An abridged version, without sections 5.3 and 6, and with other sections shortened, is in *Statistical Challenges in Modern Astronomy*, E.D. Feigelson and G.J. Babu (eds.), Springer-Verlag, New York, pp. 275–297 (1992), with discussion.

THE PROMISE OF BAYESIAN INFERENCE FOR ASTROPHYSICS

T.J. Loredo
Department of Astronomy
Space Sciences Building
Cornell University
Ithaca, New York 14853-0355 USA

ABSTRACT. The 'frequentist' approach to statistics, currently dominating statistical practice in astrophysics, is compared to the historically older Bayesian approach, which is now growing in popularity in other scientific disciplines, and which provides unique, optimal solutions to well-posed problems. The two approaches address the same questions with very different calculations, but in simple cases often give the same final results, confusing the issue of whether one is superior to the other. Here frequentist and Bayesian methods are applied to problems where such a mathematical coincidence does not occur, allowing assessment of their relative merits based on their *performance*, rather than on philosophical argument. Emphasis is placed on a key distinction between the two approaches: Bayesian methods, based on comparisons among alternative hypotheses using the single observed data set, consider averages over hypotheses; frequentist methods, in contrast, average over hypothetical alternative data samples, and consider hypothesis averaging to be irrelevant. Simple problems are presented that magnify the consequences of this distinction to where common sense can confidently judge between the methods. These demonstrate the irrelevance of sample averaging, and the necessity of hypothesis averaging, revealing frequentist methods to be fundamentally flawed. To illustrate how Bayesian and frequentist methods differ in more complicated, astrophysically relevant problems, Bayesian methods for problems using the Poisson distribution are described, including the analysis of 'on/off' measurements of a weak source in a strong background, and the analysis of time series resulting from recording event arrival times. Weaknesses of the presently used frequentist methods for these problems are straightforwardly overcome using Bayesian methods. Additional existing applications of Bayesian inference to astrophysical problems are noted.

CONTENTS

---

1. INTRODUCTION

Physics has an enormous aesthetic appeal. A great part of this appeal is the unity that physics brings to the wild variety of nature. Nowhere is this unity more apparent than in astrophysics, where physical laws are applied with astonishing success to situations that could hardly be more different than those of the earthbound laboratories in which they were first deduced and studied.

But astrophysics is an observational, not an experimental science. The phenomena under study are nearly always inaccessible to direct manipulation, and must be observed from great distances, and often indirectly. As a consequence, the scientific inferences of astrophysicists are fraught with uncertainty.

To realize the scientific potential of astrophysics thus demands an understanding, not only of the necessary physics, but also of the principles of inference that dictate how information can be optimally extracted from observational data and how theoretical predictions can be rigorously compared with such data. Indeed, for this very reason astronomers have made many of the most important early contributions to probability theory and statistics (Feigelson 1989). But in recent decades, there has been little interaction between astrophysicists and statisticians, and the majority of modern astrophysicists have little expertise in the use of sophisticated statistical methods.

In particular, few astrophysicists are aware that there is a controversy in statistics over the meaning of the most fundamental notion of the theory: probability. The traditional view, which we will call the *frequentist* view, identifies probability with the relative frequency of occurrence of an outcome in an infinite number of 'identical' repetitions of an experiment, or throughout an infinite ensemble. The older, *Bayesian* view, first clearly enunciated by

Laplace in his analyses of statistical problems in celestial mechanics, holds that probability is a measure of plausibility of a proposition, the degree to which it is supported by specified information. Though the frequentist viewpoint has been the orthodox viewpoint throughout this century, in recent decades there has been a great revival of interest in the Bayesian approach to probable inference.

Those physical scientists who are aware of the Bayesian/frequentist controversy almost universally share the opinion that these approaches are just different ways of interpreting the same calculations, and dismiss the controversy as being merely 'philosophical.' Or, aware that many Bayesian calculations require prior probabilities, they believe that Bayesian methods can differ significantly from their frequentist counterparts only when strong prior information is available. Neither of these beliefs is true.

Astronomers are not alone in having misconceptions about Bayesian inference. Many practicing statisticians also misunderstand the nature and extent of the distinction between the Bayesian and frequentist approaches. In particular, Lindley (1990) has noted that many frequentist statisticians fail to recognize Bayesian inference as 'a separate paradigm, distinct from their own, and merely... think of it as another branch of statistics, like linear models.'

On the contrary, Bayesian and frequentist methods are *fundamentally and profoundly different*. They address the same problem with different calculations, and can reach substantially different conclusions even in the absence of strong prior information. By a mathematical coincidence, the results of Bayesian and frequentist calculations for some of the most simple and most common problems are mathematically identical. However, they differ, not only in their interpretations, but in their derivations, and such an identity will not hold in general.

The key contrast between Bayesian and frequentist methods is not the use of prior information, but rather the choice of alternatives that are relevant for inference: Bayesian inference focuses on alternative hypotheses, frequentist statistics focuses on alternative data. To asses an hypothesis, H_1, Bayesian methods compare the probability of H_1 with the probabilities of other hypotheses; frequentist methods assume H_1 is true, and compare the probability of the observed data, D, with the probabilities of other data sets predicted by H_1.

A simple example will caricature this distinction. Let $p(A \mid B)$ denote the probability that some proposition, A, is true, given the truth of proposition B. If, given the truth of H_1, there are two possible data sets with probabilities $p(D_1 \mid H_1) = 0.001$ and $p(D_2 \mid H_1) = 0.999$, frequentist statistics considers observation of D_1 to be evidence against H_1 because it is so much less probable than the unobserved datum, D_2. Bayesian statistics insists that this information alone tells us nothing about H_1, since observation of D_1 is a possible consequence of the truth of H_1. Observation of D_1 can only be considered to be evidence against H_1 if there is a plausible alternative hypothesis, H_2, for which $p(D_1 \mid H_2)$ is sufficiently greater than 0.001. The priors for the two hypotheses determine precisely how much greater 'sufficiently greater' is. Priors are required to make the comparisons necessary in Bayesian inference, but they are not its essential distinguishing feature. It is the nature of the alternatives compared that distinguishes the two approaches.

I argue here that a return to the Bayesian approach promises to greatly improve the accessibility, precision, and power of statistical inference in astrophysics.

The case for Bayesian inference can be made in two complementary ways, emphasizing either its conceptual or pragmatic superiority to frequentist statistics. The compelling conceptual and mathematical foundations of Bayesian inference have been reviewed for astrophysicists in Loredo (1990), and are discussed more fully in the references cited there and in more recent works (Howson and Urbach 1989; Lindley 1990). The present paper seeks to demonstrate the superiority of Bayesian methods with more pragmatic criteria: Bayesian methods simply *perform* better in actual applications.

We will begin with a description of Bayesian and frequentist methods in Section 2 indicating how conceptual differences associated with the definition of probability lead to fundamental methodological differences in the procedures the two approaches use to address the same problem. We will emphasize the key contrast between the approaches just noted: the choice of alternatives relevant for inference. We will note that this contrast leads the two approaches to consider different kinds of averages in calculations: Bayesian calculations consider averages over hypotheses; frequentist calculations consider averages over hypothetical data (Efron 1978).

Sections 3 and 4 then present simple 'toy' problems that highlight this contrast and demonstrate that it has serious practical consequences. Section 3 will present two simple problems demonstrating that inferences based on the consideration of hypothetical data can be seriously misleading. The first problem, based on the χ^2 goodness-of-fit test, illustrates how ambiguity in the specification of data that might have been seen but were not can cause frequentist inferences to depend in a troubling way on phenomena that are irrelevant to the hypothesis under consideration. The second problem, employing common parameter estimation methods, demonstrates that the frequentist focus on good long-term behavior (averaged over many hypothetical data sets) can cause frequentist procedures to behave very poorly—even nonsensically—in individual cases. Good single-case behavior can be sacrificed for good long-term behavior.

Section 4 presents a simple problem demonstrating the necessity of explicitly considering alternative hypotheses in inference. Specifically, we address inference in the presence of *nuisance parameters*, the descriptive technical name for parameters that are a necessary element of a model for a phenomenon, but whose values are not of interest to the investigator. In an influential paper, Lampton, Margon, and Bowyer (1976) describe a frequentist method for eliminating uninteresting parameters in astrophysical data analysis; their 'projected χ^2' procedure has recently been advocated by Press, *et al.* (1986). But a simple example reveals this procedure to be seriously flawed, and traces this flaw back to the inability of frequentist methods to consider averages over alternative hypotheses.

Section 5 presents simple and astrophysically useful applications of Bayesian inference to problems involving the Poisson distribution that illustrate some of the ideas presented in Sections 3 and 4. Specifically, we discuss the measurement of a weak counting signal in the presence of a (possibly strong) background, and the analysis of event arrival times for periodicity. The results of Bayesian period detection calculations do not depend on the number of periods searched, eliminating a troubling subjective aspect of frequentist period detection procedures that has been the cause of much controversy among astronomers.

Section 6 briefly addresses the important problem of assigning direct probabilities for Bayesian inference, particularly the assignment of prior probabilities. We defer discussion of priors to this late Section in order to emphasize more fundamental points of contrast between Bayesian and frequentist methods addressed in earlier Sections.

Concluding remarks, with references for all astrophysical applications of Bayesian methods I know of, are presented in Section 7.

2. SUMMARY OF BAYESIAN AND FREQUENTIST APPROACHES: INFERRING A GAUSSIAN MEAN

Bayesians and frequentists agree on the rules for manipulating probabilities. However, their disagreement over the meaning of probability leads them to calculate the probabilities of different things in order to address the same question. Essentially, they disagree on what the arguments of probability symbols should be.

In this Section, we summarize the differences between the two approaches, illustrating each approach by applying them to a familiar statistical problem: inferring the mean of a Gaussian distribution.

2.1 Bayesian Inference

For a Bayesian, the probability $p(A \mid B)$ is a real-number-valued measure of the degree to which proposition A is supported by the information specified by proposition B. It is a numerical description of what B tells us about the truth of A, a measure of the extent to which B distinguishes between A and the alternatives to A. Any proposition is considered 'fair game' as the argument of a probability symbol. Of course, this is not to imply that the value of $p(A \mid B)$ is well-defined for every possible A and B. Indeed, one of the most important areas of research in Bayesian probability theory is the determination of the kinds of propositions for which there are well-defined probability assignments. As a bare minimum, B must specify the alternatives to A. After all, the truth of A can be uncertain only if there are alternatives that may be true in its place. Any assessment of the plausibility of A will depend on those alternatives and the extent to which any additional information specified by B distinguishes between them.

All allowed manipulations of Bayesian probabilities can be built from two basic rules. Writing $\overline{A}$ for "*not* A" (a proposition that is true if A is false), and AB for "A *and* B", these rules are the familiar sum rule,

$$p(A \mid C) + p(\overline{A} \mid C) = 1, \tag{2.1}$$

and the product rule,

$$\begin{aligned} p(AB \mid C) &= p(A \mid BC)\; p(B \mid C) \\ &= p(B \mid AC)\; p(A \mid C). \end{aligned} \tag{2.2}$$

As an abstract but important example of the Bayesian application of the sum and product rules, consider propositions H_1, H_2, and so on, asserting the truth of particular hypotheses, and a proposition D, asserting the observation of particular data relevant to those hypotheses. These propositions are all legitimate arguments for a probability symbol. Using equation (2.2), a Bayesian could calculate the probability, $p(H_1 D \mid I)$, that hypothesis 1 is true *and* that the observed data is as specified by D,

$$\begin{aligned} p(H_1 D \mid I) &= p(H_1 \mid DI)\; p(D \mid I) \\ &= p(D \mid H_1 I)\; p(H_1 \mid I). \end{aligned} \tag{2.3}$$

Here I is the background information that specifies the problem under discussion; in particular, it must, at the very least, specify the hypotheses alternative to H_1, and some logical connection between the data and each of the hypotheses. That is, I must specify sufficient information to permit unambiguous assignment of the various probabilities needed for a calculation.

It is a trivial matter to solve equation (2.3) for $p(H_1 \mid DI)$ to obtain *Bayes' Theorem*,

$$p(H_1 \mid DI) = p(H_1 \mid I)\frac{p(D \mid H_1 I)}{p(D \mid I)}. \tag{2.4}$$

Bayes' theorem, a consequence of the product rule, is the most important calculating tool in Bayesian probability theory. This is because it describes one of the most important processes in science: learning about hypotheses from data. In particular, it tells us how to update the *prior probability* of an hypothesis, $p(H_1 \mid I)$—which may summarize information about H_1 as primitive as the mere specification of alternatives or as complicated as the results of 1000 previous experiments—to its *posterior probability*, $p(H_1 \mid DI)$, which now includes the information provided by the data, D. The updating factor is the ratio of two terms. Only the numerator, $p(D \mid H_1 I)$, depends explicitly on H_1; it is called the *sampling distribution* in its dependence on D, or the *likelihood function* in its dependence on H_1. The denominator, called the *prior predictive probability* or the global likelihood, is independent of H_1 and is thus simply a normalization constant. It can be calculated by summing the product of the prior and the likelihood function over all alternative hypotheses H_i,

$$p(D \mid I) = \sum_i p(H_i \mid I) p(D \mid H_i I), \tag{2.5}$$

as shown in Loredo (1990) and references therein.

To illustrate the use of Bayes' theorem, imagine that we want to determine the distance, l, to some object from N measurements, m_i, contaminated by noise. A possible model for these data is

$$m_i = l_{\rm true} + \epsilon_i, \tag{2.6}$$

where ϵ_i is an unknown 'error' contribution. In this as in all parameter estimation problems, both frequentist and Bayesian approaches assume the truth of some parametrized model for the observations, and seek to determine the implications of the data for the values of any unknown parameters. A Bayesian does this simply by using Bayes' theorem to calculate the probabilities of various hypotheses about $l_{\rm true}$, given the data, $D = \{m_i\}$, and the background information, I, that specifies the model and anything known about $l_{\rm true}$ before consideration of the data, including the results of any previous measurements. All such probabilities could be calculated from suitable integrals of the posterior density, $p(l \mid DI)$, defined by $p(l < l_{\rm true} < l + dl \mid DI) = p(l \mid DI)dl$. Thus specifying $p(l \mid DI)$ as a function of l would completely represent the Bayesian solution to this problem. From this function we could calculate the probability that $l_{\rm true}$ is between two values, a and b, simply by integrating over l,

$$p(a < l_{\rm true} < b \mid DI) = \int_a^b p(l \mid DI)dl. \tag{2.7}$$

If our information about possible sources of error leads to the common Gaussian probability assignment for the values of ϵ_i, with zero mean and variance σ^2, a simple calculation, using a uniform prior distribution for l, yields a posterior density for l that is also Gaussian, with mean equal to the sample mean, $\overline{m}$, and standard deviation from the familiar 'root-N' rule, $\sigma/\sqrt{N}$:

$$p(l \mid DI) = \frac{N}{\sigma\sqrt{2\pi}} \exp\left[-\frac{N(l-\overline{m})^2}{2\sigma^2}\right]. \tag{2.8}$$

Using this distribution one can show that there is a probability of 0.68 that l is in the range $\overline{m} \pm \sigma/\sqrt{N}$, and that $p(l \mid DI)$ is higher within this region than outside of it. This region is called a 68% credible region, or a 68% highest posterior density (HPD) region. The details of this simple calculation, and a discussion of its robustness with respect to the choice of a prior, are available in Loredo (1990).

