]^{1/2} \exp\left[-\frac{N}{2\sigma^2}(\bar{x}-\mu)^2\right]. \tag{26}$$

This distribution is a gaussian about μ with standard deviation $\sigma/\sqrt{N}$. With μ known, we can calculate the probability that $\bar{x}$ is in any interval $[a, b]$ by integrating (26) over this region with respect to $\bar{x}$. But when μ is unknown, this is not possible. However, since $p(\bar{x} \mid \mu)$ is a function only of the difference between $\bar{x}$ and μ, we can always calculate the probability β that $\bar{x}$ lies in some interval *relative to the unknown mean*, such as the interval $[\mu + c, \mu + d]$, and the result will be independent of μ. Using equation (26), we find

$$\beta \equiv p(\mu + c < \bar{x} < \mu + d) = \frac{1}{2}\left[\operatorname{erf}\left(\frac{d}{\sigma\sqrt{2/N}}\right) - \operatorname{erf}\left(\frac{c}{\sigma\sqrt{2/N}}\right)\right]. \tag{27}$$

For a given β of interest, there are many choices of c and d that satisfy (27). For example, for the "1σ" value $\beta = 68\%$, we may choose any of $[c, d] = [-\infty, x]$, $[-\sigma/\sqrt{N}, \sigma/\sqrt{N}]$, or $[-x, \infty]$. *A priori*, there is no reason to prefer any one of these to the others in frequentist theory, and again some criterion must be invoked to select one as "best" (Lampton, Margon, and Bowyer 1976). Popular criteria are to choose the smallest interval satisfying (27), or the symmetric one. For this problem, both criteria lead to the choice $[-\sigma/\sqrt{N}, \sigma/\sqrt{N}]$

In summary, the frequentist inference about μ might be stated by estimating μ with $\bar{x}$, and giving a "1σ" confidence interval of $\bar{x} \pm \sigma/\sqrt{N}$, the familiar "root N" rule.

5.2.2. *The Bayesian Approach.* The Bayesian solution to this problem is to simply calculate the posterior distribution for μ using BT. We begin by specifying the background information I. I will contain the information leading to the MAXENT assignment of a gaussian distribution for a single datum, equation (25). I will also specify the hypothesis space, a range of possible values for μ. We will assume we know μ to be in the range $[\mu_{\rm min}, \mu_{\rm max}]$; we discuss this assumption further below.

With this I, we must assign the prior and the likelihood. A simple consistency argument (Jaynes 1968) shows that the LIP assignment for μ is the uniform density,

$$p(\mu \mid I) = \frac{1}{\mu_{\rm max} - \mu_{\rm min}}. \tag{28}$$

The likelihood follows from (25) using the product rule: the joint probability of the N independent observations is the product of their individual probabilities,

$$\begin{aligned} p(\{x_i\} \mid \mu I) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right] \\ &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left[-\frac{Ns^2}{2\sigma^2}\right] \exp\left[-\frac{N}{2\sigma^2}(\bar{x}-\mu)^2\right], \end{aligned} \tag{29}$$

where we have separated out the dependence on μ by expanding the argument of the exponential and completing the square. Here s^2 is the sample variance, $s^2 = \sum_i (x_i - \bar{x})^2/N$.

Together, the prior and the likelihood determine the marginal likelihood to be

$$p(\{x_i\} \mid I) = \frac{1}{\sqrt{N}}(\sigma\sqrt{2\pi})^{1-N} \exp\left[-\frac{Ns^2}{2\sigma^2}\right] \frac{\mathrm{erf}\left(\frac{\bar{x}-\mu_{\max}}{\sigma\sqrt{2/N}}\right) - \mathrm{erf}\left(\frac{\bar{x}-\mu_{\min}}{\sigma\sqrt{2/N}}\right)}{2(\mu_{\max} - \mu_{\min})}, \tag{30}$$

where the error functions arise from integrating (29) with respect to μ over the interval $[\mu_{\min}, \mu_{\max}]$. Equation (30) is constant with respect to μ.

With these probabilities, BT gives our complete inference regarding μ as

$$p(\mu \mid \{x_i\}I) = \left[\frac{\mathrm{erf}\left(\frac{\bar{x}-\mu_{\max}}{\sigma\sqrt{2/N}}\right) - \mathrm{erf}\left(\frac{\bar{x}-\mu_{\min}}{\sigma\sqrt{2/N}}\right)}{2}\right] \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{N}{2\sigma^2}(\bar{x}-\mu)^2\right]. \tag{31}$$

This is just a gaussian about $\bar{x}$ with standard deviation $\sigma/\sqrt{N}$, truncated at $\mu_{\min}$ and $\mu_{\max}$. The factor in brackets is the part of the normalization constant due to the truncation.

As a best fit value, we might take the mode of the distribution, $\mu = \bar{x}$ (assuming that $\bar{x}$ is in the allowed range for μ). Alternatively, we might take the mean. The mean value, and the limits of any HPD region, will depend on our prior range for μ. But as long as the prior range is large compared to $\sigma/\sqrt{N}$, the effect of the prior range will be negligible. In fact, if we are initially completely ignorant of μ, we can consider the limit $[\mu_{\min}, \mu_{\max}] \to [-\infty, \infty]$, for which the term in brackets becomes equal to 1. The mean is then the same as the mode, and the "1σ" HPD region is $\bar{x} \pm \sigma/\sqrt{N}$, the same as in the frequentist case.

5.2.3. *Comparison of Approaches.* Despite the mathematical identity of the Bayesian and frequentist solutions to this simple problem, the meaning of the results and their methods of derivation could hardly be more different.

First, the interpretations of the results are drastically different. To a Bayesian, $\bar{x}$ is the most plausible value of μ given the one set of data at hand, and there is a plausibility of 0.68 that μ is in the range $\bar{x} \pm \sigma/\sqrt{N}$. In contrast, the frequentist interpretation of the result is a statement about the long term performance of adopting the procedure of estimating μ with $\bar{x}$ and stating that the true value of μ is in the interval $\bar{x} \pm \sigma/\sqrt{N}$. Specifically, if one adopts this procedure, the average of the μ estimates after many observations will converge to the true value of μ, and the statement about the interval containing μ will be true 68% of the time. Note that this is not a statement about the plausibility of the single value of $\bar{x}$ or the single confidence region actually calculated. Frequency theory can only make statements about the long-term performance of the adopted procedure, not about the confidence one can place in the results of the procedure for the one available data set.

Mathematically, these conceptual differences are reflected in the choice of the interesting variable in the final gaussian distributions, equations (26) and (31). The frequentist approach estimates μ and finds the probability content of a confidence region by integrating over possible values of $\bar{x}$, thus taking into consideration hypothetical data sets with different sample means than that observed. The Bayesian calculation finds the estimate and the probability content of an HPD region by integrating over μ, that is, by considering different hypotheses about the unknown true value of μ. The symbolic expression of frequentist and Bayesian interval probabilities expresses this difference precisely: The frequentist calculates $p(\mu - \sigma/\sqrt{N} < \bar{x} < \mu + \sigma/\sqrt{N})$, the fraction of the time that the sample mean will be within $\sigma/\sqrt{N}$ of μ in many repetitions of the experiment. In contrast, the Bayesian calculates $p(\bar{x} - \sigma/\sqrt{N} < \mu < \bar{x} + \sigma/\sqrt{N} \mid DI)$, the probability that μ is within $\sigma/\sqrt{N}$ of the sample mean of the one data set at hand.

The second important difference between the frequentist and Bayesian calculations is the uniqueness and directness of the Bayesian approach. Frequentist theory could only produce a unique procedure by appealing to *ad hoc* criteria such as unbiasedness and shortest confidence intervals. Yet such criteria are not generally valid (Jaynes 1976; Zellner 1986). For example, there is a growing literature on biased estimators, because prior information or evidence in the sample can identify a procedure that is appropriate for the case in consideration, but that would not have optimal long term behavior (Efron 1975; Zellner 1986). In contrast, BPT provides a unique solution to any well posed problem, and this solution is guaranteed by our desiderata to be the best one possible given the information actually available, by rather inescapable criteria of rationality and consistency.

As a third important difference, we note that the frequentist calculation of the "covering probability" of the confidence region depended on special properties of the distribution for the statistic that was chosen. First, the statistic—the sample mean, $\bar{x}$—is what is called a "sufficient statistic." This means that the μ dependence of the probability of the data (*i.e.*, the likelihood, equation [29]) depends on the data only through the value of the single number $\bar{x}$, and not on any further information in the sample; a single number summarizes all of the information in the sample, regardless of the size of N. Second, the sampling probability of $\bar{x}$, equation (26), depends on μ and $\bar{x}$ only through their difference. These properties permitted the calculation of the coverage probability without requiring knowledge of the true value of μ. Unfortunately, not all distributions have sufficient statistics, and of those that do, few depend on the the sufficient statistics and the parameters only through their differences (Lindley 1958). In general, then, a frequentist confidence region can only be defined approximately. In contrast, a Bayesian can always calculate an HPD region exactly, regardless of the existence of sufficient statistics and without special requirements on the form of the sampling distribution.

As a final, fourth difference, we note that the Bayesian result that is identical to the frequentist result used a *least informative prior.* As soon as there is any cogent prior information about unknown parameter values, the Bayesian result will differ from frequentist results, since the latter have no natural means for incorporation of prior information.

In summary, Bayesian and frequentist results will only be mathematically identical if (1) there is only least informative prior information, (2) there are sufficient statistics, and (3) the sampling distribution depends on the sufficient statistic and the parameters only through their differences. Bayesian/frequentist equivalence is thus seen to be something of a coincidence (Jeffreys 1937). When these conditions are not met, Bayesian and frequentist results will generally differ (if a frequentist result exists!), and the Bayesian result will be demonstrably superior, incorporating prior information and evidence in the sample that is ignored in frequentist theory (Jaynes 1976).

5.2.4. *Improper Priors.* The Bayesian posterior becomes precisely identical to the frequentist sampling distribution when $[\mu_{\rm min}, \mu_{\rm max}] \rightarrow [-\infty, \infty]$. Interestingly, in this limit both the prior (28) and the marginal likelihood (30) vanish, but they do so in such a way that the ratio $p(\mu \mid I)/p(D \mid I)$ is nonzero. In fact, in this infinite limit, we can set the prior equal to any constant, say $p(\mu \mid I) = 1$, and we will get the same result. Such a prior is not normalized, and is therefore called *improper.* It is frequently true in estimation problems that use of improper priors gives the result that would be found by using a proper (normalizable) prior and taking the limit. Improper priors then become convenient expressions of prior ignorance of the range of a parameter. It is usually Bayesian results based on improper priors that are mathematically equivalent to frequentist results.