2.2 Frequentist Statistics

For a frequentist, the probability $p(A)$ is the long-run relative frequency with which A occurs in an infinite number of repeated experiments, or throughout an infinite ensemble. With this understanding of probability, the argument of a probability symbol cannot be an arbitrary proposition, but must be a proposition about a *random variable*, a quantity that can meaningfully be considered to vary throughout a series of repeated experiments or throughout a physically meaningful ensemble. This greatly restricts the domain of probability theory. In particular, the probability of an hypothesis is a meaningless concept in frequentist statistics. This is because a particular hypothesis is typically either true or false in *every* repetition of an experiment. Its frequentist probability, to the extent that it is meaningful, is either one or zero; but understood in this sense, the probability of an hypothesis is nearly always inaccessible (we do not know if it is true or not), and therefore scientifically uninteresting. Crudely, frequentist probabilities describe fluctuations, and hypotheses do not fluctuate.

Denied the concept of the probability of an hypothesis, frequentist statistics rejects the use of Bayes' theorem for assessing hypotheses. The mathematical correctness of the theorem is not questioned, and it is used to calculate distributions of random variables conditional on the values of other random quantities. But the application of the theorem to calculate probabilities of hypotheses is forbidden.

Of course, the principal use of probability theory in science is to assess the plausibility or viability of hypotheses. Barred from calculating the probability of an hypothesis, other ways must be found to assess hypotheses using frequencies. As a result, the discipline known as *statistics* was developed, distinct from, but relying on, probability theory. Statistics assesses hypotheses by noting that, though an hypothesis is not a random variable, data may be considered random. Thus hypotheses are assessed indirectly, by making use of the frequencies of different data sets that one might see if the hypothesis were true. This is accomplished as follows.

1. First, specify a procedure, Π_S, for selecting an hypothesis based on one or more characteristics of the data, which we denote by $S(D)$ (S may be a vector). The function, $S(D)$, is called a *statistic*.
2. Since the data are random, the function $S(D)$ is also random, and is therefore amenable to a frequentist description. Calculate the *sampling distribution* of S, $p(S \mid H)$, from $p(D \mid H)$.
3. Use the sampling distribution to characterize the long-term behavior of applying Π_S to the variety of data predicted by H.
4. Apply Π_S to the actually observed data.

The role of probability theory in frequentist statistics is limited to steps 2 and 3. In practice, these steps are sometimes performed with Monte Carlo calculations, and the overall performance of Π_S is characterized by averages over the simulated data. Unfortunately, frequentist theory offers no direct guidance for step 1—specifying a procedure and a statistic—or for choosing what characteristics of the procedure one should calculate in step 3. Intuition has been the dominant guide for specifying procedures, the relevant characteristics of the procedures, and the choice of statistic. In practice, a few procedures dominate statistical practice (estimators, confidence intervals, significance tests), and a few characteristics are considered most relevant for each (mean and variance of estimators, confidence level of confidence intervals, significance level ['false-alarm probability'] of a significance test). But there is seldom agreement on what statistic is the best to use in frequentist procedures for any but the simplest problems. Finally, note that the role of probability theory ends before the procedure is applied to the actually observed data in step 4—just where the role of probability theory begins in the Bayesian approach. Thus all probabilities quoted in frequentist analyses are to be understood as properties of the *procedure* used, not as properties of the single inference found by applying the procedure to the one observed data set.

For example, consider again the distance measurement problem whose Bayesian solution we outlined above. A frequentist would reject Bayesian probability statements about $l_{\rm true}$, arguing that $l_{\rm true}$ remains constant in repeated observations and is therefore not amenable to a frequentist description. Instead, the frequentist would develop methods that assess hypotheses about $l_{\rm true}$ using only the frequency distribution of the observed widths, w_i, which would vary from observation to observation, and which are therefore legitimate 'random variables.'

Frequentist approaches to this problem depend on what kind of hypotheses regarding $l_{\rm true}$ one may wish to consider. Two types of hypotheses are commonly considered, the specification of a single value of l to be considered as an estimate of $l_{\rm true}$ (point estimation), or the specification of a range of l values asserted to contain $l_{\rm true}$ (interval estimation). For point estimation, the procedure adopted is to assert that the true value of the parameter is equal to the value of the chosen statistic, which is here called an *estimator*. Usually several possible estimators are proposed (perhaps as a parametrized class), their sampling distributions are calculated, and one is chosen as 'best' based on how well the estimates it produces would cluster around $l_{\rm true}$ if the observations were repeated many times. For interval estimation, the procedure adopted is to assert that the true parameter value lies within an interval-valued statistic, $[l_1(\{m_i\}), l_2(\{m_i\})]$. The *confidence level* of the interval is calculated by determining the frequency with which this assertion would be correct in the long run. Again, intervals based on different possible statistics could be proposed, and one is selected based on how large the resulting intervals would be in the long run.

In the case of Gaussian errors described above, the accepted frequentist procedures give results identical to the Bayesian calculation described above: $l_{\rm true}$ is estimated to be $\overline{m}$, and a 68% confidence interval for $l_{\rm true}$ is $\overline{m}\pm\sigma/\sqrt{N}$. The sample mean is chosen as an estimator by focusing on two characteristics of the sampling distribution of possible estimators: their mean ('center of mass') and variance. Specifically, one restricts consideration to the class of estimators with mean equal to the true value of $l_{\rm true}$, so that

$$\int \hat{l}(D)\; p(D \mid l_{\rm true})\; dD = l_{\rm true}. \tag{2.9}$$

This implies that the average of the estimates found in many repetitions will converge to the true value of l. Estimators that satisfy equation (2.9) are called *unbiased*. There are an infinite number of unbiased estimators, but among them the estimator based on the sample mean converges most quickly, in the sense of having the smallest variance,

$$\int [\hat{l}(D) - l_{\rm true}]^2 \; p(D \mid l_{\rm true}) \; dD. \tag{2.10}$$

The 68% confidence interval will include $l_{\rm true}$ 68% of the time in many repeated experiments; that is, $\overline{m}$ will be within $\pm\sigma/\sqrt{N}$ of $l_{\rm true}$ with a frequency of 0.68:

$$\int_{l_{\rm true}-\sigma/\sqrt{N}}^{l_{\rm true}+\sigma/\sqrt{N}} p(\hat{l}(D) \mid l_{\rm true}) \; d\hat{l}(D) = 0.68. \tag{2.11}$$

Of course, many intervals satisfy an equation like (2.11); for example, the interval $[-\infty, \overline{m} + 0.47\sigma/\sqrt{N}]$ is also a 68% confidence interval. Additionally, an interval could be specified by using some statistic other than the sample mean. The familiar 'root-N' interval is chosen because it is the shortest 68% interval based on the estimator, or because it is the symmetric interval.

2.3 Conceptual Comparison of Bayesian and Frequentist Approaches

For the Gaussian example outlined above, Bayesians and frequentists agree on the final result. Nevertheless, it should be clear that the results mean very different things, not because Bayesians and frequentists interpret the same calculation differently, but because, in fact, they calculate very different things. Only by coincidence are their results identical in the Gaussian case. Unfortunately, the Gaussian distribution is so ubiquitous in statistics that this coincidence tends to mislead intuition. In the following sections, we will discuss examples where such a coincidence does not occur. We will first educate our misled intuition with simple 'toy' problems, where the contrast between Bayesian and frequentist methods is stark, and only then consider more complicated, practical problems.

But before moving on to examples, we will briefly elaborate on some of the conceptual and methodological differences that are already apparent in the Gaussian example. We want to know, not only *which* approach to inference is superior, but also *why* it is superior. Only with this understanding can we have confidence in the application of methods to problems whose complexity prevents our intuition from unambiguously identifying the superior result.

Much can be—and has been—written comparing the conceptual and mathematical foundations of Bayesian and frequentist statistics; references to some of this literature are provided in Loredo (1990) and in Lindley (1990), and a useful and extensive discussion is available in the recent book of Howson and Urbach (1989). Here we will emphasize two specific points of contrast between Bayesian and frequentist methods that have immediate practical consequences for scientists with real data to analyze. The examples in the remainder of this paper will concretely illustrate these points of contrast.

2.3.1 *What statistic should be used?*

The first point of contrast is of great practical significance. Bayesian inference is a problem-solving theory; given a problem, it provides a solution. In contrast, frequentist statistics is a solution-characterizing theory. It requires the user to come up with tentative solutions, and merely provides tools for characterizing them. The Bayesian solution to a well-posed problem is found by calculating the probabilities of the various hypotheses involved, using the rules of probability theory. In the process, the theory automatically identifies what statistics to calculate to optimally extract information from the data. In contrast, frequentist statistics does not provide a unique solution to a problem. Instead, it must be presented with a class of procedures. Probability theory is used to calculate certain properties of these procedures, and the investigator is left to choose from among them based on these properties. In fact, the theory does not even specify how these properties should affect the choice of a procedure. In addition, the theory cannot incorporate prior information of even the simplest kind, such as mere specification of an allowed range for a parameter, as we will see in Section 5.

For example, in the Gaussian estimation problem just considered, the Bayesian approach led directly and uniquely to the familiar result. In contrast, frequentist point estimation required specification of a class of estimators: those that are unbiased. But there is a certain arbitrariness to this specification, in that the intuitive notion behind bias would be equally well served by, say, the median or the mode of the sampling distribution, which in general will be different from the mean (though not in the Gaussian case, which is misleadingly simple). Also, by itself the notion of bias was not sufficient to specify an estimator; in fact, any estimator can be made unbiased by calculating its bias and subtracting it. Thus an additional property had to be considered to select from among competing estimators, the variance of their sampling distributions. Yet the variance of a particular estimator may increase when its bias is removed, so the relative merits of variance and bias should have been specified somehow. Finally, as appealing as criteria such as bias and variance may be to intuition, there is nothing fundamental about them, and, for example, there is a growing literature on the use of biased estimators (Efron 1975; Zellner 1986). Similar criticisms may be leveled against confidence intervals.

The practical consequence of frequentist nonuniqueness is that a complicated problem with a single Bayesian solution may have several frequentist solutions, each based on a different choice of statistic, and each giving a different answer to a particular question, with no compelling criteria for deciding between them. A striking example is provided by analyses of the neutrinos detected from supernova SN1987A. Literally dozens of analyses of the two dozen detected events have been published, many of them considering the same model, but differing in choice of statistic, and to varying degrees in the final conclusions reached, with no compelling arguments for preferring one analysis to another. In fact, most of these analyses are not even correct from the frequentist viewpoint, most investigators having confused the concept of a Type I error probability with a confidence level (Loredo and Lamb 1989, 1992). A correct frequentist analysis of these data is presented in Loredo and Lamb (1989), and a Bayesian analysis is presented in Loredo and Lamb (1992).

2.3.2 *What alternatives are relevant for inference?*

The second point of contrast between Bayesian and frequentist methods concerns the types of alternatives considered relevant for inference. In many respects, this is the essential feature distinguishing the approaches. The alternatives whose probabilities are considered in Bayesian calculations are alternative *hypotheses*. Those considered in frequentist calculations are alternative *data*. This distinction manifests itself in the Gaussian estimation problem above in two ways. First, the Bayesian result is conditional on the data (see, *e.g.*, equation [2.8]), whereas the frequentist result is conditional on a single hypothesis (see, *e.g.*, equations [2.9]-[2.11], which are conditional on $l_{\rm true}$). Second, Bayesians consider sums or averages over hypotheses (the integral in equation [2.8] is over l), whereas frequentists average over hypothetical data (the integrals in equations [2.9]-[2.11] are over possible sets of data). This point has been emphasized by Efron (1978). Only in special circumstances are frequentist data averages numerically equal to Bayesian hypothesis averages. The Gaussian distribution provides one such circumstance because of its symmetry between data and parameter.

On a purely intuitive level, this distinction should immediately raise doubts about frequentist methods. After all, in any real experiment, the observed data are the *fact*, and it is the possible hypotheses which are hypothetical. That is, we are uncertain about the hypotheses, not the data. Bayesian inference describes the uncertainty regarding the hypotheses by calculating the probabilities of those hypotheses conditional on the one fact we are certain of: the data. Frequentist statistics instead assesses procedures using the probabilities of hypothetical data conditional on the truth of a particular hypothesis. Yet the reason we perform experiments is that we do not know what hypothesis is true!

To be fair, some frequentist procedures can be cleverly designed to have characteristics that are independent of which particular hypothesis (within a class of alternatives) is true. For example, a special property of the Gaussian distribution makes equation (2.11) correct for every possible value of $l_{\rm true}$, so the true value need not be known. But this is not generally possible, particularly in multiparameter problems, or in hypothesis testing situations. In these cases, one must simply pick an hypothesis (*e.g.*, a model with its parameters fixed at their best-fit values), and assume it is true to calculate the needed distributions, despite being virtually certain that this assumption is false. The resulting statistical procedures are then not rigorously correct, but in frequentist statistics this is often 'just about the only game in town' (Press, *et al.* 1986).

On a more technical level, averaging over hypothetical data makes frequentist results depend on the precise nature of data which might have been seen, but were not. Unfortunately, factors irrelevant to the hypotheses under consideration can play an important role in determining what other data might have been seen, and these factors can affect frequentist inferences in troubling ways, as will be demonstrated in Section 3. In Bayes' theorem, the only relevant feature of the data is the dependence of its probability on the various hypotheses under consideration, not its relationship to other hypothetical data.

Finally, by seeking procedures which have good performance averaged over many hypothetical data, frequentist statistics generally trade off good performance in individual cases for good long-term performance. We will explicitly demonstrate that this occurs in Section 3. This occurs because the probabilities characterizing frequentist procedures are properties of the *procedures*, and not of the specific conclusions reached from a particular data set. For example, the probability of 0.68 assigned to the 'root-N' confidence interval in Gaussian estimation is not a property of the particular interval one would find by applying the rule to a particular data set, but is instead a property of the procedure of applying this rule to many data sets drawn from a Gaussian distribution. One can easily construct other procedures (for example, procedures that eliminate 'outlier' points that are beyond some threshold from the sample mean) that, for certain data sets, specify the *same* interval as the 'root-N' procedure, but assign it a *different* confidence level, even though the procedures assume the same Gaussian model. This is because frequentist probability theory can only characterize procedures, not particular inferences. This is made clear in the four step process outlined in the previous subsection: the role of probability theory ends when the actual data is analyzed in step 4. Bayesian probability theory works very differently. It applies probability theory directly to the observed data, seeking the best inference possible for the single case at hand.

The next two sections explore the ramifications of basing inferences on averages over hypotheses versus averages over hypothetical data. In Section 3, we demonstrate the irrelevance of averaging over hypothetical data, and in Section 4, we demonstrate the necessity of averaging over hypotheses. Along the way, the nonuniqueness of frequentist procedures will be illustrated by noting the variety of frequentist solutions to the problems considered.

3. THE IRRELEVANCE OF SAMPLE AVERAGING: STOPPING RULES AND RECOGNIZABLE SUBCLASSES

3.1 Stopping Rules: The χ^2 Goodness-of-Fit Test

After the sample mean and the 'root-N' rule, there is no statistic more familiar to astronomers than Pearson's χ^2 and its generalizations. One of the most widely used statistical methods in astronomy is the χ^2 'goodness-of-fit' test which evaluates a model based on the calculation of the probability P that χ^2 values equal to *or larger than* that actually observed would be seen if the model is true. If P is too small (the critical value is usually 5%), the model is rejected.

To create a quantity which takes on values between 0 and 1 to replace the straightforward Bayesian notion of the probability of an hypothesis, goodness-of-fit tests, and many other frequentist procedures, are forced to consider, not only the probability of seeing the actually observed data (which is almost always negligible), but the probability of seeing other hypothetical data—those that would produce a larger χ^2 value—as well. This peculiar line of reasoning has troubled scientists and statisticians for as long as such tests have been advocated. 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. Here is a simple example of this phenomenon that is widely known among statisticians, but not familiar to most physical scientists (see, *e.g.*, Berger and Berry 1988 for recent examples and references, and Lindley and Phillips 1976 for a more extensive pedagogical discussion).

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 traditional 95% critical 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 $N = 96$, theory predicts that on the average one should observe $n_A = 9.6$ 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). (A more accurate calculation using the binomial distribution gives $P = 0.073$, still acceptable at the 95% level.)

Unknown to the theorist, the observer planned the observations by deciding beforehand that they would proceed until 5 A stars were found, 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 to be 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 exact calculation using the negative binomial distribution gives $P = 0.032$.)

The two analyses differ because different ideas about the 'stopping rule' governing the observations lead to different ideas of what other data sets *might* have been observed. In this way, two different sample spaces are proposed to describe the same observation, and many others could be envisaged. Which one of all the possible calculations should be believed?

Some reflection should make us uneasy about accepting the observer's analysis. If, because of poor weather, the observing run had been cut short before 5 A stars were seen, how then should the analysis proceed? Should it include the probability of poor weather shortening the observations? If so, doesn't consistency then demand that it include the probability of poor weather in the calculation when the observations *were* able to be completed? Similar arguments can be made regarding any possible phenomenon that could have shortened the observing run (equipment failure, sickness, *etc.*), each resulting in a different choice for the sample space.

Though these arguments cast doubt on the observer's analysis, they also cast doubt on the theorist's analysis, since this analysis assumed that N was fixed, even though N could very well vary in repeated observations. In fact, it is essentially impossible to correctly characterize the frequency distribution of N that would be realized if the observations were repeated. Yet the intuitions of most scientists are not troubled by this fact: the other values of N that might have been found, but were not, have no bearing on the evidence provided by a single data set with a particular value of N.

The Bayesian calculation (Lindley and Phillips 1976; Berger and Berry 1988) satisfies this intuitive desideratum. But not only does it not consider other values of N, it also does not consider other values of n_A or n_X. In their place it considers other hypotheses about a. It thus differs from all frequentist calculations, but uses a statistic that can be approximated by the theorist's χ^2, and accepts the theory (Jeffreys 1939, §5.1). Like its frequentist counterparts, this calculation requires specification of a sample space in order to assign a likelihood function. The difference is that the Bayesian calculation focuses on the functional dependence of the likelihood on the *hypotheses*, whereas frequentist calculations focus on the dependence of sampling distributions on the *data*. As a result, Bayesian conclusions are independent of features of the sample space that can affect frequentist conclusions in troubling ways.

To be a bit more explicit, both the binomial distribution and the negative binomial distribution depend on the value of a in the same way, so Bayesian calculations by the theorist and observer would lead to the same conclusion. But the distributions differ in their dependences on N and $n_{A,obs}$, causing frequentist analyses to reach differing conclusions, depending on which distribution is used. Similarly, though variations in weather or equipment reliability could affect N and $n_{A,obs}$, the probabilities of bad weather or equipment failure are independent of a, and a Bayesian calculation with sample spaces including such phenomena will also reach the same conclusion.

3.2 Recognizable Subclasses: Estimators and Confidence Intervals

In February of 1987, approximately two dozen neutrinos were detected from supernova SN1987A in the Large Magellanic Cloud at three detectors located in Japan, the United States, and the Soviet Union. The energies and arrival times of each of the neutrinos were measured. The data are adequately modeled by thermal emission from a sphere with exponentially decaying temperature. The event arrival times and energies allow measurement of the initial temperature and decay time of the emitting surface, as well as the start time of the neutrino burst.

Many facets of the emission and detection processes make the analysis of these data very complicated (Loredo and Lamb 1989, 1992), obscuring some of the fundamental difficulties arising in a frequentist analysis of these data. To illustrate one of these difficulties, of broad significance for statistics, consider the simplified problem of inferring only the starting time of the burst from the event arrival time data, ignoring the energy data and assuming that the decay time is known. Accurate inference of arrival times is often of great interest to astronomers, particularly when observations from different detectors are to be combined; see, *e.g.*, Lobo (1990).

In this simplified problem, the probability density that an event will arrive at time t is given by a truncated exponential distribution,

$$p(t) = \begin{cases} 0, & \text{if } t < t_0; \\ \frac{1}{\tau}\exp(-\frac{t-t_0}{\tau}), & \text{if } t \geq t_0, \end{cases} \tag{3.1}$$

with τ known; we want to estimate t_0 from the sample. This problem is mathematically identical to a problem considered by Jaynes (1976). We follow his analysis here.

Consider first a frequentist analysis. Since t_0 plays the role of a location parameter, like the mean of a Gaussian, we can construct a statistic by relating the mean value of t to t_0 (this is an example of the 'method of moments' advocated by Pearson; see Eadie *et al.*

1971). The population mean is $\langle t\rangle \equiv \int t\ p(t)dt = t_0+\tau$, therefore an unbiased estimator of t_0 is

$$\hat{t} \equiv \frac{1}{N}\sum_{i=1}^{N}(t_i-\tau). \tag{3.2}$$

We can calculate the sampling distribution of $\hat{t}$ analytically using characteristic functions (related to equation (3.1) by a Fourier transform); the result is (Jaynes 1976),

$$p(\hat{t}\mid t_0) = (\hat{t}-t_0+\tau)^{N-1}\exp\left[-N(\hat{t}-t_0+\tau)\right]. \tag{3.3}$$

We can use this to get a confidence interval for t_0 of any desired size.

Now suppose that in a particular observation, with $\tau=1$, we observe three events at times $t=12, 14$, and 16. From (3.2), we would estimate t_0 to be $\hat{t}=13$, and from (3.3), the shortest 90% confidence interval for t_0 can be calculated to be

$$12.15 < t_0 < 13.83. \tag{3.4}$$

But wait; the earliest event was observed at $t_1=12$, yet both the estimate of the burst start time, and the entire 90% interval, are at later times, $t>12$, where we know it is impossible for t_0 to lie!

What is going on here? It is certainly true that if we repeat this experiment lots of times with simulated data drawn from (3.1), the average of the $\hat{t}$ values will converge to the true value of t_0, and the 90% confidence regions will include the true value in 90% of the simulations. But one can show that the confidence region will not include the true value *100% of the time* in the subclass of samples that have $\hat{t}>t_1+0.85$, and we can tell from the data whether or not any particular sample lies in this subclass. The worse-than-stated performance of the confidence region in this subclass also implies that confidence regions will be wider than they need to be for samples that do not lie in this subclass; poor behavior within the bad subclass is made up for by better-than-stated behavior outside the bad subclass.

This problem is called the problem of *recognizable subclasses:* a statistic that is good in the long-run may be poor in individual cases *that can be identified from the data.* Frequentist procedures, by seeking good performance averaged over many hypothetical data, can throw away useful information that is relevant for the single-case inference. This phenomenon is reviewed by Cornfield (1969), who argues that to avoid it, one is forced to the Bayesian approach. The phenomenon of recognizable subclasses, which arises because of the role of long-term behavior in frequentist statistics, emphasizes that,

> The value of an inference lies in its usefulness *in the individual case*, and not in its long-run frequency of success; they are not necessarily even positively correlated. The question of how often a given situation would arise is utterly irrelevant to the question how we should reason when it *does* arise. (Jaynes 1976)

The Bayesian solution to this problem is both simple and intuitively appealing. The likelihood of the data, given the model, I, specified by equation (3.1) with $\tau=1$, is just the product of N truncated exponentials at times t_i,

$$p(\{t_i\}\mid t_0 I) = \begin{cases} \exp(Nt_0)\exp\left(-\sum_{i=1}^{N} t_i\right), & \text{if } t_0 \le t_1; \\ 0, & \text{if } t_0 > t_1. \end{cases} \tag{3.5}$$

Using a uniform prior for t_0, which expresses ignorance of a location parameter (see Section 6), the normalized posterior is, from Bayes' Theorem,

$$p(t_0 \mid \{t_i\} I) = \begin{cases} N \exp\left[N(t_0 - t_1)\right], & \text{if } t_0 \le t_1; \\ 0, & \text{if } t_0 > t_1. \end{cases} \tag{3.6}$$

This reversed truncated exponential distribution is our full inference about t_0. We can summarize it in various ways. The most probable value of t_0 is just $t_1 = 12$; this is certainly reasonable, though the mode is not a good summarizing value for such a skew distribution. A more characteristic value is the mean value, which is $x_1 - 1/N = 11.66$. The 90% credible interval is $11.23 < t_0 < 12.0$; this is entirely in the allowed range, and is less than half the size of the frequentist confidence interval. The Bayesian calculation has given an entirely reasonable answer with over twice the precision of the frequentist inference; moreover, the most complicated mathematics it required was the evaluation of the integral of an exponential.

Of course, *if* the poor behavior of the $\hat{t}$ statistic was noticed, a frequentist hopefully could develop some other procedure that behaves better, probably appealing to such notions as sufficiency and ancillarity to try to reduce the effect of recognizable subclasses. But such considerations never even arise in the Bayesian analysis; probability theory automatically identifies what statistic to use, and expert knowledge of such notions as sufficiency and ancillarity (which are of limited applicability anyway) is not required.

4. THE RELEVANCE OF HYPOTHESIS AVERAGING: NUISANCE PARAMETERS

An immediate consequence of the rejection of the concept of the probability of an hypothesis is that no frequentist procedures are available for assessing propositions like, 'either hypothesis H_1 or hypothesis H_2 is true.' A Bayesian would simply calculate the probability of this proposition, denoted $H_1 + H_2$, where the '+' sign here represents logical disjunction (logical 'or'). From the sum rule and the product rule, one can derive the disjunction rule,

$$p(H_1 + H_2 \mid I) = p(H_1 \mid I) + p(H_2 \mid I) - p(H_1 H_2 \mid I). \tag{4.1}$$

If, as is often the case, the hypotheses under consideration are exclusive, so that only one of them may be true, then $p(H_1 H_2 \mid I) = 0$, and the disjunction rule takes the familiar form,

$$p(H_1 + H_2 \mid I) = p(H_1 \mid I) + p(H_2 \mid I). \tag{4.2}$$

This entire section is devoted to the important consequences of the absence of a frequentist counterpart to equation (4.2).

What types of problems have a structure requiring us to consider disjunctions of hypotheses? In fact, many, and perhaps most real statistical problems have this structure. But because frequentist methods cannot deal with such problems directly, simpler problems are often substituted for the real problem without comment. This is done so commonly that few are aware of it.

For example, whenever we want to assess the viability of a parametrized model *as a whole*, we are in a situation that logically requires consideration of disjunctions of hypotheses. This is because data never specify with certainty a single point in parameter space as the only possible one; rather, the data lead us to prefer certain regions of parameter space over others. Thus a full assessment of a model should take into account the probabilities that the model is true with every set of parameter values not ruled out by the data. Unable to do this, frequentist hypothesis tests assess a model by considering only the best-fit parameter point. A consequence is that there is no fully satisfactory way for comparing different models with differing numbers of parameters. Consider comparing two models with the χ^2 test, one of which includes the other as a special case (say, when an additional parameter is zero). When attention is fixed only on the best-fit parameter points, the more complicated model will never fit worse, and will almost always fit better. Given the two values of χ^2 resulting from the fits, frequentist attempts to justify a choice of one model over the other first try to assess whether the more complicated model fits 'significantly better' by considering some additional statistic (such as one of the possible F-statistics that can be created from the two χ^2 values). If it does not fit significantly better, there are no formal grounds for preferring either model to the other, but the simpler model is always chosen, 'Ockham's razor' being invoked as making the simpler model more plausible *a priori*.

In contrast, Bayesian model comparison methods proceed straightforwardly by calculating the probability that each model is true (assuming, as do frequentist methods, that the true model is one of the ones being considered). This probability is calculated by integrating over the entire parameter space, using the continuous counterpart to equation (4.2). Instead of invoking Ockham's razor to assign simpler models larger *a priori* probabilities, one discovers that Bayesian methods *explain and justify* Ockham's razor: even when models are assigned *equal* prior probabilities, Bayes' theorem shows that simpler models are preferred *a posteriori* unless more complicated models fit significantly better. In effect, the integration over parameter space penalizes more complicated models if their extra degrees of freedom are wasted. This is accomplished without having to invent an arbitrary comparison statistic, without having to select an arbitrary critical significance level, and without having to invoke Ockham's razor. Also, the Bayesian calculation can sensibly deal with the possibly more subtle case of comparing two models that have the same number of parameters but that differ in complexity. Further discussion of this aspect of Bayesian methods is available in Loredo (1990) and in recent tutorials by Garrett (1991) and by Jefferys and Berger (1992). Worked applications in the physical and computational sciences can be found in Bretthorst (1988, 1990a,b), MacKay (1992), and Gregory and Loredo (1992).

4.1 Nuisance Parameters

In this Section, we will focus on another somewhat simpler class of problems whose solution requires consideration of disjunctions of hypotheses. These are problems with *nuisance parameters:* multiparameter problems in which one is particularly interested in only a subset of the parameters. Such problems arise in many ways. In some cases, modeling data relevant to the phenomenon of interest may require the introduction of parameters unrelated to the phenomenon, such as the intensity of a background rate or parameters related to the detector. In other cases, some of the parameters describing the phenomenon may be intrinsically uninteresting, as may be true of the phase of a periodic signal whose frequency and amplitude are of sole interest. Even when all the parameters describing a phenomenon are intrinsically interesting, it may be necessary to consider the implications of the data for some subset of the parameters for a variety of reasons. A particular scientific question may require attention to be focused on a subset of the parameters. Graphical presentation of the implications of the data may require consideration of one- and twodimensional subsets of the parameters; in fact, this may be the most common situation in which methods summarizing inferences for subsets of parameters are required. Finally, temporary reduction of the dimensionality of a problem, particularly when such a reduction can be accomplished analytically, can greatly increase its numerical tractability.

To formulate such problems mathematically, let θ represent the parameters of interest, and let ϕ represent the nuisance parameters, also known as incidental or uninteresting parameters. Given some data, we would like to make inferences about θ without reference to ϕ.

The Bayesian approach to this problem is straightforward: we simply calculate the posterior distribution for θ, $p(\theta \mid DI)$. This is done as follows. First, with Bayes' theorem, we calculate the full joint posterior distribution, $p(\theta, \phi \mid DI)$. Then the disjunction rule yields (Loredo 1990),

$$p(\theta \mid DI) = \int p(\theta, \phi \mid DI) \, d\phi. \tag{4.3}$$

For historical reasons, the process of integrating over nuisance parameters is called *marginalization*, and the resulting distribution, $p(\theta \mid DI)$, is called the marginal distribution for θ.

In contrast to the straightforward and unique Bayesian marginalization procedure, there is no universally accepted frequentist answer to the question of how to deal with nuisance parameters. Basu (1977) remarks, 'During the past seven decades an astonishingly large amount of effort and ingenuity has gone into the search for reasonable answers to this question,' and goes on to list nine different *categories* of frequentist solutions. Dawid (1980) also discusses a variety of frequentist methods for dealing with nuisance parameters. Of these, only two are commonly used in astrophysics, and we will focus on them here.