In some estimation problems, and more frequently in model comparison problems, allowing parameter ranges in least informative priors to become infinite leads to unnormalizable or vanishing posterior probabilities. This is a signal that prior information about the allowed ranges of parameters is important in the result. In principle, we will demand that all probabilities be proper. This is never a serious restriction, for we always know *something* about the allowed parameter range. For example, in measuring the length of an object in the laboratory with a caliper, we know it can't be larger than the earth, nor smaller than an atom. We can put these limits in our prior, and we will almost always find that the posterior is independent of them to many, many significant figures; the data "overwhelms" the information in the prior range. In these cases we might as well use an improper prior as a kind of shorthand. On the other hand, if the result depends sensitively on the prior range, BPT is telling us that the information in the data is not sufficient to "overwhelm" our prior information, and so we had better think carefully about just what we know about the prior range. Or alternatively, we could try to get better data!

5.2.5. *The Robustness of Estimation.* Not only does the information in the data usually overwhelm the prior range; it also often overwhelms the actual shape of the prior, even when it is informative. This is best illustrated by example.

Suppose in our gaussian problem that our prior information indicated that μ was likely to be within some scale δ about some value μ_0. This state of knowledge could be represented by a gaussian prior with mean μ_0 and standard deviation δ,

$$p(\mu \mid I) = \frac{1}{\delta\sqrt{2\pi}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\delta^2}\right]. \tag{32}$$

Repeating the posterior calculation above with this prior, we find that the posterior mean $\hat{\mu}$ and variance σ_μ^2 are now

$$\hat{\mu} = \frac{\bar{x} + \mu_0 \frac{\alpha}{N}}{1 + \frac{\alpha}{N}}, \tag{33}$$

and

$$\sigma_\mu^2 = \frac{\sigma^2}{N+\alpha}, \tag{34}$$

where $\alpha = \sigma/\delta$. Therefore, unless $\delta \lesssim \sigma/N$ (so that $\alpha \gtrsim N$), the posterior will not be significantly different from that calculated with a least informative prior.

This is an interesting result of some practical importance. The gaussian prior is clearly much more informative than the uniform prior, but unless the prior probability is very concentrated, with $s \leq \sigma/N$, it will have little affect on the posterior. This is not a very deep result; it is just what we should expect. It merely tells us that unless our prior information is as informative as the data, it will have little effect on our inferences. Of course, it is seldom the case that we have such prior information when we analyze an experiment; our lack of such information is why we perform experiments in the first place!

The practical import of this result is that if it is not clear exactly what prior probability assignment expresses our prior information, we might as well go ahead and use some simple "diffuse" prior that qualitatively respects any prior information we have (it should vanish outside the allowed parameter range!) and see if the result depends much on the prior. Usually it will not. This phenomenon has been variously referred to as the "stability" (Edwards *et al.* 1963) or "robustness" (Berger 1984, 1985) of estimation. Berger (1984, 1985) has extensively studied the robustness of many Bayesian calculations.

This is a special case of a more general practical rule: if a problem is not well posed, in the sense of there not being obvious ways of converting information to probability assignments, just do a calculation using some approximation (a diffuse prior, a simple likelihood, a simple hypothesis space) that does not do too much violence to the information at hand. Such simplified problems are often of great use by themselves (see Section 8.3 for an example), and their solution may provide the insight one needs to put enough structure on the original problem to make it well posed.

5.2.6. *Reference Priors.* A number of investigators have developed procedures for constructing diffuse priors for estimation problems in which we are in a least informative state of knowledge about parameter values, but do not know how to find the corresponding prior distribution. The robustness of estimation implies that the detailed shape of the prior is unimportant as long as it is diffuse compared to the likelihood function, so these procedures use properties of the likelihood function to "automatically" create a diffuse prior. Such a prior is often generically referred to as a "reference prior" (Box and Tiao 1973; Zellner 1977; Bernardo 1979): it is an "off-the-shelf" diffuse prior that many consider to be a useful objective starting point for analysis.

All such priors are based on the idea that one can think of the least informative state of knowledge pragmatically as the state of having little information *relative to what the experiment is expected to provide* (Rosenkrantz 1977). Unfortunately, several different procedures can been created to express this qualitative notion. Fortunately, many of them lead to the same reference prior for many common statistical problems, and these priors are often identical to least informative priors, when the latter are known.

Though several of the proposed reference priors are often identical to least informative priors in specific problems, this will not be true in general. In particular, since the form of a reference prior depends on the likelihood function, if we are estimating the same parameter in two different experiments, the reference prior will in general be different for the two experiments. This emphasizes that a reference prior does not describe an absolute state of ignorance about a parameter, but rather specifies a state of ignorance with respect to the experiment. To the extent that we choose experiments based on our prior information about the quantity we wish to measure, we expect the prior to depend on *some* properties of the likelihood function. After all, the I that appears in the prior is the same I that appears in the likelihood; the role the parameter plays in the likelihood is an important part of our prior information about the parameter (Jaynes 1968, 1980a). But the form of the likelihood can be determined in part by information that is irrelevant to the parameter value, information that would have no influence on a least informative prior, but that could affect a reference prior.

Despite these problems, reference priors can play a useful role in Bayesian parameter estimation because they produce diffuse priors that qualitatively express ignorance about parameters, and estimation is often robust with respect to the detailed form of a diffuse prior. Some of the reference priors that have been advocated include the invariant priors of Jeffreys and Huzurbazar (Jeffreys 1939); the indifferent conjugate priors of Novick and Hall (1965); the maximal data informative priors of Zellner (1971, 1977); the data translated likelihood priors of Box and Tiao (1973); and the reference priors of Bernardo (1979, 1980).

The multiparameter marginalization priors of Jaynes (1980a), where the priors for each of the parameters in a multiparameter model are chosen to ensure that they are uninformative about the other parameters, may also be considered to be reference priors, in that they are diffuse priors determined by the form of the likelihood.

5.3 MODEL COMPARISON

We can use this signal measurement example to illustrate some key features of Bayesian model comparison. Suppose there is some model, M_1, that gives a precise prediction of the signal: $\mu_{\rm true} = \mu_1$. Suppose further that an alternative model, M_2, specifies only that $\mu_{\rm true}$ is in some interval, $[\mu_{\rm min}, \mu_{\rm max}]$. Model M_2 has a single parameter, and model M_1 is a simple hypothesis, with no parameters.

Now suppose that we obtain some data, D, with a sample mean of $\bar{x}$. Which model is more plausible in light of this data? We can answer this with Bayes' Theorem, in the form of equation (20), or in the form of posterior odds, equation (22). To use it, we need the marginal likelihoods for M_1 and M_2. Since M_1 has no parameters, the marginal likelihood is just the likelihood itself;

$$p(D \mid I_1) = \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left[-\frac{\sum_i (x_i - \mu_1)^2}{2\sigma^2}\right]. \tag{35}$$

Model M_2 is the model assumed for the estimation problem we solved above; its marginal likelihood is given by equation (30). Together, these give the Bayes factor in favor of model M_1,

$$\begin{aligned} B_{12} &= \frac{p(D \mid I_1)}{p(D \mid I_2)} \\ &\approx \frac{\mu_{\rm max} - \mu_{\rm min}}{\sigma/\sqrt{N}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{N}{2\sigma^2}(\mu_1 - \bar{x})^2\right], \end{aligned} \tag{36}$$

where we have assumed that $\mu_{\rm max}$ and $\mu_{\rm min}$ are large compared to $\sigma/\sqrt{N}$, and are far enough away from $\bar{x}$ that the last factor in equation (30) is very nearly equal to $1/(\mu_{\rm max} - \mu_{\rm min})$. This assumption amounts to saying that the experiment has measured μ more accurately than M_2 predicted it.

This result is very interesting. If $\bar{x}$ happens to equal μ_1, B_{12} will be large, favoring model M_1 which predicts that the true mean is μ_1. But B_{12} will continue to favor M_1 even when $\bar{x}$ is somewhat different from μ_1, despite the fact that model M_2 with best-fit $\mu = \bar{x}$ fits the data slightly better than M_1. In effect, M_2 is being penalized for having a parameter and therefore being more complicated than M_1.

We can see this better if we note that the ratio of the best-fit likelihoods of the models, from equations (35) and (29), is

$$R_{12} = \exp\left[-\frac{N}{2\sigma^2}(\mu_1 - \bar{x})^2\right]. \tag{37}$$

Thus the Bayes factor can be written

$$\begin{aligned} B_{12} &= \frac{1}{\sqrt{2\pi}} \frac{\mu_{\rm max} - \mu_{\rm min}}{\sigma/\sqrt{N}} R_{12} \\ &\equiv S_{12} R_{12}. \end{aligned} \tag{38}$$

The best-fit likelihood ratio, R_{12}, can never favor model M_1; the more complicated model almost always fits the data better than a simpler model. But the factor S_{12} favors the simpler model; it is called the "simplicity factor" or the "Ockham factor", and is a quantification of the rule known as "Ockham's Razor": Prefer the simpler model unless the more complicated model gives a significantly better fit (Jeffreys 1939; Jaynes 1980b; Gull 1988; Bretthorst 1990).

We can understand how the penalty for complication arises by recalling that the Bayes' factor is the ratio of the prior predictive probabilities of the models. Thus BT compares models by comparing how well each predicted the observed data. Crudely speaking, a complicated model can explain anything; thus, its prior predictive probability for any particular outcome is small, because the predictive probability is spread out more or less evenly among the many possible outcomes. But a simpler model is more constrained and limited in its ability to explain or fit data. As a result, its predictive distribution is concentrated on a subset of the possible outcomes. If the observed outcome is among those expected by the simpler model, BT favors the simpler model because it has better predicted the data.

In this sense, BT is the proper quantitative expression of the notion behind P-values: Assess an hypothesis by how well it predicts the data. To do so, BT uses only the probability of the actually observed data; additionally, it takes into account all of the possible parameter values through marginalization. This is in stark contrast to frequentist GOF tests, which consider the probabilities of hypothetical data, and assume the truth of the best-fit parameter values.

Equation (36) has a sensitive dependence on the prior range of the additional parameter that at first seems disconcerting. But a little thought reveals it to be an asset of the theory, something we might have expected and wanted. For example, suppose the alternative to model M_2 was some model M_3 which was just like M_2, but had a smaller allowed range for μ. If the sample mean, $\bar{x}$, fell in a region of overlap between the models, the likelihood ratio R_{32} would be 1, but S_{32} would lead BT to favor M_3. If the value of $\bar{x}$ fell outside of the range for μ specified by M_3, BT might still favor M_3, depending on how far $\bar{x}$ is from the prediction of M_3. In this way, BT "knows" that M_3 is simpler or more constrained than M_2, even though both models are very similar, and in particular have the same number of parameters. Such behavior could not result if the Bayes factor somehow ignored the prior ranges of model parameters. A consequence of this dependence on the prior range is that model comparison problems are not as robust as estimation problems with regard to the prior range.

Here and in other problems we can deal with sensitivity to the prior by "turning Bayes' Theorem around" and asking how different kinds of prior information would affect the conclusions. For example, if we report the likelihood ratio, R_{12}, and the posterior variance for μ, $\sigma_\mu = \sigma/\sqrt{N}$, then we know that the prior range for μ in model M_2 would have to have been smaller than $\sigma_\mu(2\pi)^{1/2}/R_{12}$ for us to just favor the more complicated model.