The first method is to estimate all of the parameters by constructing suitable estimators, $\hat{\theta}$ and $\hat{\phi}$, and then to replace ϕ by its estimate, making further inferences about θ assuming $\phi = \hat{\phi}$. We shall refer to this as the *conditional method*, since it makes inferences conditional on the hypothesis that $\phi = \hat{\phi}$. The weaknesses of this procedure, which takes no account at all of the uncertainty in the value of ϕ, should be obvious. In particular, if the model leads to strong correlations between ϕ and θ, the uncertainty in θ can be greatly underestimated by this procedure. Few careful scientists would knowingly adopt such a procedure unless ϕ were extremely well determined by the data. Yet some widely used procedures implicitly treat nuisance parameters in just this way. For example, most periodogram-based methods for measuring the frequencies of periodic signals implicitly assume that the true phase of the signal is the best-fit phase (see, *e.g.*, Scargle 1982). If these procedures were otherwise optimal, they could underestimate the uncertainty of the frequency and strength of a signal.

The second method is more sophisticated, and is based on the joint likelihood function for the parameters, $\mathcal{L}(\theta, \phi) \equiv p(D \mid \theta, \phi)$. The intuitively appealing notion on which all likelihood methods are based is that the hypothesis with the highest likelihood among those being considered is to be preferred. Thus to eliminate ϕ from consideration, a new *projected likelihood function*, $\mathcal{L}_p(\theta)$, is created by maximizing $\mathcal{L}(\theta, \phi)$ with respect to ϕ at each value of θ, that is,

$$\mathcal{L}_p(\theta) = \max_{\phi} \mathcal{L}(\theta, \phi), \tag{4.4}$$

where $\max_\phi$ denotes the maximum with respect to ϕ with θ fixed. Inferences about θ are then made from $\mathcal{L}_p(\theta)$ as if it were a normal likelihood function. The projected likelihood function is also called the profile likelihood or the eliminated likelihood.

As suggested by its name, contours of the projected likelihood function correspond to the geometric projection of contours of the full likelihood function into the subspace spanned by the interesting parameters. If the conditional method described above were used with the likelihood function, it would correspond to making inferences based on a cross section of the likelihood function rather than a projection. Just as the shadow of an object onto a plane is generally larger than, and never smaller than, any cross section of the object parallel to that plane, so projected likelihood functions are generally broader than conditional likelihood functions, and in this manner attempt to account for the uncertainty in the nuisance parameters.

When the errors are described by a Gaussian distribution, the likelihood function is proportional to $\exp(-\chi^2/2)$, with the familiar χ^2 statistic appearing in the exponential. Then use of the projected likelihood function corresponds to use of a projected χ^2 statistic, where the projection is found by minimizing χ^2 with respect to the nuisance parameters. This method for treating nuisance parameters is widely used in astrophysics, and particularly in X-ray astronomy, where it has been carefully described by Lampton, Margon, and Bowyer (1976). More recently, its use has been advocated by Press, *et al.* (1986). Both groups of authors are careful to note the inadequacy of the conditional method.

For linear models with Gaussian errors and uniform prior densities for the parameters, Bayesian marginalization over ϕ leads to a marginal distribution for θ that is precisely proportional to the frequentist projected likelihood function, both functions being themselves Gaussian. Thus both methods lead to the same final inferences. But again, this correspondence is a consequence of special properties of the Gaussian distribution, and will not hold for nonlinear models or for non-normal errors. Thus inferences based on marginal distributions and on projected likelihoods, even in the absence of important prior information, will differ, and we must ask: which inferences should we believe?

To address this question, we will analyze a simple problem with both methods, but arrange the problem to magnify the difference between them to the point where our intuition can unambiguously decide which method is superior.

4.2 Coin Tossing With a Nuisance Parameter

One of the most common problems for illustrating statistical ideas is the determination of the bias of a coin. We will now consider a biased coin problem, introducing a nuisance parameter associated with the date of the coin. The following problem is adapted from Basu (1975).

Imagine that we have a bucket with 50 coins in it. We know that 40 of the coins, one each from the years 1951 to 1990, are biased in the same (unknown) direction, and that the remaining 10 coins are biased in the other direction. Further, we know that the 10 'minority' coins all have the same (unknown) date, also in the period 1951 to 1990. We want to know the direction of the bias—heads or tails—of the majority of coins.

In terms of the obligatory greek letters, we can formulate this problem as follows. Let θ be the unknown bias of the majority, $\theta = H$ (heads) or T (tails). Let ϕ be the year of the coins with the opposite bias, $\phi = 1951$ to 1990. We want to determine θ; the additional parameter, ϕ, is a nuisance parameter.

Consider inferring θ from two possible data sets. To obtain the first data set, we take a single coin from the bucket and, either by measuring it or flipping it many times, determine that the direction of its bias, B, is heads: $B = H$. But we throw it back into the bucket before noting its date. We obtain the second data set by taking a single coin from the bucket, finding that for it $B = H$, and additionally noting that it's date, Y, is 1973. If we have to draw a conclusion about θ from such meager data, what conclusion should we draw in each case?

Intuition suggests that, since we don't know the date identifying the 10 coins with the different bias, each data set provides us with the same information about the bias of the majority of the coins. Since we are much more likely to choose a coin from the majority, we would conclude in both cases that the majority are biased towards heads: $\theta = H$. We wouldn't feel certain of this, but it is clearly the most plausible choice.

Now we will analyze the two data sets using likelihood methods. The first data set is the single datum, $B = H$. We begin by noting that knowledge of the date, ϕ, specifying the special coins cannot help us to determine the bias of a particular coin of unknown date, so the probability of this data is independent of ϕ: $p(B \mid \theta, \phi) = p(B \mid \theta)$. Thus for the first data set, the problem is reduced to one without a nuisance parameter. The likelihoods of the two possible values of θ are just the probabilities of seeing $B = H$ for each value:

$$\mathcal{L}(\theta = H) \equiv p(B = H \mid \theta = H) = \frac{40}{50} = 0.8 \tag{4.5a}$$

$$\mathcal{L}(\theta = T) \equiv p(B = H \mid \theta = T) = \frac{10}{50} = 0.2. \tag{4.5b}$$

Thus our best guess, in the maximum likelihood sense, is that the bias of the majority is toward heads, in agreement with our intuition.

Now consider the analysis of the second data set, $B = H$, $Y = 1973$. Here we have information about the date, so we will calculate the joint likelihood of θ and ϕ, and then eliminate ϕ by maximizing with respect to it. By the product rule, we can break up the joint likelihood, $\mathcal{L}(\theta, \phi) \equiv p(B = H, Y = 1973 \mid \theta, \phi)$, into two factors,

$$\mathcal{L}(\theta, \phi) = p(Y = 1973 \mid B = H; \theta, \phi)\, p(B = H \mid \theta, \phi). \tag{4.6}$$

The second factor we've already calculated for the analysis of the first data set, so we need only calculate the first factor. This is the probability that a coin biased toward heads has the year 1973. If the bias of the majority is toward heads, then this probability is simply 1/40, independent of the value of ϕ. But if the bias of the majority is toward tails, then the coin must be in the minority; so $p(Y = 1973 \mid B = H, \theta = T, \phi)$ will be 1 if $\phi = Y$, and zero otherwise. Thus,

$$p(Y = 1973 \mid B = H, \theta, \phi) = \begin{cases} 1/40, & \text{for } \theta = H,\ \phi = 1973; \\ 1/40 & \text{for } \theta = H,\ \phi \neq 1973; \\ 1, & \text{for } \theta = T,\ \phi = 1973; \\ 0, & \text{for } \theta = T,\ \phi \neq 1973. \end{cases} \tag{4.7}$$

Multiplying equations (4.7) by the appropriate factors in equations (4.5), the joint likelihoods for the various values of θ and ϕ are,

$$\mathcal{L}(\theta, \phi) = \begin{cases} 1/50 = 0.02, & \text{if } \theta = H, \text{ for any } \phi; \\ 10/50 = 0.2, & \text{if } \theta = T \text{ and } \phi = 1973; \\ 0, & \text{otherwise.} \end{cases} \tag{4.8}$$

Maximizing over ϕ, the projected likelihood for θ has the values,

$$\mathcal{L}_p(\theta) = \begin{cases} 0.02, & \text{if } \theta = H; \\ 0.2, & \text{if } \theta = T. \end{cases} \tag{4.9}$$

The projected likelihood method leads to the conclusion that the direction of the bias of the majority is toward tails!

This is obviously a ridiculous conclusion. What has happened here? Consider the Bayesian solutions to these problems. To calculate the posterior probabilities of the various possibilities, we must assign prior probabilities to the values of θ and ϕ. Before considering the data, we have no reason to prefer $\theta = H$ to $\theta = T$, thus we assign $p(\theta = H) = p(\theta = T) = 1/2$. Similarly, we assign $p(\phi) = 1/40$, since there are 40 possible choices for ϕ, and we have no reason to prefer one to any other *a priori*. With uniform prior probabilities, Bayes' theorem gives posterior probabilities that are proportional to the likelihood functions, with the constant of proportionality chosen to make the distributions normalized.

For the first data set, Bayes' Theorem, using the likelihoods calculated above, gives posterior probabilities proportional to equations (4.5); in fact, the proportionality constant is just 1:

$$p(\theta \mid B = H) = \begin{cases} 0.8, & \text{for } \theta = H; \\ 0.2, & \text{for } \theta = T. \end{cases} \tag{4.10}$$

We reach the same conclusion as the frequentist: the bias of the majority of coins is probably toward heads.

For the second data set, we similarly find joint posterior probabilities proportional to equations (4.8),

$$p(\theta, \phi \mid B = H, Y = 1973) = \begin{cases} 0.02, & \text{if } \theta = H, \text{ for any } \phi; \\ 0.2, & \text{if } \theta = T \text{ and } \phi = 1973; \\ 0, & \text{otherwise.} \end{cases} \tag{4.11}$$

But to summarize the implications of the data for θ alone, we marginalize with respect to ϕ, summing (*not* projecting) the joint distribution over ϕ, to find,

$$p(\theta \mid B = H, Y = 1973) = \begin{cases} 0.8, & \text{for } \theta = H; \\ 0.2, & \text{for } \theta = T. \end{cases} \tag{4.12}$$

This posterior distribution is the same as equation (4.10), so the Bayesian concludes that both data sets yield the same information regarding θ, nicely agreeing with our intuition.

This example demonstrates that the projected likelihood method advocated by Lampton, Margon, and Bowyer (1976) and by Press *et al.* (1986) is fundamentally flawed. The reason is that, though projection identifies which particular hypothesis is most favored by the data, it takes no account of the actual number of alternative hypotheses. In the example just discussed, of all the possible hypotheses, the one hypothesis with $\theta = T$ and $\phi = 1973$ has higher probability than any particular one with $\theta = H$. But there are so many more of the latter that it is more plausible that one of them is true than that the $\theta = T$ hypothesis is true. The only way this can be taken into account is by *summing over hypotheses*, a process denied to frequentists by the frequency definition.

4.3 Frequentist Use of the Likelihood Function

In closing this Section, we note in passing that the discussion of this and the previous Sections addresses an issue that arises for many when first exposed to the Bayesian approach. Noting that the data enter Bayes' theorem through the likelihood function, it is tempting to assert that use of likelihood methods, though perhaps frequentist, should produce results essentially equivalent to Bayesian methods.

At least three objections can be made to this claim. First, likelihood methods 'ignore' prior information, or more precisely, assume very specific, uniform prior information. In many problems, our prior information is so much less informative than the data that 'ignoring' the prior information is completely reasonable, but this is not always the case, particularly when data are sparse, or when there is important prior information (as in the analysis of inverse problems, or the combination of results of different experiments).

Second, even when the use of a uniform prior is justified, likelihood methods still share many of the defects of other frequentist procedures to the extent that they rely on averages over hypothetical data (as in the construction of confidence intervals or the calculation of significance levels). In particular, the results of such methods may be sensitive to irrelevant features of the sample space, as discussed in Section 3. There, the observer and the theorist used different likelihood functions for the same data and theory (negative binomial and binomial likelihoods, respectively). But though these likelihood functions differ in their dependences on the data, they agree in their dependences on the parameter, which is the only aspect of the likelihood function relevant to the Bayesian calculation. These two objections are further elaborated upon in Basu (1975), Berger and Wolpert (1984), and Berger and Berry (1988).

The third objection is related to equation (4.2). It is that the likelihood is a point function, not a set function or measure function (Basu 1975). That is, it makes no sense to speak of the likelihood that either H_1 or H_2 is true, because a likelihood is a probability, not of an hypothesis, but of data. Thus likelihoods of different hypotheses cannot be added. For hypotheses labeled by a continuous parameter, say θ, this is reflected in that $\mathcal{L}(\theta)$ is a point function ascribing a likelihood to *each value* of θ, whereas a Bayesian posterior distribution, $p(\theta \mid DI)$, is a density, ascribing probabilities to *intervals* of θ. A consequence is that the integral of $\mathcal{L}(\theta)$ need not be unity; in fact, such an integral has no meaning within frequentist statistics. This property of the likelihood function is an objection to its use because many problems require a set function for their solution—just those problems that a Bayesian would address with equation (4.2).

The examples we present here illustrating these objections demonstrate that, though likelihood methods focus on the correct function, they use it improperly, and that a fully Bayesian calculation is required.

5. INFERENCE WITH THE POISSON DISTRIBUTION

Astronomers frequently observe distributions of discrete events, be they macroscopic events such as the occurrence of a supernova or the location of a galaxy, or microscopic events such as the detection of particles or radiation quanta from an astrophysical source. Often our information (or lack thereof!) about the relevant processes leads us to model the data with the Poisson distribution (see Jaynes 1990 for an instructive derivation of this distribution). In this Section we will discuss some simple but common problems requiring inferences based on the Poisson distribution. These show how the considerations of the previous Sections affect inferences in real, practical problems. For concreteness, we will discuss temporally distributed events; analyses of events distributed in space, angle, redshift, or energy, for example, would proceed in an analogous fashion.

The basic datum for temporally distributed events is the number of events, n, detected in some time interval, T. The Poisson distribution relates this datum to a rate function, $r(t;\theta)$, with parameters θ, such that the probability of the datum (the likelihood function) is,

$$p(n \mid \theta I) = \frac{(rT)^n}{n!} e^{-rT}. \tag{5.1}$$

If $r(t)$ varies significantly with time over the interval T, then rT should be replaced by the integral of $r(t)$ over the interval.

Two characteristics of the Poisson distribution complicate frequentist analyses. First, the distribution connects a real-valued quantity, r, with an integer-valued datum, n; it is not at all symmetric between data and parameters, as is the Gaussian distribution. Second, there is important prior information: r must be non-negative.

When n and rT are large, the Poisson distribution can be accurately approximated with a Gaussian distribution, allowing straightforward use of frequentist statistics such as Gaussian confidence intervals or χ^2. But astronomers frequently find themselves in situations where such an approximation is unjustified, but where a rigorous and precise inference is still required. Indeed, this is the rule rather than the exception in some fields, such as my own field of high energy astrophysics, where the great expense of an experiment or the uniqueness of the observed phenomenon makes it impossible to 'go out and get more data' so that Gaussian approximations can be used. In such cases, we must work directly with the Poisson distribution, without approximation.

5.1 Estimating a Poisson Mean

We first consider the simplest case: inferring the magnitude of a *constant* rate, r, from a single measurement of n events in a time T. The Bayesian inference for r is found by using Bayes' theorem, equation (2.4), to find a posterior distribution for r, $p(r \mid nI)$, where I represents the information that r and n are connected by the Poisson distribution, and any other information we may have about r and n.

Bayes' theorem requires a likelihood function $p(n \mid rI)$, a prior density $p(r \mid I)$, and a prior predictive distribution, $p(n \mid I)$. The likelihood function is given by equation (5.1). In addition to the likelihood function, the background information, I, must allow us to assign a prior density to r, from which we can calculate the prior predictive probability according to equation (2.5). Deferring a more careful discussion of priors to the following Section, we here adopt a uniform prior for r between the lower limit of $r = 0$ and some upper limit

$r_{\rm max}$,

$$p(r \mid I) = \frac{1}{r_{\rm max}}. \tag{5.2}$$

The upper limit is required to allow us to make the prior 'proper,' that is, normalized. In principle, there is always some known upper limit; for example, we know the radiation intensity from some source could not be so large that it would have vaporized the detector! In practice, the actual value of this limit will usually have a negligible affect on the resulting inferences. Indeed, in most problems where limits are required to make priors proper, Bayes' theorem is perfectly well behaved in the limit where the prior range is infinite. That will be the case here.

The prior predictive distribution—the normalization constant for Bayes' theorem—is found by integrating the prior times the likelihood with respect to r. This gives,

$$\begin{aligned} p(n \mid I) &= \frac{T^n}{n!} \int_0^{r_{\rm max}} dr\, r^n e^{-rT} \\ &= \frac{1}{T r_{\rm max}} \cdot \frac{\gamma(n+1, T r_{\rm max})}{n!}, \end{aligned} \tag{5.3}$$

where $\gamma(n,x) \equiv \int_0^x dx x^{n-1} e^{-x}$ is the incomplete Gamma function. Combining this equation with the prior and the likelihood, the posterior distribution for r is,

$$p(r \mid nI) = \frac{T(rT)^n e^{-rT}}{n!} \cdot \frac{n!}{\gamma(n+1, T r_{\rm max})}, \qquad \text{for} \quad 0 \le r \le r_{\rm max}. \tag{5.4}$$

Note that the normalization constant, $r_{\rm max}$, has cancelled out, so that the posterior depends on the prior range only very weakly, through the incomplete gamma function factor (which appears so that the posterior is normalized over the finite prior range). In fact, if the prior range is large, so that $T r_{\rm max} \gg n$, then $\gamma(n+1, T r_{\rm max}) \approx \Gamma(n+1) = n!$, and the posterior is just the first factor in equation (5.4),

$$p(r \mid nI) = \frac{T(rT)^n e^{-rT}}{n!}, \qquad \text{for} \quad r \ge 0. \tag{5.5}$$

Rainwater and Wu (1947) proposed using this distribution for analyzing nuclear particle counting data.

We could have derived this result in one line by arguing that, with a constant prior, the posterior is simply proportional to the likelihood; equation (5.5) is simply the likelihood function with a factor of T added so that it is normalized when integrated over r. We have followed a more rigorous path to demonstrate two facts of some practical significance. First, we have demonstrated that prior information may only weakly affect posterior inferences: equation (5.4) has a very weak dependence on $r_{\rm max}$, so that it is usually well approximated the simpler equation (5.5). Second, we have shown rigorously that a legitimate, normalized posterior results from considering the limit of infinite $r_{\rm max}$, and that this posterior is the same posterior one would find by taking the prior to be constant over an infinite range, and hence improper. This is important because improper priors, considered as the limit of a sequence of proper priors, are often convenient expressions of initial ignorance, and it is convenient to know that the limiting process need not be explicitly carried out. If the limit does not exist, the product of the (improper) prior and the likelihood will not be normalizable, signaling the investigator that more information must be specified to make the problem well-posed.

The posterior distribution is the full Bayesian inference for r. But for graphical or tabular display, it can be summarized in various ways. The mode (most probable value of r) is n/T, the posterior mean is $\langle r\rangle \equiv \int dr\ rp(r \mid nI) = (n+1)/T$, and the posterior standard deviation is $\langle r^2 - \langle r\rangle\rangle^{1/2} = \sqrt{n+1}/T$. When n is large, the Bayesian inference thus agrees with the standard $(n \pm \sqrt{n})/T$ estimate from a Gaussian approximation. But when n is small, the posterior is not at all symmetric about the mode, and these numbers do not adequately summarize the implications of the data. A more descriptive summary would be the posterior mode and the boundaries of a credible region containing, say, 95% of the posterior density. For example, if one event were seen, the most probable value of r would be $1/T$, and the 95% credible interval would extend from $r = 0$ to $r = 4.74/T$.

It is nearly always the case that the signal whose intensity, s, we are interested in is measured with contamination by some background signal with rate b. If b is known, the Bayesian inference for s follows from the above analysis by setting $r = s + b$, so that

$$p(s \mid nbI) = C\frac{T[(s+b)T]^n e^{-(s+b)T}}{n!}, \tag{5.6}$$

where $C = 1/p(n \mid bI)$ is a normalization constant that can be easily found by expanding the binomial $(s+b)^n$ and integrating over s. This gives

$$C = \left[\sum_{i=0}^{n} \frac{(bT)^i e^{-bT}}{i!}\right]^{-1}. \tag{5.7}$$

Helene (1983) proposed using equation (5.6) for analyzing multichannel spectra produced in nuclear physics experiments. Kraft, Burrows, and Nousek (1991; hereafter KBN) provide an excellent discussion of its application to astrophysical data, including a comparison with a frequentist solution to this problem.

Equations (5.5) and (5.6) are both useful and interesting. What is perhaps most interesting is the ease with which they are found within the Bayesian paradigm. Indeed, the problems in this Section are among the very first I tried to solve when I first learned about Bayesian inference, and here I am not alone (Gull 1988 describes a similar experience). But despite these being 'beginner' Bayesian problems, finding frequentist solutions to these problems is difficult. The frequentist counterpart to equation (5.5) (the $b = 0$ case) has been thoroughly discussed by Gehrels (1986), who points out that the discreteness of the data makes it impossible to find a rigorous confidence interval for r (one which covers the true value of r with a long-term frequency that is independent of r). Instead, a conservative interval must be used, so that the 68% interval covers the true value of r *at least* 68% of the time. In fact, when the true rate is small, the standard intervals described by Gehrels cover the true rate 100% of the time.

Physicists have not found it straightforward to extend this frequentist result to cases with $b \neq 0$. Hearn (1969) presented an incorrect procedure for use by nuclear physicists that O'Mongain (1973) applied to observations of astrophysical gamma-ray sources; only recently have Li and Ma (1983) pointed out that Hearn's method is incorrect. Intuition suggests a simple procedure: estimate the total rate, and simply shift the estimate and its confidence region down by b. Sard and Sard (1949) show that the resulting point estimate of s is a good frequentist estimate in the sense of being unbiased. KBN further note that the shifted confidence interval, truncated at $s = 0$, is also a rigorously correct confidence interval; that is, following this procedure will cover the true value with the stated frequency. Yet simple subtraction of the background will often lead to negative signal estimates when the signal rate is small. Similarly, KBN emphasize that though shifted truncated intervals have the stated frequency coverage, such intervals collapse to zero size (at $s = 0$) if, as can happen, the signal is weak and the number of background counts in the sample is somewhat below bT. These intervals are therefore unacceptable.

The strange behavior of these $b \neq 0$ frequentist estimates occurs for two reasons: the presence of irrelevant subclasses in the sample space, and the inability of frequentist methods to deal with cogent prior information of even the simplest kind. The sample space used to derive the interval includes hypothetical data sets in which the number of background counts is larger the the total number of background and signal counts in the actual data. This class of data is clearly irrelevant to the analysis of the actual data; considering it can cause the interval to include negative values of s. The prior information that $s \geq 0$ must then be used to truncate the region at $s = 0$; but though this leads to regions with the stated frequency of coverage, it causes the region to collapse to zero size when the uncorrected region lies completely below $s = 0$.

Only very recently has Zech (1989) found a frequentist procedure that avoids this behavior. He accomplishes this by making the sample space depend on the actually observed data: only hypothetical data sets with numbers of background counts $\leq N$ are considered. By a mathematical coincidence similar to that arising with the Gaussian distribution, his result gives intervals identical to Bayesian credible intervals calculated with equation (5.6) (Jaynes 1976 discusses a similar such coincidence with the Poisson distribution). In fact, Zech found the procedure by looking for a frequentist interpretation of the Bayesian result. Zech's procedure is essentially what a statistician would call 'conditioning on the value of an ancillary statistic,' the ancillary statistic here being N. Though this leads to an acceptable procedure in this case, ancillary statistics are not in general available in problems with recognizable subclasses. Finally, the use of ancillaries is somewhat inconsistent from the frequentist viewpoint, and thus controversial: if one accepts that inferences should be conditional on *some* feature of the data, why not condition completely on the data, as Bayesian procedures do?

5.2 Analysis of 'On/Off' Measurements

In our derivation of equation (5.6), we assumed that the background rate was known. More frequently, the background rate is itself measured by counting events in some time interval, and so is known imprecisely. Inferring the signal strength when the background is itself imprecisely measured is called an 'On/Off' signal measurement: one points the detector 'off source' to measure the background, and then 'on source' to measure the signal plus background. From these data we seek inferences about the signal strength alone, without reference to the background strength. Such inferences might be summarized as points with error bars on a plot of count rates versus time or energy, for example. Thus this is a problem with a nuisance parameter: the background rate.

The usual approach to this problem is to use the 'off' measurement to obtain an estimate of the background rate, $\hat{b}$, and its and standard deviation, σ_b, and to use the 'on' measurement to find an estimate of the signal plus background rate, $\hat{r}$, and its standard deviation, σ_r. The signal rate is then estimated by $\hat{s} = \hat{r} - \hat{b}$, with variance $\sigma_s^2 = \sigma_r^2 + \sigma_b^2$ (see, *e.g.*, Nicholson 1966). This procedure gives the correct result when applied to signals which can be either positive or negative, and for which the Gaussian distribution is appropriate. Thus it works well when the background and signal rates are both large enough so that the Poisson distribution is well-approximated by a Gaussian. But when either or both of 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 investigators believe 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).

Two frequentist alternatives to the above procedure have been proposed by gamma-ray astronomers (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, following Hearn (1969), both procedures interpret a likelihood ratio as the covering probability of a confidence region, and thus are not even correct frequentist procedures (Li and Ma 1983). Second, none of the procedures correctly accounts for the uncertainty in the background rate. 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.

More recently, Zhang and Ramsden (1990) have addressed this problem, using known results in the statistics literature to improve on the Gaussian approximation. The calculations involved are complicated and will not be described further here. But their confidence regions have a peculiar behavior that calls into question their reliability. When no background counts are observed over a long time interval, we become essentially certain that the background rate is zero. In this case, then, the 'On/Off' result should reduce to the well-known result for the measurement of a signal with no background (see, *e.g.*, Gehrels 1986). Instead, the Zhang and Ramsden interval collapses to zero size about $s = 0$, even when counts have been observed on-source: the procedure indicates certainty that $s = 0$ when we are certain that $s > 0$. This happens regardless of the value of $T_{\rm off}$.

The presence of prior information, the presence of a nuisance parameter, and the discrete/continuous character of the Poisson distribution all conspire to make this a difficult research problem for a frequentist. In contrast, the Bayesian solution to this problem is again a straightforward 'beginner' problem, as we now show.

First we consider the 'off' measurement. Suppose we count $N_{\rm off}$ events in a time $T_{\rm off}$ from an 'empty' part of the sky. These data lead to a posterior distribution for the background rate, b, of exactly the same form as equation (5.5):

$$p(b \mid N_{\rm off} I_b) = \frac{T(bT)^{N_{\rm off}} e^{-bT}}{N_{\rm off}!}. \tag{5.8}$$

Now consider the 'on' measurement. Suppose we count $N_{\rm on}$ events in a time $T_{\rm on}$ from measurements on source. This measurement provides us with information about both b and the source rate s. From Bayes' theorem, the joint posterior density for s and b is,

$$\begin{aligned} p(sb \mid N_{\rm on} I) &= p(sb \mid I) \frac{p(N_{\rm on} \mid sbI)}{p(N_{\rm on} \mid I)} \\ &= p(s \mid bI) p(b \mid I) \frac{p(N_{\rm on} \mid sbI)}{p(N_{\rm on} \mid I)}. \end{aligned} \tag{5.9}$$

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. We can express this symbolically by writing $I = N_{\rm off} I_b I_s$.

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

$$p(N_{\rm on} \mid sbI) = \frac{[(s+b)T_{\rm on}]^{N_{\rm on}} e^{-(s+b)T_{\rm on}}}{N_{\rm on}!}. \tag{5.10}$$

The prior for s, $p(s \mid bI)$, we will again take to be uniform,

$$p(s \mid bI) = 1. \tag{5.11}$$

To be rigorous, we should set a range and normalize this prior, and later consider the limit of large range, but the posterior we will find using this improper prior will be proper, so we need not go through the trouble of explicitly taking the limit. 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 $p(b \mid N_{\rm off} I_b)$, the posterior for b from the background estimation problem; it is given by equation (5.8). We can now calculate the joint posterior for s and b by normalizing the product of equations (5.7), (5.9), and the uniform prior for s. Ignoring the normalization for now, Bayes' theorem (equation (5.9)) gives the dependence of the joint posterior on the parameters as

$$p(sb \mid N_{\rm on} I) \propto (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})}. \tag{5.12}$$

To find the posterior density for the source strength, *independent of the background*, we just marginalize with respect to b, calculating $p(s \mid nI) = \int db\, p(sb \mid nI)$. Helene (1983) noted that the background uncertainty can be accounted for in this manner; but he only treated the case where the number of counts is large enough to justify a Gaussian approximation. The exact integral can be easily calculated after expanding the binomial, $(s+b)^{N_{\rm on}}$, in equation (5.12). The resulting normalized posterior (in the limit of large $s_{\rm max}$) is,

$$p(s \mid N_{\rm on} I) = \sum_{i=0}^{N_{\rm on}} C_i \frac{T_{\rm on}(sT_{\rm on})^i e^{-sT_{\rm on}}}{i!}, \tag{5.13}$$

with

$$C_i \equiv \frac{\left(1 + \frac{T_{\rm off}}{T_{\rm on}}\right)^i \ \frac{(N_{\rm on}+N_{\rm off}-i)!}{(N_{\rm on}-i)!}}{\sum_{j=0}^{N_{\rm on}} \left(1 + \frac{T_{\rm off}}{T_{\rm on}}\right)^j \ \frac{(N_{\rm on}+N_{\rm off}-j)!}{(N_{\rm on}-j)!}}. \tag{5.14}$$

Note that the denominator of C_i is simply the numerator summed over i, so that $\sum_{i=1}^{n} C_i = 1$.

This result is very appealing. Comparing it with equation (5.5), we see that Bayes' theorem estimates s by taking a weighted average of the posteriors one would obtain attributing 0, 1, 2,..., $N_{\rm on}$ events to the signal. The weights depend on $N_{\rm on}$, $T_{\rm on}$, $N_{\rm off}$, and $T_{\rm off}$ so that the emphasis is placed on a weak signal or a strong signal, depending on how $N_{\rm on}/T_{\rm on}$ compares with $N_{\rm off}/T_{\rm off}$.