This kind of analysis can give us some insight into the common practice of accepting a new parameter if its value is significant at the "2σ" level. Taking $\mid \bar{x} - \mu_1 \mid = 2\sigma_\mu$, then $R_{12} = e^{-2}$, and the prior range that would make the Bayes factor indifferent between the models (giving $B_{12} = 1$) has a size of $\sigma_\mu(2\pi)^{1/2}/R_{12} = 18.5\sigma_\mu$. Thus the common practice of accepting a parameter significant at about the 2σ level corresponds to an initial state of uncertainty regarding the parameter value that is about one to two orders of magnitude greater than the uncertainty after the experiment.

The simple example we have worked here is more sensitive to the prior range than most realistic model comparison problems. Good examples of realistic model comparison problems in the physical sciences are discussed by Bretthorst (1988b, 1989a,b,c,d). Many additional model comparison problems have been worked in the Bayesian literature under the name, "significance testing". Important references include Jeffreys (1939), Zellner (1980), and Bernardo (1980).

6. Case Study: Measuring a Weak Counting Signal

We need only generalize the gaussian measurement problem slightly to obtain a problem that is both astrophysically interesting and resistant to frequentist analysis. We will consider in this section the measurement of a signal in the presence of a background rate that has been independently measured. We will consider signals that are measured by counting particles (photons, neutrinos, cosmic rays), so that the Poisson distribution is the appropriate sampling distribution.

The usual approach to this problem is to obtain an estimate of the background rate, $\hat{b}$, and its and standard deviation, σ_b, by observing an empty part of the sky, and an estimate of the signal plus background rate, $\hat{r}$, and its standard deviation, σ_r, by observing the region where a signal is expected. The signal rate is then estimated by $\hat{s} = \hat{r} - \hat{b}$, with variance $\sigma_s^2 = \sigma_r^2 + \sigma_b^2$. This procedure is the correct one for analyzing data regarding a signal which can be either positive or negative, when the gaussian distribution is appropriate. Thus it works well when the background and signal rates are both large so that the Poisson distribution is well-approximated by a gaussian. But when the rates are small, the procedure fails. It can lead to negative estimates of the signal rate, and even when it produces a positive estimate, both the value of the estimate and the size of the confidence region are corrupted because the method can include negative values of the signal in a confidence region.

These problems are particularly acute in gamma-ray and ultra-high energy astrophysics, where it is the rule rather than the exception that few particles are counted, but where one would nevertheless like to know what these sparse data indicate about a possible source. Given the weaknesses of the usual method, it is hardly surprising that more sophisticated statistical analyses of reported detections conclude that "not all the sources which have been mentioned can be confidently considered to be present" (O'Mongain 1973) and that "extreme caution must be exercised in drawing astrophysical conclusions from reports of the detection of cosmic γ-ray lines" (Cherry *et al.* 1980).

Three frequentist alternatives to the above procedure have been proposed by gamma-ray astronomers (Hearn 1969; O'Mongain 1973; Cherry *et al.* 1980). They improve on the usual method by using the Poisson distribution rather than the gaussian distribution to describe the data. But they have further weaknesses. First, all three procedures interpret a likelihood ratio as the covering probability of a confidence region, and thus are not even accurate frequentist procedures. Second, none of the procedures correctly accounts for the uncertainty in the background rate. Hearn (1969) uses the best-fit estimate of the background in his calculation, correcting the result afterward by using the gaussian propagation of error rule. O'Mongain (1973) tries to find 'conservative' results by using as a background estimate the best-fit value plus one standard deviation. Cherry *et al.* (1980) try to more carefully account for the background uncertainty by a method similar to marginalization; but strangely they only include integral values of the product of the background rate and the observing time in their analysis.

There are several reasons for the difficulty in finding a unique, optimal frequentist solution to this problem. First, there is important prior information in this problem: neither the signal nor the background can be negative. Second, there is a nuisance parameter: we want to estimate the signal, but to do so we must also consider possible values of the background. Third, the appropriate distribution is not the gaussian distribution, and cannot be written as a function of the difference between sufficient statistics and the relevant parameters; thus frequentist methods for finding confidence regions and dealing with nuisance parameters in the gaussian case do not apply.

Bayesian probability theory can deal with all these complications straightforwardly. The Bayesian solution to this problem is as follows.

First, the background rate, b, is measured by counting n_b events in a time T from an "empty" part of the sky. If we were interested in the value of b, we could estimate it from these data by taking prior information I_b specifying the connection between b, n_b, and T; I_b will identify the Poisson distribution as the likelihood function (see Jaynes 1990a for an instructive Bayesian derivation of the Poisson distribution). The likelihood function is thus

$$p(n_b \mid bI_b) = \frac{(bT)^{n_b} e^{-bT}}{n_b!}. \tag{38}$$

The least informative prior for the rate of a Poisson distribution can be derived from a simple group invariance argument, noting that $1/b$ plays the role of a scale for measurement of time (Jaynes 1968). The result is

$$p(b \mid I_b) = \frac{1}{b}. \tag{39}$$

This is called the "Jeffreys prior", since it was first introduced in similar problems by Jeffreys (1939). It corresponds to a prior that is uniform in $\log b$, and expresses complete ignorance regarding the scale of the background rate. As written here, it is improper. We can bound b to make the prior proper, and take limits after calculating the posterior for b, but as long as n_b is not zero, the limit will exist and be the same as if we just used equation (39) throughout the calculation. Of course, the prior probability for negative values of b will be taken to be zero.

Given these probability distributions, the marginal likelihood is

$$\begin{aligned} p(n_b \mid I_b) &= \frac{T^{n_b}}{n_b!} \int_0^\infty db\, b^{n_b - 1} e^{-bT} \\ &= \frac{1}{n_b}. \end{aligned} \tag{40}$$

The posterior density for b is then,

$$p(b \mid n_b I_b) = \frac{T(bT)^{n_b - 1} e^{-bT}}{(n_b - 1)!}. \tag{41}$$

If we are interested in the background, we might summarize this posterior by noting its mean, $\langle b \rangle = n_b/T$, and its standard deviation, $n_b^{1/2}/T$, the usual "root N" result expected from a Poisson signal. With a prior that is different from equation (39), these values would be different, but not substantially so if n_b is reasonably large. For example, a uniform prior would give a mean value of $(n_b + 1)/T$ and a standard deviation of $\sqrt{n_b + 1}/T$.

Now we count n events in a time t from a part of the sky where there is a suspected source. This measurement provides us with information about both b and the source rate s. From BT, the joint posterior density for s and b is,

$$p(sb \mid nI) = p(sb \mid I)\frac{p(n \mid sbI)}{p(n \mid I)}$$
$$= p(s \mid bI)p(b \mid I)\frac{p(n \mid sbI)}{p(n \mid I)}. \tag{42}$$

Of course, the information I includes the information from the background measurement, as well as additional information I_s specifying the possible presence of a signal. Symbolically, $I = n_b I_b I_s$.

The likelihood is the Poisson distribution for a source with strength $s + b$:

$$p(n \mid sbI) = \frac{t^n(s+b)^n e^{-(s+b)t}}{n!}. \tag{43}$$

The prior for s, $p(s \mid bI)$, is the least informative prior for a Poisson rate $(s + b)$, with the value of b given,

$$p(s \mid bI) = \frac{1}{s+b}. \tag{44}$$

Again, we take the prior probability to be zero for negative values of s. The prior for b in this problem is *informative*, since we have the background data available. In fact, since I_s is irrelevant to b, the prior for b in this problem is just the posterior for b from the background estimation problem, and is given by equation (41). Ignoring the normalization for now, BT gives the dependence of the posterior on the parameters as

$$p(sb \mid nI) \propto (s+b)^{n-1} b^{n_b - 1} e^{-st} e^{-b(t+T)}. \tag{45}$$

Usually, we are only interested in the source strength. To find the posterior density for the source strength, *independent of the background*, we just marginalize with respect to b, calculating $p(s \mid nI) = \int db\ p(sb \mid nI)$. After expanding the binomial, $(s + b)^{n-1}$, the integral can be easily calculated. The resulting normalized posterior is,

$$p(s \mid nI) = \sum_{i=1}^{n} C_i \frac{t(st)^{i-1} e^{-st}}{(i-1)!}, \tag{46}$$

with

$$C_i \equiv \frac{(1+\frac{T}{t})^i \quad \frac{(n+n_b-i-1)!}{(n-i)!}}{\sum_{j=1}^{n}(1+\frac{T}{t})^j \quad \frac{(n+n_b-j-1)!}{(n-j)!}}. \tag{47}$$

Note that $\sum_{i=1}^{n} C_i = 1$.

This result is very appealing. Comparing it with equation (41), we see that BT estimates s by taking a weighted average of the posteriors one would obtain attributing 1, 2,. . ., n events to the signal. The weights depend on n, t, n_b, and T so that the emphasis is placed on a weak signal or a strong signal, depending on how n/t compares with n_b/T. Further development of this result, including application to real data, will appear elsewhere.

7. Case Study: Neutrinos from SN 1987A

The simple example of the previous section shows how straightforwardly Bayes' Theorem provides a solution to a well-posed problem that, despite its simplicity, has so far evaded straightforward frequentist analysis. Now we will discuss another problem that at first appears to be much more complicated, but which we will see is no more complicated in principle than the gaussian estimation problem discussed in Section 5.

In February of 1987, a supernova was observed in the Large Magellanic Cloud. This supernova, dubbed SN 1987A, was the closest one observed in the history of modern astronomy. Setting it apart from all other supernovae ever observed—indeed, from all other astrophysical sources ever observed, except for the Sun—is the fact that it was detected, not only in electromagnetic radiation, but also in neutrinos. Roughly two dozen neutrinos were detected from the supernova by the Japanese Kamiokande II (KII), Irvine-Michigan-Brookhaven (IMB), and Soviet Baksan detectors.

Neutrinos are believed to carry away about ninety-nine percent of the energy released by a supernova; the KII, IMB, and Baksan detections thus represent the first direct measurement of the energy of a supernova. In addition, neutrinos interact with matter so weakly that once they leave the collapsing stellar core, they pass unimpeded through the massive envelope of the supernova. Thus the detected neutrinos provide astrophysicists with their first glimpse of a collapsing stellar core. The analysis of the observations is therefore of great significance for testing supernova theory.

In addition, important information about intrinsic properties of the neutrino, such as its rest mass and electric charge, is contained in the data. This is because the 50 kpc path length between the Large Magellanic Cloud and Earth is vastly larger than that accessible in terrestrial laboratories.

Unfortunately, the weakness of neutrino interactions responsible for their usefulness as probes of stellar core dynamics also makes them extremely difficult to detect once they reach Earth. Of the approximately 10^{16} supernova neutrinos that passed through the detectors, only about two dozen were actually detected. Even these few events were not detected directly, but only by detecting tertiary photons they produced in the detectors. The small size of the data set, and the complicated relationship between properties of the incident neutrino signal and properties of the detected tertiary particles, demand careful, rigorous analysis of the implications of these data.