The form of equation (5.13) suggests that C_i is the probability that i of the events observed on-source are from the source, taking into account the information about the background provided by the off-source measurement. In fact, this probability can be calculated directly. Given the posterior distribution for the background rate, equation (5.8), the *posterior predictive distribution* that n' background events will be observed in an interval $T_{\rm on}$ can be calculated according to

$$p(n' \mid I_b) = \int_0^\infty p(n' \mid bI_b) p(b \mid I_b) db, \tag{5.15}$$

where $p(n' \mid bI_b)$ is given by the Poisson distribution with expectation $bT_{\rm on}$, and $p(b \mid I_b)$ is given by equation (5.8). The reader may verify that assigning $n' = n - i$ events to the background with probability given by equation (5.15) leads to an alternate derivation of equation (5.13) that explicitly identifies the C_i with the probability that $n - i$ events are background events. This demonstrates both the consistency and the 'sophisticated subtlety' of Bayesian inference; in fact, requirement of this kind of consistency plays an important role in the foundation of the theory (see, *e.g.*, Loredo 1990). Application of this result to actual data, including a comparison with frequentist results, will be presented elsewhere.

5.3 Poisson Spectrum/Shape Analysis

So far we have considered the simplest possible Poisson problems: inferring the value of a constant rate. It is astonishing that these problems are so difficult from the frequentist point of view that no satisfactory frequentist solution to the simple problem of analyzing on/off measurements has yet appeared in the astrophysical literature, and that in fact several demonstrably unsatisfactory procedures have been presented. Regardless of whether a fully satisfactory procedure exists elsewhere in the frequentist literature, it is troubling that skilled scientists have been thwarted in finding the solution to such apparently simple problems as those just discussed. In contrast, we have easily found the Bayesian solutions to these problems.

Now we will briefly consider a class of more complicated problems: inferring characteristics of a time-varying rate function, $r(t;\theta)$ from the arrival times, t_i, of N events detected in a time interval T. Astronomers frequently focus attention on periodic rate functions, with one of the model parameters being a frequency, ω. Two commonly considered problems astronomers address with such data are, (1) the detection problem: is there evidence that a periodic signal is present, and (2) the estimation problem: what is the frequency. Bayesian and frequentist methods for addressing these problems differ greatly.

5.3.1 *Frequentist Analyses*

The frequentist procedures used for these problems calculate the value of some statistic, $S(\omega)$, at some finite set of frequencies, ω_i ($i = 1$ to N_ω). The detection problem is addressed by calculating the distribution for S assuming that the signal is constant; the location of the maximum observed value of $S(\omega_i)$ in this distribution is then used to decide whether the assumption of a uniform rate should be rejected. If there is evidence for a periodic signal, the estimation problem is addressed using the functional dependence of $S(\omega)$ on ω.

Three choices for S dominate current frequentist analyses of arrival time data by astronomers (see, *e.g.*, Leahy *et al.* 1983; and Leahy, Elsner, and Weisskopf 1983), though several other statistics have been used as well. One choice requires binning the data into uniformly spaced bins containing several events each; the power spectrum of the binned time series, calculated using the discrete Fourier transform (DFT), is used as the statistic. A 'χ^2' statistic is used in the 'epoch folding' (EF) approach. Here one folds the arrival times modulo some trial period to produce a phase, ϕ_i, for each event in the interval $[0, 2\pi]$. The phases are then binned, and the χ^2 statistic comparing the resulting histogram with a uniform distribution is calculated. The third choice is the Rayleigh statistic, R^2, used in the Rayleigh test (RT) for uniformity on a circle (Mardia 1972). R^2 is calculated by folding the arrival times modulo a trial period, placing N unit vectors in the (r, ϕ) plane at the resulting phases ϕ_i, and calculating the mean squared length of the resultant vector,

$$\begin{aligned} R^2 &= \frac{1}{N}\left[\left(\sum_{i=1}^{N} \sin\phi_i\right)^2 + \left(\sum_{i=1}^{N} \cos\phi_i\right)^2\right] \\ &= 2 + \frac{4}{N}\sum_{i\neq j} \cos(\phi_j - \phi_i) \end{aligned} \tag{5.16}$$

R^2 is proportional to the Fourier power in the time series at the trial frequency; generalizations of the RT are also used that add the powers at various harmonics (Protheroe 1987).

Several choices the investigator must make complicate the interpretation of procedures based on these statistics, particularly when they are used for signal detection. Both the binned DFT and EF methods require a choice of bin size and a choice of initial phase (*i.e.*, the location of the bin boundaries). Also, if such methods are used to search for an unknown frequency (as opposed to being used to detect pulsation at a frequency known *a priori* from other observations), the number of frequencies searched, N_ω, must be carefully considered in the assessment of the significance of a result.

These difficulties arise because the frequentist analysis does not analyze the data set as if it were a single sample from an ensemble of sets of arrival times. Instead, it considers *folded versions* of the single data set to be samples from an ensemble of sets of phases (for the RT and EF methods) or sets of binned phases (for the binned DFT method): the sample space is the space of phases or binned phases, not the space of arrival times. This complicates the analysis because each folded version of the data set is derived from *one* set of observed arrival times, and thus difficult questions of fairness and independence arise. For example, if we find a '4σ' bump at frequency ω in the DFT, χ^2 function, or Rayleigh statistic, we would consider this significant if we examined only a few other frequencies, but hardly surprising if we examined 10^6 'well-separated' frequencies. Thus the implications of the value of $S(\omega)$ for the existence of a signal depend not only on the value itself, but on the number of other values that have been examined. This number cannot be chosen arbitrarily. If it is too large, we may not be able to consider each of the values of the chosen statistic to be independent samples from a uniform phase distribution; indeed, if two of the examined frequencies are close enough, the values of the statistic could be identical. Also, the actual values of the examined frequencies must be specified without consideration of the data, for if the frequencies are chosen to ensure that $S(\omega)$ is maximized at one of the ω_i, the resulting value of S is not a fair sample (Leahy *et al.* 1983), and the significance of the signal will be overestimated. Similarly, if the initial phase for the binned DFT and EF methods is chosen so that the resulting value of S is maximized, the resulting set of binned phases is not a 'fair' sample from the space of random phases, a point whose importance for time series analysis has only recently been appreciated by astronomers (Protheroe 1985; Collura *et al.* 1987), though Press and Schecter (1974) and Hillas (1975) earlier noted the importance of this affect in other astrophysical problems.

The need to carefully specify the set of frequencies or phases to be examined is the spectrum analysis analogue to the stopping rule problem discussed in Section 3. More fundamentally, these difficulties arise because of the presence of nuisance parameters. In estimating the frequency of a periodic signal, all the other parameters required to describe the signal (such as its phase and amplitude) are nuisance parameters. In assessing the evidence that a periodic signal is present at all, the frequency itself becomes a nuisance parameter. Unable to deal with nuisance parameters in the hypothesis space, the frequentist methods just described redefine what is to be considered the data so that nuisance parameters can be accounted for to some extent by adding additional structure to the sample space. But such structure so complicates the analysis that the significance of detections is often incorrectly assessed. Analyses of arrival time data are thus often greeted with suspicion, and have sometimes been the subject of vehement debate.

5.3.2 *Bayesian Analyses*

Bayesian analyses of such data proceed very differently. The detection problem is addressed by assuming the data can be described by one member of a class of model rate functions, including among them a constant rate model, and calculating the probabilities of each model; there is no significant evidence for a periodic signal if the probability of the constant model is large. The estimation problem is addressed by calculating the full posterior distribution for the parameters of a model, and then integrating away all the model parameters except ω, producing a marginal distribution for the frequency.

The Bayesian calculations are much more straightforward than their frequentist counterparts, in part because the sample space is the space of arrival times, not the space of phases or binned phases, but largely because of Bayesian facility in dealing with nuisance parameters, which alleviates the need to modify the sample space to attempt to account for nuisance parameters. As a result, the statistics arising in the Bayesian calculation can be evaluated at any number of frequencies; questions of independence are eliminated because alternative hypotheses are dealt with in the hypothesis space, not in the sample space. We might summarize the distinction between the two methods as follows: frequentist period searches *maximize* and then correct the result for the amount of parameter space searched; Bayesian period searches *average* over the parameter space.

One apparent drawback of the Bayesian approach is the need to assume specific models for the shape of the periodic signal in order to address the detection problem. This contrasts sharply with frequentist methods that seek to reject a uniform model rather than choose between a uniform model and periodic models. However, we will see that each statistic used in frequentist procedures arises naturally in a Bayesian calculation that assumes the signal is of a very specific form. In this sense these apparently alternative-free procedures implicitly assume very specific classes of alternatives. This is recognized in thorough frequentist comparisons of the various statistics, where the *power* of a statistical test—a measure of its ability to correctly identify a signal from among two alternatives—is calculated and used to choose from among competing statistics. Unfortunately, few studies by astronomers consider the power of a test (Leahy, Elsner and Weisskopf 1983; Protheroe 1987; and Buccheri and DeJager 1989 are notable exceptions).

The key ingredient in the Bayesian analysis of arrival time data is the likelihood function. We now describe how the likelihood function can be constructed for any desired periodic rate function. Then we will briefly note how specific choices of this function lead us to consider the same statistics used in some frequentist procedures, but to use them in different ways.

The likelihood function for arrival time data can be built from the Poisson distribution as follows. Divide the observing interval into many small intervals of size Δt; we will ultimately consider the limit in which these intervals become infinitesimal, but finite intervals could represent the precision of the clock recording the arrival times. From the Poisson distribution, the probability that no event will be detected in an interval Δt about time t is,

$$P_0(t) = e^{-r(t)\Delta t}. \tag{5.17}$$

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

$$P_1(t) = r(t)\Delta t \; e^{-r(t)\Delta t}. \tag{5.18}$$

We will assume that the intervals are small enough that no more than one event is observed in any interval.

The likelihood function, $\mathcal{L}(\theta) \equiv p(\{t_i\} \mid \theta I)$, is the product of the probabilities of detecting each of the observed events, times the product over all intervals not containing an event of the probability of no detection. That is,

$$\mathcal{L}(\theta) = \left[\prod_{i=1}^{N} P_1(t_i)\right] \prod_j P_0(t_j), \tag{5.19}$$

where j runs over all intervals not containing an event. From the definitions of P_0 and P_1 it follows that

$$\mathcal{L}(\theta) = \left[\prod_{i=1}^{N} r(t_i)\Delta t\right] \; e^{-\sum_j r(t_j)\Delta t}, \tag{5.20}$$

where j now runs over *all* intervals. As noted following equation (5.1), $r\Delta t$ is really a shorthand for the integral of $r(t)$ over the interval Δt. Thus the sum in the exponential is equal to the integral of $r(t)$ over the observed interval and the likelihood becomes,

$$\mathcal{L}(\theta) = \Delta t^N \; e^{-\int_T r(t)dt} \; \prod_{i=1}^{N} r(t_i). \tag{5.21}$$

Combined with prior probability distributions for the parameters, this likelihood function yields a posterior distribution for the rate function parameters, θ. This likelihood function was studied by Cleveland (1983) for frequentist analyses of solar neutrino data; several investigators used it for analyzing the neutrinos observed from supernova SN 1987A (see Loredo and Lamb 1989 for a review).

To proceed further, we must specify parametrized models for $r(t)$ and priors for the parameters. In this limited space we can only briefly indicate the results of a few simple choices and their relationship to frequentist statistics. We will focus on estimation problems (*e.g.*, estimating the unknown frequency of an assumed pulsation), only briefly discussing the equally important and logically prior detection problem (deciding whether or not a periodic signal is present) which we discuss in greater detail elsewhere (Gregory and Loredo 1992).

If a constant model is studied, $r(t) = r$, the likelihood function becomes $(r\Delta t)^N e^{-rT}$. In its dependence on r, this is proportional to the likelihood function studied earlier, equation (5.1). A uniform prior density for r thus leads to a posterior density for r of the form of equation (5.5), as we should expect.

Moving on to periodic models, perhaps the simplest such model one might study is one with a sinusoidal variation. Noting that $r(t)$ must be everywhere positive, we write the rate function as

$$r(t) = A\left[1 + f\sin(\omega t + \phi)\right], \tag{5.22}$$

where A is the time averaged rate, f is the pulsed fraction in the interval $[0,1]$, ω is the (angular) frequency, and ϕ is the phase. With this choice, the likelihood function takes the form,

$$\mathcal{L}(\omega, A, f, \phi) = (A\Delta t)^N e^{-AT} \prod_{i=1}^{N} \left[1 + f\sin(\omega t_i + \phi)\right], \tag{5.23}$$

where here and throughout this Section we approximate the integrated rate in the exponential by the duration times the time average, AT. If we assign uniform priors to all the parameters (those for f and ϕ being bounded between 0 and 1, and 0 and 2π, respectively), the joint posterior distribution for the parameters is $p(\omega, A, f, \phi \mid DI) \propto \mathcal{L}(\omega, A, f, \phi)$. The marginal distribution for ω can be found analytically by integrating out the other parameters; it has the form,

$$\begin{aligned} p(\omega \mid DI) = \; & C_1 \quad + \\ & C_2\left[\cos\omega(t_2 - t_1) + \cos\omega(t_3 - t_1) + \ldots\right] \quad + \\ & C_3\left[\cos\omega(t_2 - t_1)\cos\omega(t_4 - t_3) + \ldots\right] \quad + \\ & C_4\left[\cos\omega(t_2 - t_1)\cos\omega(t_4 - t_3)\cos\omega(t_6 - t_5) + \ldots\right] \quad + \ldots, \end{aligned} \tag{5.24}$$

where the C_i are constants. Crudely, the marginal distribution counts the numbers of pairs of events separated by an integral number of periods, and the numbers of distinct *pairs of pairs* of events so separated, and so on. This is to be compared with the Rayleigh statistic, equation (5.16), which counts only pairs of events. In fact, the leading ω-dependent term in $p(\omega)$ *is* the Rayleigh statistic.

When the shape of the signal is not known to be sinusoidal, it is necessary to consider more complicated signal models. One possibility is a sum of harmonically related sinusoids, with different amplitudes and phases. As was done above, the amplitudes and phases can be integrated out, leaving a marginal distribution for ω. Unfortunately, numerical evaluation of the marginal distribution for such models is prohibitively computationally expensive. But other models can be considered that combine computational simplicity with usefully general lightcurve shapes. We will discuss two choices that are related to the frequentist statistics discussed above.

First, consider a rate function that has the form of an *exponentiated* sinusoid,

$$r(t) = \frac{A}{2\pi I_0(\kappa)} e^{\kappa \cos(\omega t+\phi)}. \tag{5.25}$$

This function is proportional to the vonMises distribution, a circular generalization of the Gaussian distribution (Mardia 1972). This rate function has one peak per period, with a location determined by ϕ and a width determined by κ. Thus it is useful when the lightcurve is expected to have one peak, but with unknown width or 'duty cycle.' The Bessel function, $I_0(\kappa)$, appears so that the parameter A is the time-averaged rate. When $\kappa = 0$, the rate is constant.

Because this function is the exponential of a sinusoid, the likelihood function has a convenient form:

$$\mathcal{L}(\omega, A, \kappa, \phi) = \left[\frac{A\Delta t}{I_0(\kappa)}\right]^N e^{-AT} e^{\kappa \sum_i \cos(\omega t_i+\phi)}. \tag{5.26}$$

With uniform priors, the phase and amplitude can be integrated out to yield a marginal distribution for ω and κ of the form,

$$p(\omega, \kappa \mid DI) = C \frac{I_0\left[\kappa N R(\omega)\right]}{I_0^N(\kappa)}, \tag{5.27}$$

where C is a normalization constant, and $R(\omega)$ is the square root of the Rayleigh statistic, equation (5.16). A marginal distribution for ω can be found by integrating this with respect to κ numerically.