7.1 A BEWILDERING VARIETY OF FREQUENTIST ANALYSES

Within days after the landmark detection, the first contributions to what would soon become a vast literature analyzing the detected neutrinos appeared. Today, the two dozen supernova neutrinos are probably the most analyzed data set in the history of astrophysics, the number of published analyses far outnumbering the number of data. Unfortunately, nearly all of these analyses have *ad hoc* methodological elements, due to their frequentist inspiration.

With the exception of several qualitative moment analyses, most investigators analyzed the data by comparing them with parametrized models for the neutrino signal. With so few data, only the simplest signal models can be justified. But despite the simplicity of the models, the complexity of the detection process greatly complicates any frequentist analysis of the data, because the sampling distribution is extremely complex even for simple models.

No obvious sufficient statistics exist, and it would be difficult, if not impossible, to analyze the frequency behavior of statistics to identify unbiased, efficient estimators. A consequence of the lack of sufficient statistics is that frequentist confidence regions for parameters can only be found approximately.

All the usual frequentist criteria therefore founder on this problem, and investigators have been forced to rely on their intuitions and their Monte Carlo codes to create and calibrate statistics for their analyses. It is no wonder, therefore, that a bewildering variety of statistics and statistical methodologies has been applied to these data, yielding a similarly bewildering variety of results (see Loredo and Lamb 1989 for a review). Though many investigators used the maximum likelihood method to find best-fit parameters—a method with a Bayesian justification— several employed Pearson's method of moments, or invented their own statistics. A wide variety of methods were invented to calculate "confidence regions" for parameters, most of them confusing GOF P-values with covering probabilities. The majority of these methods relied on one-dimensional or two-dimensional Kolmogorov-Smirnov (KS) statistics, or similar goodness-of-fit statistics based on the cumulative distribution for the events, rather than the likelihood, even when the likelihood was used to find best-fit parameter values. Finally, very few studies considered more than one model for the neutrino emission. Usually, the adequacy of a single model was assumed without question; in some cases, adequacy was justified with an "alternative-free" goodness of fit test. A few studies explored several models, attempting to compare them with maximum likelihood ratios, but more complicated models always had larger likelihoods.

Testimony to the robustness of this problem, the results of many of these studies agree, if not precisely, at least qualitatively. But there is still troubling variety in the conclusions reached. For example, some investigators conclude that the observations are in conflict with soft equations of state for neutron star matter, though most conclude that the data are consistent with all reasonable equations of state, soft or hard. Some investigators claim the data indicate a small, nonzero electron antineutrino mass of a few eV, while most claim that the data only indicate an upper limit on the mass in the 15 to 20 eV range. The wide variety of statistical methods used in these investigations, and the variety in the models assumed for the neutrino emission and detection processes, make the literature on the supernova neutrinos appear muddled and confused. In the context of frequentist theory, there is no compelling criterion for making a judgement about the relative soundness of one analysis compared to another. Some scientists, in an attempt to summarize the analyses, have been forced to do "statistical statistics", averaging the results of different studies.

The majority of these studies were not even good frequentist analyses. In particular, many investigators identified "95% confidence regions" with the range of parameter values that had goodness-of-fit P-values of greater than 5%, based on a flawed definition of a confidence region. These investigators did not notice that their best-fit P-values of ≈ 0.80 implied that "confidence regions" with probability smaller than about 20% *could not even be defined* with their methods. But this is almost beside the point. The emphasis of frequentist statistics on averages over hypothetical random experiments, and the lack of a clear rationale for the choice of statistics, has led to a "Monte Carlo Mystique" in astronomical statistics whereby almost any calculation relying on a sufficient number of simulated data sets is deemed a "rigorous" statistical analysis.

Alone among these analyses is the work of Kolb, Stebbins, and Turner (KST, 1987). They focus on one interesting parameter—the mass of the electron antineutrino, $m_{\bar{\nu}_e}$—and setting aside all of the fancy statistics and Monte Carlo codes, ask instead what careful intuitive reasoning about the data can reveal about $m_{\bar{\nu}_e}$. They conclude that at best, the data can put an upper limit on $m_{\bar{\nu}_e}$ of the order of 25 to 30 eV, not significantly better than current laboratory limits. Later detailed statistical studies found "95% confidence" limits ranging from 5 eV to 19 eV. Significantly, some recent reviews of the observations downplay these later studies and emphasize the qualitative KST limit, testimony to the lack of confidence scientists have in the statistical methods of astrophysicists.

7.2 THE BAYESIAN ANALYSIS

The Bayesian analysis of the neutrino data has been presented by Loredo and Lamb (1989; 1990a,b). They estimate parameters for simple neutrino emission models using Bayes' Theorem with uniform priors. This calculation is as straightforward in principle as the gaussian calculation of Section 5; the only complications are computational, arising from the complexity of the detector response and the dimensionality of the parameter spaces.

The data produced by the detectors are the detected energies, $\epsilon_i^{\rm det}$, and arrival times, $t_i^{\rm det}$, of the detected neutrinos. To analyze these data, Loredo and Lamb consider a variety of parametrized models for the neutrino emission, and use Bayes' Theorem to estimate the model parameters and to compare alternative models. Given a model for the neutrino emission rate, a predicted detection rate per unit time and unit energy, $d^2N_{\rm det}/d\epsilon^{\rm det}dt^{\rm det}$, can be calculated using the response function of the experiment. From this detection rate, the likelihood function needed in Bayes' Theorem can be constructed as follows.

The expected number of neutrinos detected in a small time interval, Δt, and a small energy interval, $\Delta\epsilon$, is just the detection rate times $\Delta t\Delta\epsilon$. From the Poisson distribution, the probability that no neutrinos will be detected within these intervals about a specified energy and time is

$$P_0(\epsilon^{\rm det},t^{\rm det}) = \exp\left[-\frac{d^2N_{\rm det}(\epsilon^{\rm det},t^{\rm det})}{d\epsilon^{\rm det}dt^{\rm det}}\Delta\epsilon\Delta t\right]. \tag{48}$$

Similarly, the probability that a single neutrino will be detected in the interval is

$$P_1(\epsilon^{\rm det},t^{\rm det}) = \frac{d^2N_{\rm det}(\epsilon^{\rm det},t^{\rm det})}{d\epsilon^{\rm det}dt^{\rm det}}\Delta\epsilon\Delta t\exp\left[-\frac{d^2N_{\rm det}(\epsilon^{\rm det},t^{\rm det})}{d\epsilon^{\rm det}dt^{\rm det}}\Delta\epsilon\Delta t\right]. \tag{49}$$

The intervals are chosen small enough that the probability of detecting more than one neutrino is negligible compared to P_0 and P_1.

The likelihood of a particular observation is the product of the probabilities of detection of each of the $N_{\rm obs}$ observed neutrinos, times the product over all intervals not containing a neutrino of the probability of no detection. That is,

$$\mathcal{L} = \left[\prod_{i=1}^{N_{\rm obs}} P_1(\epsilon_i^{\rm det},t_i^{\rm det})\right]\prod_j P_0(\epsilon_j^{\rm det},t_j^{\rm det}), \tag{50}$$

where j runs over all intervals not containing an event. It is more convenient to work with the log likelihood, $L=\ln(\mathcal{L})$. From the definitions of P_0 and P_1 it follows that

$$L = \sum_{i=1}^{N_{\rm obs}}\ln\left[\frac{d^2N_{\rm det}(\epsilon_i^{\rm det},t_i^{\rm det})}{d\epsilon^{\rm det}dt^{\rm det}}\Delta\epsilon\Delta t\right] - \sum_j\frac{d^2N_{\rm det}(\epsilon_j^{\rm det},t_j^{\rm det})}{d\epsilon^{\rm det}dt^{\rm det}}\Delta\epsilon\Delta t, \tag{51}$$

where j now runs over all intervals. In the limit of small $\Delta\epsilon$ and Δt, the second term becomes the integral of the rate function over all time and all energy. Thus the log likelihood is

$$\begin{aligned} L &= \sum_{i=1}^{N_{\rm obs}} \ln\left[\frac{d^2 N_{\rm det}(\epsilon_i^{\rm det}, t_i^{\rm det})}{d\epsilon^{\rm det}\, dt^{\rm det}}\right] - \int_0^{t_{\rm dur}} dt \int_0^{\infty} d\epsilon^{\rm det}\, \frac{d^2 N_{\rm det}(\epsilon^{\rm det}, t^{\rm det})}{d\epsilon^{\rm det}\, dt^{\rm det}} \\ &= \sum_{i=1}^{N_{\rm obs}} \ln\left[\frac{d^2 N_{\rm det}(\epsilon_i^{\rm det}, t_i^{\rm det})}{d\epsilon^{\rm det}\, dt^{\rm det}}\right] - N_{\rm det}, \end{aligned} \tag{52}$$

where $t_{\rm dur}$ is the duration of the time interval under study, and $N_{\rm det}$ is the total number of events expected to be detected in that interval. In this equation, the intervals $\Delta\epsilon$ and Δt have been omitted because they are constants that do not affect the functional dependence of L on the detected rate function.

Equation (52) is the final form for the likelihood function. Combined with prior probability densities for the parameters (Loredo and Lamb [1989] assume uniform priors), it yields a posterior distribution for the model parameters. The calculation, though straightforward in principle, is complicated in practice because the response functions of the detectors are complicated. This is because the neutrinos are not detected directly; rather, tertiary photons produced in the detectors by the neutrinos are detected, leading to a complicated relationship between detected photon energy and the energy of the incident neutrino. As a result, calculation and study of the posterior distribution requires the resources of a supercomputer. Details are presented in Loredo and Lamb (1989, 1990a,b).

These calculations show that the observations are in spectacular agreement with the salient features of the theory of stellar collapse and neutron star formation which had developed over several decades in the absence of direct observational data. In particular, the inferred radius and binding energy of the neutron star formed by the supernova are in excellent agreement with model calculations based on a wide range of equations of state, despite earlier indications to the contrary.

These calculations also show that the upper limit on the mass of the electron antineutrino implied by the observations is 25 eV at the 95% confidence level, 1.5 to 5 times higher than found previously, and not significantly better than current laboratory limits.

This work demonstrates the value of using correct and rigorous Bayesian methods for the analysis of astrophysical data, and shows that such an analysis is not only possible, but straightforward, even when the data are related to the physical quantities of interest in a very complicated manner.

8. Where to Go from Here

Bayesian probability theory, as described here, is impressive in its simplicity and its scope. Desiderata of appealing simplicity lead to its rules for assignment and manipulation of probabilities, which are themselves extremely simple. Its identification of probability with plausibility makes it a theory of drastically broader scope than traditional frequentist statistics. This broad scope adds to the simplicity and unity of the theory, for whenever we wish to make a judgement of the truth or falsity of *any* proposition, A, the correct procedure is to calculate the probability, $p(A \mid E)$, that A is true, conditional on all the evidence, E, available, regardless of whether A refers to what would traditionally be called a random variable or a more general hypothesis (Jaynes 1990b). In most cases, this calculation will involve the use of Bayes' Theorem.

Because of its broad scope, BPT is more than merely a theory of statistics. It is a theory of *inference*, a generalization of deductive inference to cases where the truth of a proposition is uncertain because the available information is incomplete. As such, it deserves to be a familiar element of every scientist's collection of general methods and tools.

Of course, the theory is ideally suited for application to problems traditionally classified as "statistical". There, it promises to simplify and unify statistical practice. Indeed, it is already doing so in the fields of mathematical statistics, econometrics, and medicine. Astrophysicists have been slow to reap the benefits of the theory, but several applications relevant to astrophysics have been worked out. We will describe some here, as an entrance to the expanding literature on Bayesian methods.

8.1 ASTROPHYSICAL ESTIMATION AND MODEL COMPARISON PROBLEMS

Because of the prevalence of the gaussian distribution in statistical problems, many frequentist parameter estimation calculations will be equivalent to their Bayesian counterparts, provided that there are no nuisance parameters and that there is no important prior information about parameter values. But when there are nuisance parameters, or when there is important prior information, Bayesian methods should prove superior to frequentist methods, if the latter even exist for such problems. Also, if the relevant distributions are more complicated than gaussian, lacking obvious sufficient statistics, Bayesian methods will almost certainly prove superior to frequentist methods, and will be easier to derive.

Problems for which Bayesian methods will provide demonstrable advantages are only beginning to be be identified and studied. All such problems are approached in a unified manner using Bayes' Theorem, eliminating any nuisance parameters through marginalization. The signal measurement and supernova neutrino problems mentioned above are examples.

Another example is the analysis of "blurred" images of point sources in an attempt to resolve closely spaced objects (Jaynes 1988; Bretthorst and Smith 1989). In this problem, some of the parameters specifying the locations of objects are nuisance parameters, since it is their *relative* positions that are of interest. Further, the noise level is not always known; in the Bayesian calculation it, too, can be a nuisance parameter to be eliminated by marginalization, effectively letting Bayes' Theorem estimate the noise from the data. Finally, the brightnesses of the two or more possible objects can be marginalized away, leaving a probability density that is a function only of relative position between objects, and which answers the question, "Is there evidence in the data for an object at this position relative to another object?" In analyzing an image for the presence of two objects, the Bayesian procedure can thus reduce the dimensionality of the problem from seven (two two-dimensional positions, two brightnesses, and the noise level) to one (the relative separation of the objects). Of course, once the relative separation posterior is studied and found to reveal the presence of closely spaced objects, their intensities and positions can be calculated, using knowledge of their relative separation to simplify analysis of the full posterior.

Analytical work (Jaynes 1988) and numerical work analyzing simulated data (Bretthorst and Smith 1989) indicate that the Bayesian algorithm can easily resolve objects at separations of less than one pixel, depending on the signal-to-noise ratio of the data. Further, model comparison methods can be used to determine the number of point sources for which there is significant evidence in the data. Significantly, the calculation also reveals that the usual practice of apodizing an optical system to smooth out the sidelobes of the point spread function destroys significant information that the Bayesian calculation can use to resolve objects (Jaynes 1988). Apodizing leads to a smoother image that is less confusing to the eye, but it destroys much of the information in the sidelobes that probability theory can use to improve resolution. This work awaits application to real data, and extension to other similar problems, such as the analysis of data from optical interferometers.

8.2 BAYESIAN SPECTRUM ANALYSIS

One class of statistical problems is of such great importance in astrophysics that it deserves special consideration: the analysis of astrophysical time series data for evidence of periodic signals. This problem is usually referred to as *spectrum analysis.* In the past three years, new Bayesian spectrum analysis methods have been developed that offer order-of-magnitude greater frequency resolution than current methods based on the discrete Fourier transform (DFT). Additionally, they can be used to detect periodicity in amplitude modulated signals or more complicated signals with much greater sensitivity than DFT methods, without requiring the data to be evenly spaced in time.

Current frequentist methods seek information about the spectrum of the *signal* by calculating the spectrum of the *data* via the discrete Fourier transform (DFT). But the presence of noise and the finite length of the data sample make the data spectrum a poor estimate of the signal spectrum. As a result, *ad hoc* methods are used to "correct" the data spectrum, involving various degrees of smoothing (to eliminate spurious peaks). The statistical properties of the result are analyzed assuming the signal is just noise, to try to find the "false alarm" probability of an apparent spectral feature being due to noise. (Good reviews of these methods are in Press, *et al.* 1986, and van der Klis 1989.)

In contrast, Bayesian methods (Jaynes 1987; Bretthorst 1989, 1990) assess the significance of a possible signal by directly calculating the probabilities that the data are due to a periodic signal or to noise, and comparing them. To estimate the frequency of a signal, these methods simply calculate the probability of a signal as a function of its frequency, marginalizing away the phase and amplitude of the signal.

Using these methods, Jaynes (1987) derived the DFT as the appropriate statistic to use when analyzing a signal with a single sinusoid present. His work shows how to manipulate the DFT without smoothing to get an optimal frequency estimate that can have orders-of-magnitude greater resolution than current methods. Bretthorst (1989, 1990) has extended Jaynes' work, showing analytically and with simulated and actual data that the DFT is *not* appropriate for the analysis of signals with more complicated structure than a single sinusoid, and that Bayesian methods give much more reliable and informative results. In particular, Bayesian methods can easily resolve two frequencies that are so close together that there is only a single peak in the DFT of the data, simply by considering a model with more than one sinusoid present. As was the case in the analysis of blurred images just discussed, probability theory uses information in the sidelobes to improve resolution, information that is thrown away by the standard Blackman-Tukey smoothing methods. Model comparison calculations can be used to identify how many sinusoids there is evidence for in the data. Bretthorst (1988a,b) has applied these methods to Wolf's sunspot data, comparing the results of the Bayesian analysis with conventional DFT results.

For signals that are not stationary, such as chirped or damped signals, the DFT spreads the signal power over a range of frequencies. However, if the general form of the signal is known, Bayesian generalizations of the DFT can be constructed that take into account the possibility that the signal has some unknown chirp or decay rate, effectively concentrating all of the signal power into a single frequency, thereby greatly improving detection sensitivity for such signals. These methods should prove to be of immense value for the study of nonstationary astrophysical time series, such as those observed from the "quasi-periodic oscillator" x-ray sources, or those expected from sources of gravitational radiation. In particular, the gravitational radiation signal expected from coalescing binaries is chirped, so the "chirpogram" introduced by Jaynes (1987) and further studied by Bretthorst (1988a,b) should play an important role in the analysis of gravitational wave signals. An integrated circuit is currently being developed to facilitate rapid calculation of the chirpogram (Erickson, Neudorfer, and Smith 1989).

8.3 INVERSE PROBLEMS

Problems that are mathematically ill-posed in the sense of being underdetermined arise frequently in astrophysics; they are usually called *inverse problems.* Examples include calculating the interior structure of the sun from helioseismology data, calculating radio images from interferometric data, "deblurring" optical or x-ray images, or estimating a spectrum from proportional counter or scintillator data. Abstractly, all of these problems have the following form. Some unknown signal, s, produces data, d, according to

$$d = Rs + e, \tag{53}$$

where R is a complicated operator we will call the response function of the experiment, and e represents an error or noise term. Given d, R, and some incomplete information about e, we wish to estimate s. Such problems can be ill-posed in three senses.

First, the response operator is usually singular in the sense that a unique inverse operator, R^{-1}, does not exist. Put another way, there exists a class, A, of signals such that $Rs = 0$ for any s in A. Thus d contains no information about such signals, so that even the noiseless "pure inverse problem" of solving $d = Rs$ for s does not have a unique solution: any element of A can be added to any solution to give another solution. The set A is called the *annihilator* of R. It exists because the "blurring" action of R destroys information about finely structured signals.

Second, the presence of noise effectively enlarges the annihilator of R, since signals s such that $Rs = \epsilon$, with ϵ small compared to the expected noise level, can be added to any possible solution to obtain another acceptable solution. In practice, this is revealed by instability in any attempt to directly invert equation (48), small changes in the data resulting in large changes in the estimated signal.

Finally, the data, d, are usually discrete and finite in number, and the signal, $s = s(x)$, is usually continuous. Thus, even if R were not singular and there were no noise, estimating $s(x)$ from d would still be severely underdetermined.

One approach to such ill-posed problems is to make them well-posed by studying simple parameterized models for the signal. The resulting estimation problem can be addressed straightforwardly with Bayes' Theorem. But often, one would like "model-independent" information about the signal, $s(x)$.

Frequentist approaches to this problem fall into two classes. *Regularization methods* estimate the signal by invoking criteria to select one member of the set of all possible signals that are consistent with the data as being "best" in some sense. *Resolution methods* try to determine what features all the feasible signals have in common by estimating resolvable averages of them. All such methods have obvious *ad hoc* elements—the choice of regularizer, or the choice of a measure of resolution—and there are usually many methods available for solving a particular problem. In recent years, the importance of using prior information to guide development of an inverse method has been greatly emphasized (Frieden 1975; Narayan and Nityanada 1986). Unfortunately, it is not clear how to optimally use even the simplest prior information, such as the positivity of the signal, to develop a frequentist inverse method.

The Bayesian approach to inverse problems is to *always* address them as estimation problems via Bayes' Theorem. They differ from other more common estimation problems only in the character of the model assumed. In particular, the model will usually have more parameters than there are data. Prior information, taken into account through prior probabilities, is what makes such problems well-posed despite the discrepancy between the number of data and the number of parameters.

Bayesian solutions to inverse problems are only beginning to be developed and understood. Only the simplest kinds of models and prior information have yet been explored. Surprisingly, the resulting methods are usually as good as any existing frequentist methods, and are sometimes significantly better. These methods are the *Maximum Entropy Methods* prominent in these Proceedings, though the "entropy" which plays such an important role in these methods is *not* the entropy described in Section 3, above.

Bayesian inversion methods, including the popular maximum entropy methods, can be developed as follows (Jaynes 1984a,b). Consider estimating a one-dimensional signal, $s(x)$. Begin by discretizing the problem, seeking to estimate the finite number of values $s_j \equiv s(x_j)$, $j = 1$ to M; M may be much larger than the number of data. The "parameters" of our model are thus just the M values of the discrete signal. Using Bayes' theorem, we can calculate the posterior probability of a signal, given the data, the response function, and information about the noise:

$$p(\{s_j\} \mid DI) = p(\{s_j\} \mid I)\frac{p(D \mid \{s_j\}I)}{p(D \mid I)}. \tag{54}$$

The likelihood function will be determined by our information about the noise; if the information leads to a gaussian noise distribution, the log likelihood will just be proportional to χ^2. The critical element of the problem is the assignment of prior probabilities to the s_j. Uniform priors clearly will not do, for then all of the possible signals that fit the data will be equally likely, and the problem will remain underdetermined. Intuitively, we reject many of the possible signals—for example, wildly oscillating signals—because our prior information about the nature of the true signal makes it extremely unlikely that it could have been one of the many unappealing but possible signals. We must find a way to encode some of this information numerically in a prior probability assignment over the s_j.