The data enter this marginal posterior only through $R(\omega)$. Thus from a Bayesian point of view, the Rayleigh statistic exhausts the information about periodicity when the rate function is of the form of equation (5.25). In the frequentist literature, the Rayleigh statistic has been criticized because it has been found to be insensitive to signals of narrow width (Leahy, Elsner, and Weisskopf 1983). In the context of the vonMises model, however, these frequentist analyses assume that $\kappa = 1$. Allowing κ to vary freely may alleviate this weakness.

For signals that may have more than one peak per period, we need a still more general model. One possibility is the exponential of the sum of two or more harmonically related sinusoids. When amplitudes and phases are marginalized, the resulting statistic is closely related to the Z_n^2 generalizations of the Rayleigh statistic (Buccheri and DeJager 1989 review the Z_n^2 statistics). But here we will consider a different, simpler model (Gregory and Loredo 1992). We approximate the lightcurve with a piecewise constant rate function with M pieces, each covering an equal fraction of the period. Then we can write the rate function as

$$r(t) = AMf_j, \quad \text{with} \quad j(t) = \mathrm{int}\left[1 + M[(\omega t+\phi) \mathrm{mod} 2\pi]/2\pi\right]. \tag{5.28}$$

Here j identifies in which of the M pieces the time t falls. The parameter A is again the time-averaged rate, and the M parameters f_j are the fractions of the rate in each of the M bins. These parameters must be in the range $[0,1]$, and only $M-1$ of them are really free, since $\sum_j f_j = 1$. Thus A parametrizes the amplitude of the rate, and the f_j parametrize the shape of the lightcurve.

In terms of this piecewise constant rate, the likelihood is,

$$\mathcal{L}(\omega, A, \phi, \{f_i\}) = M^N (A\Delta t)^N e^{-AT} \prod_{j=1}^{M} f_j^{n_j}, \tag{5.29}$$

where $n_j = n_j(\omega, \phi)$ is the number of events that lie in piece j of the lightcurve, given the phase and frequency. These numbers correspond to the number of events that lie in bin j in the EF method.

This piecewise constant model covers a wide variety of shapes, but this variety comes with a cost: there are lots of parameters. However, the shape parameters can be integrated out analytically using the generalized Beta integral. Using uniform priors, the resulting marginal distribution for the frequency and phase is,

$$p(\omega, \phi \mid DI) = C \frac{M^N \ (M-1)!}{(N+M-1)!} \left[\frac{n_1! \, n_2! \, \ldots n_M!}{N!} \right]. \tag{5.30}$$

This distribution takes into account information about *all* of the wide variety of shapes parametrized by the f_j. Again, C is a normalization constant, and only the term in brackets depends on ω and ϕ. This term is just the reciprocal of the multiplicity of the set of n_j values—the number of ways N events can be distributed in M bins with n_j events in each bin—also called the configurational entropy of the n_j. This multiplicity is largest, and the marginal density is thus smallest, when the n_j are all equal. Thus there is strong evidence of a period only if the resulting set of n_j values is not uniform. In fact, Gregory originally proposed studying the multiplicity as a statistic because of this intuitively appealing behavior (Gregory and Loredo 1992). The Bayesian calculation just described illuminates the assumptions leading to this statistic, and specifies precisely how to use it to make probability statements about the signal.

The behavior of the multiplicity as a measure of nonuniformity is remarkably similar to the idea behind the frequentist EF method. In fact, using Stirling's approximation, one can show that

$$\log p(\omega, \phi \mid DI) \approx \frac{1}{2}\chi^2 + \frac{1}{2}\sum_j \log n_j + C(M), \tag{5.31}$$

where $C(M)$ is a constant depending on M, and χ^2 is the same statistic used in the EF method. Thus we see that, to terms of order $\log n_j$, the χ^2 statistic exhausts the information in the data when a piecewise constant model is assumed.

This calculation provides a Bayesian interpretation of the χ^2 statistic that leads us to use it in a manner very different from that in the EF method. First, as we noted earlier, the choice of phase affects the interpretation of the χ^2 statistic in the EF method. This led Collura *et al.* (1987) to suggest using χ^2 averaged over phase to eliminate this subjective aspect of the EF method. But equation (5.31) reveals that the proper way to eliminate the phase from consideration is to average the *exponential* of $\chi^2/2$ over phase, not χ^2 itself.

Second, equation (5.31) shows how to use χ^2 to estimate the frequency of the signal: the exponential of $\chi^2/2$, averaged over ϕ, is the marginal density for ω, and integrals and moments of this function can be straightforwardly used to estimate the uncertainty with which the frequency is determined.

Finally, we have so far discussed only the estimation problem of inferring ω, assuming a periodic signal is present. Bayesian methods can be easily developed for addressing the detection problem as well. For models like the sinusoid model, equation (5.22), the simplest way to address the detection problem is to simply estimate the amplitude of the pulsed part of the signal by calculating the marginal distribution for the pulsed fraction, f. If $f = 0$ lies outside, say, the 95% credible region for f, there is significant evidence for a signal. More rigorously, though, one must perform a Bayesian model comparison calculation, calculating the probability that the data are from a constant model, or one of the other models we have discussed. There is not enough space here to describe such calculations (see Gregory and Loredo 1992), but they are straightforward. They lead to results that depend on the range of frequencies examined, but *not on the number of frequencies examined* in that range, eliminating this subjective aspect of all frequentist tests for periodicity. Further, for the piecewise constant model, model comparison calculations can be used to determine not only if there is evidence for a periodic signal, but also the number of bins needed to model the signal shape, using the M-dependent terms in equation (5.30). Essentially, M becomes a parameter of the model, to be estimated from the data like the other parameters. No frequentist method for doing this has yet been offered.

We have only briefly outlined part of the theory of Bayesian spectrum/shape analysis of event location data. Preliminary investigations of the application of such methods to simulated and real data show these methods to have great promise. Further development and study of these methods will be presented elsewhere.

6. ASSIGNING DIRECT PROBABILITIES: PRIORS AND LIKELIHOODS

We have demonstrated that application of the sum and product rules to probabilities of hypotheses straightforwardly leads to procedures that perform as well as, and often better than, their frequentist counterparts. In our calculations, however, we have taken a rather cavalier attitude toward priors. The presence of priors in Bayesian calculations, and the historical lack of compelling assignments for priors, has led many to assert that Bayesian inference is too subjective for use by scientists. Many, too, believe that priors are the primary element distinguishing Bayesian and frequentist methods, so that Bayesian methods are only relevant when there is strong prior information. These beliefs have prevented many scientists from even considering the application of Bayesian methods to their statistical problems. As a result, Bayesian methods have been dismissed without examining their performance.

I have deferred a discussion of priors to this late Section for two reasons. First, I want to emphasize that priors are far from the only distinguishing feature of Bayesian inference; Bayesian inference differs from frequentist statistics in a much more fundamental way that drastically affects the calculations one must perform to address a problem regardless of whether important prior information is available. Second, a proper discussion of prior probabilities *must* raise conceptual issues that I wanted to avoid until I demonstrated the pragmatic superiority of the Bayesian approach, and the serious shortcomings of the frequentist approach. Independent of all philosophical argument, it is a *fact* that frequentist methods suffer from serious problems and inconsistencies, and that Bayesian methods avoid these problems. With this purely pragmatic motivation, we now study the problem of assigning priors. For it is a further fact that logical consistency requires Bayesian calculations to use priors.

6.1 Priors: One Type of Direct Probability

We begin our discussion by noting the broader context in which the problem of assigning priors appears in Bayesian inference. The sum and product rules that we have used so frequently in this paper tell us how to combine known probabilities to find other, related, probabilities. But before they can give us the numerical values we require in any practical application of the theory, they require as 'input' the numerical values of those probabilities from which others are calculated. Thus, in a sense, the sum and product rules are only half of probability theory, the missing half being the rules that specify how some information I about a proposition A directly leads to a numerical value for the probability $p(A \mid I)$.

Probabilities that are assigned directly are called *direct probabilities*. In principle, any probability could be a direct probability. We might, for example, search for rules that allow us to directly assign a numerical value to a posterior probability, $p(H \mid DI)$, and so avoid using Bayes' theorem, priors, and likelihoods. In practice, it has proved more straightforward to use Bayes' theorem, so the most studied direct probabilities are priors and likelihoods.

Thus the first point we want to emphasize about priors is that both priors *and likelihoods* are direct probabilities. The same kind of logical analysis will be needed to justify assignments of both kinds of probabilities in the formal development of probability theory. On the other hand, we reserve the same freedom to use scientific judgement, informed by experience, to assign approximate priors that qualitatively express information in practical calculations, just as we use judgement to assign approximate likelihoods, both in Bayesian and frequentist calculations. The next two subsections discuss some of the formal apparatus available for assigning direct probabilities. But then we discuss some guidelines to help us decide if and when much effort should be spent on formality and rigor in practical calculations.

6.2 Abstract Probability

Much of the confusion and uneasiness felt toward priors results from the failure to recognize how radically different the Bayesian concept of probability is from the frequentist concept. Superficially, Bayesian calculations appear similar to their frequentist counterparts, and it is tempting to merely plow ahead and calculate as before, only allowing more freedom regarding the arguments of probability symbols. But on a deeper level Bayesian calculations only make sense if we drastically change the way we think about probability.

Frequentist statistics identifies probability with frequency—an empirical concept—and thus seems almost to be a physical theory. It gives randomness and probability the character of properties of nature. Bayesian probability theory is more abstract. Bayesian probabilities describe a state of knowledge specified by the information placed to the right of the bar in a probability symbol. No reference is made to frequency, repetition, randomness, or any empirical phenomenon. Of course, one is free to study probabilities of propositions referring to frequencies or other observable phenomena, as is frequently done in practice; this will lead to derivable mathematical relationships between probabilities and frequencies (see, *e.g.*, Jaynes 1978). But probabilities themselves are not empirical.

Frequentist reference to properties of nature makes frequentist statistics appear more objective than Bayesian inference. This objectivity is illusory, however. The frequencies required by the theory are those from an infinite number of 'identical' repetitions of an experiment. Such frequencies are never available. Thus frequentists must confront a problem similar to that Bayesians encounter: the assignment of frequencies from incomplete information. But this differs from the Bayesian problem in that there is one 'true' frequency distribution realized in nature; the frequentist is not interested in describing incomplete information, but instead needs to identify the 'true' distribution despite incomplete information.

For a Bayesian, there is no such thing as an empirically meaningful 'true' distribution for anything, neither for hypotheses nor for data. Probability distributions always describe an incomplete state of knowledge, not properties of an hypothetical infinite population. This is true even for situations where we may feel comfortable thinking about randomness and frequencies as objective properties of a system. In most (possibly all) such situations, 'randomness' (*i.e.*, unpredictability) is not a physical property; it is a consequence of lacking the information necessary to predict outcomes with certainty. For example, the randomness of coin flipping does not refer to a property of coins; it is a consequence of incomplete information regarding the properties of the coin being flipped and the precise conditions of the flip, information that, if available, would enable us to predict outcomes with certainty.

In particular, a Bayesian prior distribution for a parameter, $p(\theta \mid I)$, does not refer to a population of experiments or worlds, each with different values of θ. It simply describes what the information I tells us about the various possible values θ might take in the one case at hand. There is no 'true' prior, realized in nature as a frequency distribution that we must discover (as is assumed by Annis, Cheston, and Primakoff 1953, for example). The prior expresses what we know or are willing to assume about the single case at hand.

These considerations are as true of likelihood functions as they are of priors. In particular, Bayesian inference is not concerned with the identification of 'true' models or parameter values. Rather, models and their parameters are viewed as simplified descriptions of a phenomenon in the context of which we describe past data and predict future data (West and Harrison 1989; Hestenes 1989). Bayesian probability theory is the mathematical language used for such description. The sum and product rules are the 'grammar' of this language, and the direct probabilities, specified by I, are its 'vocabulary.' When it occurs, the word 'true' has meaning only in the context of the information I specifying the problem, and not in any absolute sense. Our notation notes this explicitly: following Jeffreys (1939), all our probability symbols explicitly show the dependence of the results on the information assumed.

The distinction between Bayesian and frequentist notions of probability is reflected in an interesting way in the language used to refer to distributions (Jaynes 1986). Frequentists speak of $p(x)$ as the distribution *of* the quantity x: the quantity that is 'distributed' in a frequentist distribution is the argument, which takes on various values with a frequency distribution $p(x)$. Bayesians speak of $p(x \mid I)$ as a probability distribution *for* the quantity x: the quantity that is 'distributed' is the probability, the plausibility assigned to various possible values of x. The symbol x may refer to an unknown constant, incapable of taking on any but one value, but if that value is unknown, we distribute plausibility among the various possible values according to $p(x \mid I)$.

A useful analogy can be made between probability, a numerical encoding of the qualitative notion of plausibility, and temperature, a numerical encoding of the qualitative notion of hot and cold (Jaynes 1993). The Kelvin temperature, T, which appears in thermodynamics, is an abstract quantity: we do not naturally assign the number '273' to the coldness of ice! But this abstract mathematical quantity is required to develop a quantitative theory of heat. Similarly, Bayesian probability is an abstract, primitive notion required if we want to reason consistently and with mathematical precision in the presence of uncertainty (Jaynes 1990). For both probability and temperature, we can relate the abstract concept to the intuitive one by application to special cases. Thermodynamics assigns the temperature $T = 273$K to ice at the melting point, so if the temperature of something is 273K, we know it is as cold as ice. Similarly, probability theory assigns a probability of 1/36 to the "snake's eyes" outcome of a fair roll of a pair of dice, so an hypothesis with a probability of 1/36 is as plausible as "snake's eyes." In the case of probability theory, relating mathematics to intuition may involve consideration of frequencies, though more often our intuition about probability derives, not from observations of frequencies, but from counting the number of outcomes that seem equally plausible *a priori*.

Part of the value of the abstract mathematical theories is that their abstraction, which disconnects them from intuition, allows us to apply them to situations far beyond those accessible to our intuition. Temperatures of 10^9 degrees Kelvin are intuitively meaningless, but very meaningful to a physicist; a probability of 10^{-9} is similarly outside the range of human experience, but can be very meaningful to a scientist.

6.3 From Information to Probability

The thermal analogy breaks down when we note that temperature is a physical property whereas probability describes a state of knowledge. Temperatures can be measured, but probabilities must instead be *assigned*. The Bayesian counterpart to a thermometer must be developed; it will not be an instrument, but rather a collection of rules for converting different kinds of information to probability assignments. These rules cannot be arbitrary, for though Bayesian distributions are not empirically verifiable frequency distributions, neither are they arbitrary descriptions of the opinions or whims of an individual. The rules must ensure that the assigned probabilities satisfy the consistency requirements that underly the sum and product rules in the foundations of the theory.

Only recently have statisticians recognized that the problem of assigning probabilities is fully half of probability theory, and requires tools beyond the sum and product rules for its solution. As a result, this part of the theory is not yet fully developed. Indeed, it is unlikely that it will ever be complete, since there is probably no end to the kinds of information one may have about propositions. At present, we can only convert very limited kinds of information into probability assignments; we are not yet at the point where we can consider background propositions like, 'everything expert X knows about A.' Instead, we can only consider simple caricatures of the full information a scientist may have about a phenomenon. But this limited capability is already sufficient to duplicate all of the successes of frequentist theory, and to move well beyond frequentist capabilities in some problems.