The natural way to proceed is to specify precisely the available information, I, and use the principles discussed in Section 3.3 to assign the prior, $p(\{s_j\} \mid I)$. The information will probably be of the form of a specification of the nature of the alternatives, I_0, and some additional testable information, E. The information I_0 will lead to a least informative distribution, $p(\{s_j\} \mid I_0)$. For example, if the signal $s(x)$ must by nature be positive, the LIP distribution for $\{s_j\}$ might be a product of Jeffreys priors, $p(\{s_j\} \mid I_0) = \prod 1/s_j$. The testable information, E, could include, for example, information about the expected scale of detail in the signal, in the form of prior covariances among the s_j. This information would be used to identify the appropriate informative prior for the signal by MAXENT. The entropy of the distribution $p(\{s_j\})$ needed to use MAXENT is calculated by integrating over the values of the s_j variables,

$$H[p(\{s_j\})] = -\int ds_1 \ldots \int ds_M \; p(\{s_j\}) \log \left[\frac{p(\{s_j\})}{m(\{s_j\})} \right], \tag{55}$$

where $m(\{s_j\})$ is the LIP assignment for $\{s_j\}$. The informative distribution is the one with highest entropy, $H[p(\{s_j\})]$, among all those that satisfy the constraints imposed by E, and could be found (at least in principle) by the method of Lagrange multipliers.

For historical reasons, this is not the approach that has been taken in assigning a prior for the signal, though it is a promising direction for future research. Instead, a prior has been constructed by choosing an alternative space of hypotheses than the s_j, from which the s_j values can be derived, but whose nature permits an unambiguous and appealing prior probability assignment.

The well-known maximum entropy inversion methods arise from a particularly simple alternative hypothesis space created as follows (Gull and Daniel 1978; Jaynes 1982, 1984a,b; Skilling 1986). First, discretize the M signal values into some large number, N, of independent "signal elements" of size δs.* Then build a signal by taking the N signal elements one at a time and putting them in one of the M signal bins. A signal is built once each of the N elements have been placed into a bin; we will call such a signal a "microsignal". The new hypothesis space is the set of the M^N possible resulting microsignals, and as a least informative assignment, we will consider each of them to be equally probable, with probability M^{-N}. If we label each of the signal elements with an index, δs_i, then we can describe each microsignal by a set of M lists of the indices corresponding to the elements in each of the M bins. For example, for a two bin signal built from five signal elements, a particular microsignal could be described by the set $\{(2,3),(1,4,5)\}$.

Of course, the model leading to the microsignal hypothesis space is not the only model one could imagine for constructing a signal; further, it is not clear exactly what information about the signal is being assumed by this model. Nevertheless, the resulting prior for $\{s_j\}$ has some intuitively pleasing properties, and leads to inversion methods that have proved extremely useful for the analysis of complicated data.

The least informative distribution for microsignals implies a prior probability distribution for the "macrosignals" specified by the M numbers, s_j, as follows. In terms of the basic signal element, we can write $s_1 = n_1 \delta s$, $s_2 = n_2 \delta s$, and so on, with $\sum_j s_j = N \delta s$. An element of the original hypothesis space can thus be specified by a set of integers, n_j. Now the key is to note that, in general, each of the possible macrosignals—each of the possible set of n_j values—will correspond to *many* possible microsignals. For example, a macrosignal with $n_1 = 2$ signal elements in bin 1 is equally well described by microsignals with signal elements (1, 2) in bin 1, or (1, 3) in bin 1, or (1437, 3275) in bin 1.

Denote the number of microsignals that correspond to a given macrosignal by the *multiplicity* $W(\{n_j\})$ of the macrosignal. The prior probability we will assign to each macrosignal is just its multiplicity times the probability, M^{-N}, of each of its constituent microsignals;

* These elements are not to be identified with any physical "quantum" in the problem; for example, they should not be identified with photons detected by an experiment. They should reflect our prior information about the interesting scale of variation in the *signal*, not the data.

$p(\{n_j\} \mid I) = W(\{n_j\})M^{-N}$. The multiplicity of a macrosignal is given by the multinomial coefficient,

$$W(\{n_j\}) = \frac{N!}{n_1!n_2!\dots n_M!}. \tag{56}$$

Using Stirling's formula, the log of the multiplicity is well approximated by

$$\begin{aligned}\log W(\{n_j\}) &\approx N\log N - \sum_{j=1}^{M} n_j \log n_j \\ &= N\left[-\sum_{j=1}^{M}\frac{n_j}{N}\log\frac{n_j}{N}\right] \\ &= NH(\{n_j\}),\end{aligned} \tag{57}$$

where we have defined the *combinatorial entropy of the signal*, $H(\{n_j\})$, as

$$H(\{n_j\}) \equiv -\sum_{j=1}^{M}\frac{n_j}{N}\log\frac{n_j}{N}. \tag{58}$$

In terms of the entropy, the prior probability of a macrosignal can now be written,

$$p(\{n_j\} \mid I) = M^{-N}e^{NH(\{n_j\})}. \tag{59}$$

This prior has some intuitively appealing properties. In particular, it favors smoothly varying signals in the following sense. A priori, the most probable signal using this particular signal model is the signal with maximum combinatorial entropy; a simple calculation shows that the completely uniform signal, with all n_j equal, has maximum entropy. Similarly, a signal with all N signal elements in one bin—the "least uniform" signal—is a priori the least probable; it has a multiplicity of one. When combined with a likelihood function, this prior assignment will thus tend to favor the most uniform of all those signals consistent with the data.

To use the entropic prior (59), the values of M and N must be specified. Their values should express prior information we have about the signal and the experiment's ability to measure it. M will be related to the resolution we expect is achievable from our data. N might be related to how well the data can resolve differences in the signal level; it therefore seems reasonable that the choice of N should be tied to the noise level. Finding ways to convert prior information into choices for M and N is a current research problem (see, *e.g.*, Jaynes 1985b, 1986b; Gull 1989). Fortunately, the results of inversion with entropic priors do not depend sensitively on these numbers.

Despite the simplicity of the information leading to entropic inversion, it has proved enormously successful for analyzing a wide variety of astrophysical data. Some impressive recent examples include the calculation of radio images from interferometric data (Skilling and Gull 1985); imaging accretion discs from emission line profiles (Marsh and Horne 1989); estimating distances to clusters of galaxies from angular positions and apparent diameters of galaxies (Lahav and Gull 1989); and deconvolution of x-ray images of the Galactic center region (Kawai *et al.* 1988). An extensive bibliography of earlier applications of entropic inversion in astronomy is available in Narayan and Nityanada (1986), and in the physical sciences in general in Smith, Inguva, and Morgan (1984).

Entropic inverses like that described here were first introduced in astrophysics by Gull and Daniel (1978), based on earlier work by Frieden (1972) and Ables (1974). In these works, entropic inverses are presented as regularization methods, that is, as methods for producing a single "best" estimate of the signal from the data. Most later work has emphasized this regularization interpretation of the combinatorial entropy of an image (see Narayan and Nityanada 1986 for a review). In this context, entropic inverses are referred to as "maximum entropy methods", since they focus attention on what we would here identify as the most probable (maximum entropy) signal. Only recently has the Bayesian interpretation of these methods been clarified (Jaynes 1984b, 1985b, 1986b; Gull 1989; Skilling 1986, 1989, 1990). As valuable as the regularization interpretation may be, the Bayesian interpretation should prove even more valuable, for the following reasons.

First, as a regularization method, it is not clear why maximum entropy methods should be preferred to other regularization methods. Many have argued that entropy should be preferred as a regularizer by making analogies between the combinatorial entropy of a signal and the entropy of a probability distribution. As we have shown above, a probability distribution with maximum entropy consistent with the available information is the uniquely correct distribution to choose to represent that information. The mathematical similarity of equations (12) and (58) has led some to claim the same status for a signal with maximum combinatorial entropy. But since a signal is not a probability distribution, the arguments identifying the entropy of a distribution as the uniquely correct measure of its information content do not apply to signals. (See Skilling 1989 for a different viewpoint.)

Second, when entropy is viewed as a regularizer and not a prior probability, the manner in which it should be used to address an inverse problem is not clear. It should be combined with some statistical measure of the goodness-of-fit of a signal to the data, but the choice of statistic and the relative weighting of the entropy factor and the goodness-of-fit is arbitrary in frequentist regularization theory. Thus entropy has been combined, not only with the likelihood of the signal, as dictated in the Bayesian approach, but also with other goodness-of-fit statistics, such as the Kolmogorov-Smirnov statistic, adding a new element of arbitrariness and subjectivity to the results. Further, the connection of the parameter N with prior information is lost the regularization approach, where it plays the role of a relative weighting between entropy and goodness-of-fit. No compelling criteria for the specification of the value of such a "regularization parameter" have yet been introduced in regularization theory.

Third, as a regularization method, entropic inverses can provide only a single "best" signal. When viewed as Bayesian methods, however, they can not only produce a "best" (most probable) signal, but can also provide measures of the statistical significance of features in the inverted signal. This aspect of Bayesian entropic inverses is an important element of the "Quantified Maximum Entropy" approach described by Skilling (1990) and Sibisi (1990) in these proceedings.

Finally, the Bayesian interpretation of entropic inverses reveals their dependence on prior information and a specific model for the signal, indicating ways they may be improved for specific problems. For example, though maximum entropy methods impressively reconstruct signals with point sources against a weak background, it is well known that they often poorly reconstruct signals that have a strong smoothly varying component, producing spurious features (Narayan and Nityanada 1986). To deal with such situations, several *ad hoc* modifications have been advanced (see, *e.g.*, Frieden and Wells 1978; Narayan and Nityanada 1986; Burrows and Koornneef 1989). Yet from a Bayesian perspective, it is apparent that such poor behavior is simply the result of the minimal amount of information assumed in calculating entropic inverses. The microsignal model assumes little more than the positivity of a signal; in particular, it ignores possible correlations between values of the signal in adjacent bins. Incorporation of such information should improve restorations; initial studies by Gull (1989a) reveal the promise of such an approach.

Entropic inverses are only one particularly simple example of a Bayesian inverse method. Others can be created, either by incorporating additional information into the prior (59) through MAXENT, by considering some hypothesis space other than that of the microsignal model that leads to the entropic inverse (Jaynes 1984b, 1986b), or especially by using MAXENT to find the prior for the s_j directly (using the entropy of the *distribution*, equation [55], not that of the signal). Further research into Bayesian inversion should yield methods superior to entropic inversion in particular problems, though the simplicity of the entropic inverse will no doubt recommend it as a useful "jackknife" method, useful in the preliminary analysis of a wide variety of problems.

8.4 JAYNESIAN PROBABILITY THEORY

Bayesian methods are playing an increasingly important role in many areas of science where statistical inference is important. They have had a particularly powerful impact in mathematical statistics and econometrics, and there is much a physical scientist can learn from the statistical and econometric Bayesian literature. Particularly rich sources of information are the books by Tribus (1969), Zellner (1971), Box and Tiao (1973), and Berger (1985), and the influential review article of Edwards *et al.* (1963). Many important references to the literature are available in the reviews of Lindley (1972), Zellner (1989), and Press (1989).

But with the exception of the much neglected work of Jeffreys (1939), Bayesian methods have had little impact in the physical sciences until very recently. This has been due in large part to the lack of compelling rationale for the assignment of prior probabilities. The majority of the Bayesian literature (including most of the references just mentioned) regards prior probabilities as purely subjective expressions of a person's opinions about hypotheses, allowing individuals in possession of the same information to assign different probabilities to propositions. With this subjective element, Bayesian probability theory was viewed as being of little value to physical science.

Virtually alone among statisticians, Jaynes has emphasized that an *objective* probability theory can be developed by requiring that probability assignments satisfy the desideratum that we have here called *Jaynes Consistency:* Equivalent states of knowledge should be represented by equivalent probability assignments. This principle is the key to finding objective solutions to the problem of assigning direct probabilities—both prior probabilities and sampling probabilities—which is fully half of probability theory. The resulting theory remains subjective in the sense that probabilities represent states of knowledge, and not properties of nature. But the theory is objective in the sense of being completely independent of personalities or opinions. It is this objective aspect that makes the *Jaynesian Probability Theory* outlined here the appropriate tool for dealing with uncertainty in astrophysics, and indeed in all sciences.

9. Acknowledgements

It is a great pleasure to thank Don Lamb, Larry Bretthorst, and Ed Jaynes for many very valuable discussions. This work was supported in part by NASA grants NGT-50189, NAGW-830, and NAGW-1284.

10. References

Ables, J.G. (1974) 'Maximum Entropy Spectral Analysis', *Astron. Astrophys. Supp.* **15**, 383.

Bayes, T. (1763) 'An Essay Towards Solving a Problem in the Doctrine of Chances', *Phil. Trans. Roy. Soc. London* **53**, 370. Reprinted in *Biometrika* **45**, 293, and in Press (1989).

Berger, J.O. (1984) 'The Robust Bayesian Viewpoint', in J.B. Kadane (ed.), *Robustness of Bayesian Analyses*, Elsevier Science Publishers, B.V., p. 63.

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

Berger, J.O., and D. A. Berry (1988) 'Statistical Analysis and the Illusion of Objectivity', *Amer. Scientist* **76**, 159.

Berger, J.O., and R. Wolpert (1984) *The Likelihood Principle*, Institute of Mathematical Statistics, Hayward, CA.

Bernardo, J.M. (1979) 'Reference Posterior Distributions for Bayesian Inference', *J. Roy. Stat. Soc.* **B41**, 113.

Bernardo, J.M. (1980) 'A Bayesian Analysis of Hypothesis Testing', in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.), *Bayesian Statistics*, University Press, Valencia, Spain, p. 605.

Bevington, P.R. (1969) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill Book Company, New York.

Birnbaum, A. (1962) *J. Amer. Statist. Assoc.* 'On the Foundations of Statistical Inference', **57**, 269; and following discussion.

Box, G.E.P., and G.C. Tiao (1973) *Bayesian Inference in Statistical Analysis*, Addison-Wesley Publishing Co., Reading, MA.

Bretthorst, G.L. (1988a) 'Excerpts from Bayesian Spectrum Analysis and Parameter Estimation', in G.J. Erickson and C.R. Smith (eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1*, Kluwer Academic Publishers, Dordrecht, p. 75.

Bretthorst, G.L. (1988b) *Bayesian Spectrum Analysis and Parameter Estimation*, Springer-Verlag, New York.

Bretthorst, G.L. (1989a) 'Bayesian Model Selection: Examples Relevant to NMR', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, p. 377.

Bretthorst, G.L. (1989b) 'Bayesian Analysis I: Parameter Estimation Using Quadrature NMR Models', *J. Magn. Reson.*, in press.

Bretthorst, G.L. (1989c) 'Bayesian Analysis II: Signal Detection and Model Selection', *J. Magn. Reson.*, in press.

Bretthorst, G.L. (1989d) 'Bayesian Analysis III: Applications to NMR Signal Detection, Model Selection and Parameter Estimation', *J. Magn. Reson.*, in press.

Bretthorst, G.L. (1990) 'An Introduction to Parameter Estimation Using Bayesian Probability Theory', these proceedings.

Bretthorst, G.L., and C.R. Smith (1989) 'Bayesian Analysis of Signals from Closely-Spaced Objects', in R.L. Caswell (ed.), *Infrared Systems and Components III*, Proc. SPIE 1050.

Burrows, C., and J. Koornneef (1989) 'The Application of Maximum Entropy Techniques to Chopped Astronomical Infrared Data', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht.

Cherry, M.L., E.L. Chupp, P.P. Dunphy, D.J. Forrest, and J.M. Ryan (1980) 'Statistical Evaluation of Gamma-Ray Line Observations', *Ap. J.* **242**, 1257.

Cox, R.T. (1946) 'Probability, Frequency, and Reasonable Expectation', *Am. J. Phys.* **14**, 1.

Cox, R.T. (1961) *The Algebra of Probable Inference*, Johns Hopkins Press, Baltimore.

Dawid, A.P. (1980) 'A Bayesian Look at Nuisance Parameters', in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.), *Bayesian Statistics*, University Press, Valencia, Spain, p. 167.

Eadie, W.T., D. Drijard, F.E. James, M. Roos, and B. Sadoulet (1971) *Statistical Methods in Experimental Physics*, North-Holland Publishing Company, Amsterdam.

Edwards, W., H. Lindman, and L.J. Savage (1963) 'Bayesian Statistical Inference for Phychological Research', *Psych. Rev.* **70**, 193; reprinted in J.B. Kadane (ed.), *Robustness of Bayesian Analyses*, Elsevier Science Publishers, B.V., p. 1.

Efron, B. (1975) 'Biased Versus Unbiased Estimation', *Adv. Math.* **16**, 259.

Erickson, G.J., P.O. Neudorfer, and C.R. Smith (1989) 'From Chirp to Chip, A Beginning', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrect, p. 505.

Feigelson, E.D. (1989) 'Statistics in Astronomy', in S. Kotz and N.L. Johnson (eds.), *Encyclopedia of Statistical Science, Vol. 9*, in press.

Fougere, P.F. (1988) 'Maximum Entropy Calculations on a Discrete Probability Space', in G.J. Erickson and C.R. Smith (eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1*, Kluwer Academic Publishers, Dordrecht, p. 205.

Fougere, P.F. (1989) 'Maximum Entropy Calculations on a Discrete Probability Space: Predictions Confirmed', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrect, p. 303.

Frieden, B.R. (1972) 'Restoring with Maximum Likelihood and Maximum Entropy', *J. Opt. Soc. Am.* **62**, 511.

Frieden, B.R. (1972) 'Image Enhancement and Restoration', in T.S. Huang (ed.), *Picture Processing and Digital Filtering*, Springer-Verlag, New York, p. 177.

Frieden, B.R., and D.C. Wells (1978) 'Restoring with Maximum Entropy. III. Poisson Sources and Backgrounds', *J. Opt. Soc. Am.* **68**, 93.

Good, I.J. (1980) 'The Contributions of Jeffreys to Bayesian Statistics', in A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics*, North-Holland, Amsterdam, p. 21.

Grandy, W.T. (1987) *Foundations of Statistical Mechanics Vol. 1: Equillibrium Theory*, D. Reidel Publishing Company, Dordrecht.

Gull, S.F. (1988) 'Bayesian Inductive Inference and Maximum Entropy', in G.J. Erickson and C.R. Smith (eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1*, Kluwer Academic Publishers, Dordrecht, p. 53.

Gull, S.F. (1989) 'Developments in Maximum Entropy Data Analysis', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, p.

53.
Gull, S.F., and G.J. Daniell (1978) 'Image Reconstruction from Incomplete and Noisy Data', *Nature* **272**, 686.
Hearn, D. (1969) 'Consistent Analysis of Gamma-Ray Astronomy Experiments', *Nuc. Inst. and Meth.* **70**, 200.
Iverson, G.R. (1984) *Bayesian Statistical Inference*, Sage Publications, Beverly Hills, California.
Jaynes, E.T. (1957a) 'Information Theory and Statistical Mechanics', *Phys. Rev.* **106**, 620.*
Jaynes, E.T. (1957b) 'How Does the Brain Do Plausible Reasoning?', Stanford Univ. Microwave Laboratory Report No. 421, reprinted in G.J. Erickson and C.R. Smith (eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1* (1988), Kluwer Academic Publishers, Dordrect, p. 1.
Jaynes, E.T. (1958) *Probability Theory in Science and Engineering*, Colloquium Lectures in Pure and Applied Science No. 4, Socony Mobil Oil Co. Field Research Laboratory, Dallas.
Jaynes, E.T. (1963) 'New Engineering Applications of Information Theory', in J.L. Bogdanoff and F. Kozin (eds.), *Proc. of the 1st Symp. on Engineering Applications of Random Function Theory and Probability*, John Wiley and Sons, Inc., New York, p. 163.
Jaynes, E.T. (1968) 'Prior Probabilities', *IEEE Trans.* **SSC-4**, 227.*
Jaynes, E.T. (1973) 'The Well-Posed Problem', *Found. of Phys.* **3**, 477.*
Jaynes, E.T. (1976) 'Confidence Intervals vs. Bayesian Intervals', in W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, D. Reidel Pub. Co., Dordrecht, p. 252.*
Jaynes, E.T. (1978) 'Where Do We Stand on Maximum Entropy', in R.D. Levine and M. Tribus (eds.), *The Maximum Entropy Formalism*, MIT Press, Cambridge, p. 15.*
Jaynes, E.T. (1980a) 'Marginalization and Prior Probabilities', in A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics*, North-Holland, Amsterdam, p. 43.*
Jaynes, E.T. (1980b) 'Review of *Inference, Method, and Decision* (R.D. Rosenkrantz)', *J. Am. Stat. Assoc.* **74**, 740.
Jaynes, E.T. (1982) 'On the Rationale of Maximum Entropy Methods', *Proc. IEEE* **70**, 939.
Jaynes, E.T. (1983) *Papers on Probability, Statistics, and Statistical Physics* (ed. R.D. Rosenkrantz), D. Reidel Pub. Co., Dordrecht.
Jaynes, E.T. (1984a) 'The Intuitive Inadequacy of Classical Statistics', *Epistemologia* **VII**, 43.
Jaynes, E.T. (1984b) 'Prior Information and Ambiguity in Inverse Problems', *SIAM-AMS Proc.* **14**, 151.
Jaynes, E.T. (1985a) 'Some Random Observations', *Synthese* **63**, 115.
Jaynes, E.T. (1985b) 'Where Do We Go From Here?', in C.R. Smith and W.T. Grandy, Jr. (eds.), *Maximum-Entropy and Bayesian Methods in Inverse Problems*, D. Reidel Publishing Company, Dordrecht, p. 21.
Jaynes, E.T. (1985c) 'Highly Informative Priors', in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.), *Bayesian Statistics 2*, Elsevier Science Publishers, Amsterdam, p. 329.

* Reprinted in Jaynes (1983).

Jaynes, E.T. (1986a) 'Bayesian Methods: General Background', in J.H. Justice (ed.), *Maximum-Entropy and Bayesian Methods in Applied Statistics*, Cambridge University Press, Cambridge, p. 1.

Jaynes, E.T. (1986b) 'Monkees, Kangaroos, and N', in J.H. Justice (ed.), *Maximum-Entropy and Bayesian Methods in Applied Statistics*, Cambridge University Press, Cambridge, p. 26.

Jaynes, E.T. (1987) 'Bayesian Spectrum and Chirp Analysis', in C.R. Smith and G.J. Erickson (eds.), *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems*, D. Reidel Publishing Company, Dordrecht, p. 1.

Jaynes, E.T. (1988a) 'The Relation of Bayesian and Maximum Entropy Methods', in G.J. Erickson and C.R. Smith (eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1*, Kluwer Academic Publishers, Dordrecht, p. 25.

Jaynes, E.T. (1988b) 'Detection of Extra-Solar System Planets', in G.J. Erickson and C.R. Smith (eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1*, Kluwer Academic Publishers, Dordrecht, p. 147.

Jaynes, E.T. (1989a) 'Clearing Up Mysteries — The Original Goal', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht.

Jaynes, E.T. (1989b) 'Probability in Quantum Theory', in *Proceedings of the Workshop on Complexity, Entropy, and the Physics of Information*, in press.

Jaynes, E.T. (1990a) 'Probability Theory as Logic', these proceedings.

Jaynes, E.T. (1990b) *Probability Theory – The Logic of Science*, in preparation.

Jeffreys, H. (1937) 'On the Relation Between Direct and Inverse Methods in Statistics', *Proc. Roy. Soc.* **A160**, 325.

Jeffreys, H. (1939) *Theory of Probability*, Oxford University Press, Oxford (3d revised edition 1961).

Kawai, N., E.E. Fenimore, J. Middleditch, R.G. Cruddace, G.G. Fritz, and W.A. Snyder (1988) 'X-Ray Observations of the Galactic Center by Spartan 1', *Ap. J.* **330**, 130.

Kolb, E. W., A. J. Stebbins, and M. S. Turner (1987) 'How Reliable are Neutrino Mass Measurements from SN 1987A?', *Phys. Rev.* **D35**, 3598; **D36**, 3820.

Lahav, O., and S.F. Gull (1989) 'Distances to Clusters of Galaxies by Maximum Entropy Method', *M.N.R.A.S.***240**, 753.

Lampton, M., B. Margon, and S. Bowyer (1976) 'Parameter Estimation in X-Ray Astronomy', *Ap. J.* **208**, 177.

Laplace, P.S. (1812) *Theorie Analytique des Probabilités*, Courcier, Paris.

Laplace, P.S. (1951) *Philosophical Essay on Probability*, Dover Publications, New York (originally published as the introduction to Laplace [1812]).

Lindley, D.V. (1958) 'Fiducial Distributions and Bayes' Theorem', *J. Roy. Stat. Soc.* **B20**, 102.

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

Lindley, D.V. (1972) *Bayesian Statistics, A Review*, Society for Industrial and Applied Mathematics, Philadelphia.

Loredo, T.J. and D.Q. Lamb (1989) 'Neutrinos from SN 1987A: Implications for Cooling of the Nascent Neutron Star and the Mass of the Electron Antineutrino', in E. Fenyves (ed.), *Proceedings of the Fourteenth Texas Symposium on Relativistic Astrophysics*, *Ann. N. Y. Acad. Sci.* **571**, 601.

Loredo, T.J. and D.Q. Lamb (1990a) 'Neutrinos from SN 1987A: Implications for Cooling of the Nascent Neutron Star', submitted to *Phys. Rev. D.*

Loredo, T.J. and D.Q. Lamb (1990b) 'Neutrinos from SN 1987A: Implications for the Mass of the Electron Antineutrino', submitted to *Phys. Rev. D.*

Mardia, K.V. (1972) *Statistics of Directional Data*, Academic Press, London.

Marsh, T.R., and K. Horne (1989) 'Maximum Entropy Tomography of Accretion Discs from their Emission Lines', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, p. 339.

Martin, B.R. (1971) *Statistics for Physicists*, Academic Press, London.

Mendenhall, W., R. L. Scheaffer, and D. D. Wackerly (1981) *Mathematical Statistics with Applications*, Duxbury Press, Boston.

Narayan, R., and R. Nityanada (1986) 'Maximum Entropy Image Restoration in Astronomy', *Ann. Rev. Astron. Astrophys.*, **24**, 127.

Novick, M., and W. Hall (1965) 'A Bayesian Indifference Procedure', *J. Am. Stat. Assoc.* **60**, 1104.

O'Mongain, E. (1973) 'Appplication of Statistics to Results in Gamma Ray Astronomy', *Nature* **241**, 376.

Press, S.J. (1989) *Bayesian Statistics: Principles, Models, and Applications*, John Wiley and Sons, New York.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling (1986) 'Numerical Recipes', Cambridge University Press, Cambridge.

Rényi, A. (1972) *Letters on Probability*, Wayne State University Press, Detroit.

Rosenkrantz, R.D. (1977) *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*, D. Reidel Publishing Company, Dordrect.

Runcorn, K. (1989) 'Sir Harold Jeffreys (1891-1989)', *Nature* **339**, 102.

Shore, J.E., and R.W. Johnson (1980) 'Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy', *IEEE Trans. Inf. Th.* **IT-26**, 26; erratum in **IT-29**, 942.

Sibisi, S. (1990) 'Quantified MAXENT: An NMR Application', these proceedings.

Skilling, J. (1986) 'Theory of Maximum Entropy Image Reconstruction', in J.H. Justice (ed.), *Maximum Entropy and Bayesian Methods in Applied Statistics*, Cambridge University Press, Cambridge, p. 156.

Skilling, J. (1989) 'Classic Maximum Entropy', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, p. 45.

Skilling, J. (1990) 'Quantified Maximum Entropy', these proceedings.

Skilling, J. and S.F. Gull (1985) 'Algorithms and Applications', in C.R. Smith and W.T. Grandy, Jr. (eds.), *Maximum-Entropy and Bayesian Methods in Inverse Problems*, D. Reidel Publishing Company, Dordrecht, p. 83.

Smith, C.R., and G. Erickson (1989) 'From Rationality and Consistency to Bayesian Probability', in J. Skilling (ed.), *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, p. 29.

Smith, C.R., R. Inguva, and R.L. Morgan (1984) 'Maximum-Entropy Inverses in Physics', *SIAM-AMS Proc.* **14**, 151.

Tribus, M. (1962) 'The Use of the Maximum Entropy Estimate in the Estimation of Reliability', in R.E. Machol and P. Gray (eds.), *Recent Developments in Information and Decision Processes*, The Macmillan Company, New York, p. 102.

Tribus, M. (1969) *Rational Descriptions, Decisions and Designs*, Pergamon Press, New York.

Van Campenhout, J.M., and T.M. Cover (1981) 'Maximum Entropy and Conditional Probability', *IEEE Trans. on Info. Theory* **IT-27**, 483.

van der Klis, M. (1989) 'Fourier Techniques in X-Ray Timing', in H. Ögelman and E.P.J. van den Heuvel (eds.), *Timing Neutron Stars*, Kluwer Academic Publishers, Dordrect, p. 27.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, J. Wiley and Sons, New York.

Zellner, A. (1977) 'Maximal Data Informative Prior Distributions', in A. Aykac and C. Brumat (eds.), *New Developments in the Application of Bayesian Methods*, North-Holland Publishing Co., Amsterdam, p. 211; reprinted in A. Zellner (1984) *Basic Issues in Econometrics*, University of Chicago Press, Chicago, p. 201.

Zellner, A. (1986) 'Biased Predictors, Rationality, and the Evaluation of Forecasts', *Econ. Let.* **21**, 45.

Zellner, A. (1988) 'A Bayesian Era', in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.), *Bayesian Statistics 3*, Oxford University Press, Oxford, p. 509.

Zellner, A., and A. Siow (1980) 'Posterior Odds Ratios for Selected Regression Hypotheses', in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.), *Bayesian Statistics*, University Press, Valencia, Spain, p. 585.

INDEX

Bayes' theorem, 86, 101, 103, 107, 132.
Bayes, T., 85, 87.
Bayesian inference, 81, 94.
Bernoulli's theorem, 85, 103.
Bernoulli, J., 85.
Boolean algebra, 96.
Bretthorst, G.L., 130.

chi-squared statistic, 108, 109.
classical statistics, 84.
conditionality principle, 110.
consistency, 96.
Cox, R.T., 81, 94, 95.

decision theory, 106.
direct probabilities, 99.

entropy, 102, 133.
estimation, 81, 104, 104, 114, 129.

Fourier transform, 130.
frequency theory, 84, 88, 89, 102.
frequentist statistics, 84.

gaussian distribution, 113.
global likelihood, 87, 105.
goodness-of-fit test, 109.

improper prior, 117, 123.
informative probabilities, 99, 101.
inverse problem, 131.

Jaynes consistency, 96.
Jaynes, E.T., 81, 94, 95, 130, 136.
Jaynesian probability theory, 136.
Jeffreys prior, 123, 132.
Jeffreys, H., 83, 94, 109, 136.

Laplace, P.S., 81, 85, 87.
law of large numbers, 85.
least informative probabilities, 99.
likelihood principle, 110.
likelihood, 87.

marginal likelihood, 105.
marginalization, 105, 107, 124.
MAXENT, 101, 103, 113, 132, 133.
model comparison, 81, 104, 107, 120, 129.

neutrinos, 81, 125.
nuisance parameter, 107, 123, 129.

Ockham factor, 121.
Ockham's razor, 121.
optional stopping, 93, 109.
orthodox statistics, 84.

parametrized models, 104.
Poisson distribution, 122.
posterior predictive distribution, 112.
posterior probability, 87.
principle of indifference, 88.
principle of insufficient reason, 88.
prior predictive distribution, 105.
prior probability, 87, 88.
probability axioms, 86, 96.
probability, definition of, 84.
product rule, 86, 97.

random variable, 89, 91, 110.
randomness, 91.
reference prior, 119.
regularization method, 131.
resolution method, 132.
robustness, 118.

sampling distribution, 87.
significance test, 104.
simplicity, 121.
spectrum analysis, 130.
sufficient statistic, 110, 117, 129.
sum rule, 86, 98.
supernova SN 1987A, 81, 125.

weak signal, 81, 122.

ERRATA

1. In the two paragraphs following equation (34) on page 118, make the following corrections:
   - Replace $\alpha = \sigma/\delta$ with $\alpha = \sigma^2/\delta^2$.
   - Replace $\delta \lesssim \sigma/N$ with $\delta \lesssim \sigma/\sqrt{N}$.
   - Replace $s \leq \sigma/N$ with $s \leq \sigma/\sqrt{N}$.
2. Jaynes (1980b) appeared in 1979, not 1980; the remainder of the reference is correct. The missing reference to Jaynes (1985d) is: Jaynes, E.T. (1985d) 'Macroscopic Prediction', in H. Haken (ed.), *Complex Systems – Operational Approaches*, Springer-Verlag, Berlin, p. 254.
3. Bretthorst (1989b, c, d) have since appeared in 1990 in *J. Mag. Res.*, **88**, on pages 533, 552, and 571, respectively.