In fact, it is often the case that we want the data to 'speak for themselves,' and thus seek a prior that expresses ignorance, not one which expresses expert knowledge. Much of the existing literature focuses on such 'uninformative' priors; we will review some of this work here.

The first point to emphasize is that we are never in a state of complete ignorance about a parameter. We always know *something* about it. In particular, since the I that appears in the prior, $p(\theta \mid I)$, is the same as that appearing in the likelihood, $\mathcal{L}(\theta) = p(D \mid \theta I)$, we at least know the role θ plays in the likelihood function. Our task is to assign priors to parameters, not to greek letters; the parameters have some meaning in the context of the model we are studying, and this meaning is an important piece of information that must guide our probability assignment (Jaynes 1968; Lindley 1990). In actuality, then, there is no such thing as an 'uninformative' probability assignment; what we seek is an assignment that is in some sense 'least informative,' expressing as little beyond mere specification of the meaning of the parameter as possible. The background information, I, must specify precisely the nature of the parameter and what we mean by being ignorant of its value in enough detail to make the problem of assigning direct probabilities mathematically well-posed.

There is so much confusion over the issue of 'uninformative' priors that it is perhaps worth taking the space to put this another way. If we really want the data to 'speak for themselves,' all we can do is present the data. Once we introduce a model and parameters into a problem, we have already used information that is not in the data by themselves, and to be consistent, we must use this information throughout our analysis. Further, since we are performing a mathematical analysis, we must state this information in a mathematically precise manner, even though such a mathematically precise specification may be merely a caricature of our human information. Our problem, then, is not to find some magic function that is a uniquely correct specification of what we mean by 'uninformative,' but rather, to *define what we mean by 'uninformative'* with enough precision to allow unambiguous calculation of the prior corresponding to the chosen definition. Since, as we've already noted, mathematical models are only caricatures of reality, it should come as no surprise that there are several useful definitions of what one may mean by 'uninformative,' and thus several methods for finding uninformative or least informative priors. If we find that our results depend sensitively on which definition we choose, then we need to very carefully consider the relationship between our real-world knowledge and our mathematical definitions. But more often than not, our results will not depend that sensitively on which definition we use, as we note in section 6.4 below.

We will discuss here two methods for finding least informative assignments from information that may at first seem too vague to allow precise mathematical description. Each method will be appropriate when we have a particular kind of information about a model. First we will discuss the *group invariance method* developed by Jaynes (1968, 1973, 1980, 1993). This method is often appropriate when the parameters have a physical interpretation that allows us to identify two mathematical descriptions of a situation as being equivalent. Then we will discuss a *predictive method* that may be appropriate when the parameters do not have an obvious physical meaning, but are instead primarily useful for summarizing or predicting data, as when we fit data to a straight line (*e.g.*, in the Tully-Fisher or Faber-Jackson relations). This was the method used by Bayes in his famous paper introducing a special case of what is today known as Bayes' theorem (Stigler 1982); some recent applications are reviewed by Geisser (1988).

The key to understanding the group invariance method is to be careful to distinguish between an actual problem and its mathematical representation. Suppose two investigators with the same information, I, about a phenomenon analyze the same data set, but choose to parametrize their models differently. Investigator A chooses θ as the parameter, and investigator B chooses $\phi = \phi(\theta)$. The actual problems these investigators address are identical, but the mathematical problems they face—the conversion of the information I to direct probability assignments for the symbols θ and ϕ—differ because of the different ways each have chosen to label the same actual hypotheses. Thus the functional forms of their priors will in general differ; A will assign a prior density with functional form $p(\theta < \theta_{\rm true} < \theta + d\theta \mid I) = f(\theta)d\theta$, and B will assign a different function, $p(\phi < \phi_{\rm true} < \phi + d\phi \mid I) = g(\phi)d\phi$. But since the two investigators are addressing the same problem, we demand that the probability A assigns to a region of θ be equal to the probability B assigns to the corresponding region of ϕ. This leads to the *transformation equation*,

$$f(\theta)d\theta = g[\phi(\theta)]d\phi(\theta). \tag{6.1}$$

This equation simply states that the two different mathematical problems describe the same actual problem. It must be true for any choices of θ and ϕ.

Now suppose the information I identifies particular choices of θ and ϕ that make the *mathematical* problem of assigning $p(\theta < \theta_{\rm true} < \theta + d\theta \mid I)$ equivalent to that of assigning $p(\phi < \phi_{\rm true} < \phi + d\phi \mid I)$; that is, the manner in which I distinguishes between different values of θ is identical to the manner in which it distinguishes between values of ϕ. Symmetries of the likelihood function may help identify such mathematically equivalent parametrizations. For such parameter choices, the actual functions A and B assign must be the same. This leads to the *symmetry equation*,

$$f(x) = g(x), \tag{6.2}$$

for all values of the argument, x. This equation can only be true for certain choices of θ and $\phi = \phi(\theta)$ identified by I.

Combining the transformation and symmetry equations leads to a functional equation (an equation whose solution is a function) for f (or g),

$$f(\theta) = f[\phi(\theta)]\frac{d\phi(\theta)}{d\theta}. \tag{6.3}$$

Such a functional equation can be solved for the form of $f(\theta)$ (see Aczel 1966 for some methods), thus identifying the prior expressing the information I.

Perhaps the simplest possible example of an assignment resulting from group invariance is Laplace's *Principle of Indifference* (PI) for assigning a least informative probability distribution to a exhaustive set of N discrete, exclusive hypotheses, H_i (one, and only one, of the H_i is true). Laplace suggested that the distribution expressing 'complete ignorance' about such hypotheses assigns them each the same probability: $p(H_i \mid I) = 1/N$. Jaynes (1993) presents a careful discussion of the PI, deriving it from group invariance. We can illustrate the principles most simply by considering the case $N = 2$.

Let the information I specify only that there are two hypotheses, and that they are exclusive. Let investigator A label the hypotheses A_1 and A_2, and let B label them B_2 and B_1, respectively (B reverses the numbering of the index). Investigator A might express the information symbolically as the proposition $I = (A_1 \oplus A_2)$: 'Either A_1 or A_2 is true' (we here use '$\oplus$' to indicate exclusive 'or'). Similarly, B would write $I = (B_1 \oplus B_2)$.

Write $p(A_i \mid I) = f_i$ for A, and $p(B_i \mid I) = g_i$ for B. Since A and B are in the same state of information, we demand that they assign the same probabilities to the same actual hypotheses; thus the transformation equations for this problem read,

$$f_1 = g_2 \quad \text{and} \quad f_2 = g_1. \tag{6.4}$$

Now note that I distinguishes between the symbols used for the hypotheses in the same way, so that the mathematical problems A and B face are symbolically equivalent: $I = (A_1 \oplus A_2)$ distinguishes among the A_i in precisely the same way that $I = (B_1 \oplus B_2)$ distinguishes among the B_i. Thus we require that A and B assign the same functions, giving the symmetry equations,

$$f_i = g_i. \tag{6.5}$$

Combining equations (6.4) and (6.5), we find the functional equation $f_1 = f_2$. Requiring the distribution to be normalized then gives $f_1 = f_2 = 1/2$, the PI assignment for $N = 2$. This result is straightforwardly generalized to $N > 2$ by letting B number the hypotheses with an order-N cyclic permutation of A's numbering.

As a simple example with a continuous parameter, consider assigning a prior density to a location parameter, l, like that considered in the Gaussian estimation problem discussed in Section 2. Intuitively, if we are ignorant of a location, a displacement of a small amount does not change our state of knowledge. Thus I will distinguish among values of l in precisely the same manner in which it distinguishes among values of $l' = l + C$. In this way, the physical meaning of a location parameter identifies a class of parameterizations—those differing by translations—that lead to mathematically equivalent probability assignment problems. For such parameters, the functional equation (6.3) takes the form, $f(l) = f(l + C)$, with C a constant. The solution to this functional equation, unique up to a constant factor, is $f(l) = constant$. This is the prior we used in the Gaussian estimation problem and in the truncated exponential problem in Section 3. Priors for several other types of parameters have been found by group invariance; Jaynes (1968, 1973, 1980) and Bretthorst (1988) discuss several important examples. Future development of this method may be hastened by finding ways to symbolically express the information I in a manner that explicitly identifies mathematical equivalence between problems, as we did for the PI problem above.

Frequently a model parameter will have no direct physical significance, as is often true when we are fitting lines or polynomials to data. In such cases, there may not be an obvious choice of transformation corresponding to prior ignorance. Such parameters have meaning only insofar as they are useful for summarizing or predicting data. In such cases priors can sometimes be identified by specifying ignorance about *predictions* rather than about the parameters themselves. That is, prior ignorance may best be formulated, not in reference to the prior, but in reference to the prior predictive distribution, $p(D \mid I)$. Even parameters with obvious physical meaning may not have an obvious group invariance; priors for these parameters, too, may best be found by focusing on their predictive aspects. Specifically, recall that the prior predictive distribution can be calculated according to

$$p(D \mid I) = \int p(\theta \mid I) p(D \mid \theta I) d\theta. \tag{6.4}$$

Predictive methods seek to find the prior by specifying $p(D \mid I)$, and solving the integral equation (6.4) for the prior. They are particularly useful when the data are discrete, which can greatly simplify the task of assigning a least informative predictive distribution.

For example, consider the estimation of a Poisson rate, r, discussed in Section 5. When we are so uncertain of r that we cannot even exclude the possibility that $r = 0$, it is not clear what group invariance is relevant (scale invariance, usually invoked for such problems, is excluded if r may vanish). On the other hand, intuition suggests that ignorance of the rate corresponds to not having any prior preference for seeing any particular number of counts; $p(n \mid I)$ should be constant with respect to n. The prior predictive for this Poisson problem is given by

$$p(n \mid I) = \frac{1}{n!T} \int_0^\infty d(rT)\; p(r \mid I)\; (rT)^n e^{-rT}. \tag{6.5}$$

For $p(n \mid I)$ to be constant with respect to n, the integral must be proportional to $n! = \Gamma(n+1)$. But if $p(r \mid I)$ is constant with respect to r, the integral is, up to a constant factor, the definition of $\Gamma(n+1)$; further, since the integral is of the form of a Laplace transform, this is the unique solution, up to a constant factor. Thus the prior expressing ignorance about the number of counts we expect to observe is the constant prior we used throughout Section 5.

Some other methods available for assigning both least informative and informative priors are briefly mentioned in Loredo (1990).

6.4 Prior Robustness

Before one worries too much about the precise functional form of a prior, it is worthwhile to investigate to what extent details of the prior will influence the posterior in the problem of interest. Consider again the Gaussian estimation problem discussed in Section 2. There we found the posterior for l to be a Gaussian distribution with a mean and standard deviation of $\overline{m}$ and $\sigma/\sqrt{N}$, respectively. These results were found using a uniform prior for l, which we have just argued correctly expresses ignorance about a location parameter.

Suppose we had much stronger prior information about l, perhaps from a previous measurement, that was itself described by a Gaussian distribution with mean l_0 and standard deviation δ. From Bayes' theorem it is easy to show that the resulting posterior for l remains Gaussian, but with a mean and standard deviation of,

$$\hat{l} = \frac{\overline{m} + \alpha^2 l_0}{1+\alpha^2}, \quad \text{and} \quad \sigma_l = \frac{\sigma}{\sqrt{N}}(1+\alpha^2)^{-1/2}, \tag{6.1}$$

where $\alpha = \sigma/(\delta\sqrt{N})$. From these equations we see that unless $\delta \lesssim \sigma/\sqrt{N}$ (so that $\alpha^2 \gtrsim 1$), the posterior calculated with the Gaussian prior is not significantly different from that calculated with a uniform prior, even though the priors are very different.

This result should come as no surprise. It simply says that the prior will have little effect on our inferences unless our prior information is as informative as the data. Savage has elevated this observation to a principle, the 'principle of stable estimation:' if the likelihood is large in a region where the prior does not change strongly, and if the prior nowhere enormously exceeds its value in this region, then it is a good approximation to use a flat prior (see Edwards, Lindman, and Savage 1963). In such cases the information provided by the data overwhelms the prior information, and the data essentially "speak for themselves." This will usually be the case when there is a lot of data. In such cases there is no need to bother about whether our prior information is translation invariant, scale invariant, or diffuse in some other specific manner; it will simply not matter in the end.

Of course, when we do not have a lot of data, the precise form of the prior may strongly affect our inferences. For example, consider the estimation of a Poisson mean discussed in Section 5.1. Use of a uniform prior, which we justified above with a predictive argument, led to the posterior given by equation (5.5). If instead we were sure *a priori* that $r > 0$, but had no knowledge of the scale of r, a group invariance argument would lead to a prior proportional to $1/r$, rather than a constant prior (Jaynes 1968). The net affect on the posterior is to replace every occurrence of n on the right hand side of equation (5.5) with $n - 1$. When many counts are observed, this change has little effect on inferences: the posterior mode moves from n/T to $(n+1)/T$, and the standard deviation changes from $\sqrt{n-1}/T$ to $\sqrt{n}/T$. But when only a single count is observed, the posteriors differ significantly. The posterior based on the scale-invariant prior decays exponentially from its maximum at $r = 0$, whereas that from the uniform prior vanishes at $r = 0$, rises to a maximum at $r = 1/T$, and then decays.

This example shows us that when the data do not tell us much, what we know after consideration of the data strongly depends on what we knew without the data. When in doubt about how informative the data are, one should perform calculations with several priors to determine how robust posterior inferences are with respect to prior knowledge. If the posterior depends sensitively on the prior, we still learn something important: we learn that the data provide little information. Precise conclusions will then only follow if prior information can be precisely specified.

Posteriors *can* depend sensitively on the prior. Far from a weakness of the Bayesian approach, we consider this to be an important asset of the theory. In this manner, it automatically warns us when the data are uninformative.

6.5 Objective Bayesian Inference

The sense that priors make Bayesian methods too subjective for use by scientists seeking scientific objectivity has been exacerbated by the insistence on the part of many Bayesian statisticians that Bayesian probabilities describe the personal opinions or beliefs of individuals, and that two individuals possessing the same factual information can nevertheless assign different probabilities. This 'subjective Bayesian' viewpoint is perhaps most closely identified with L.J. Savage (see, *e.g.*, Edwards, Lindman, and Savage 1963). In the astrophysical literature, Sturrock (1973) has proposed the use of subjective probabilities to systematize personal evaluations of astrophysical theories.

Here we have taken the viewpoint, sometimes called the 'objective Bayesian' viewpoint, that probabilities are an encoding of *information*, not opinions or beliefs. Within this viewpoint, we may still consider probabilities that encode the state of knowledge of an individual if we wish; by understanding I to be the proposition, 'everything person X knows about A', $p(A \mid I)$ then becomes a description of the state of knowledge of person X. But we insist that these probabilities describe, not the opinions or beliefs of X, but the consequences of the knowledge or assumptions on which these are based. Thus two people in the same state of knowledge about a proposition must assign it the same probability. The importance of this simple consistency principle, the key to finding objective prior probability assignments, has been emphasized by Jaynes (1968, 1983, 1993); I have called it *Jaynes Consistency* (Loredo 1990). Surely a mathematical theory of uncertainty must satisfy this simple consistency requirement if it is to have any claim at all to scientific "objectivity."

7. BAYESIAN INFERENCE IN ASTROPHYSICS

The applications discussed here demonstrate the pragmatic and conceptual superiority of Bayesian inference for the analysis of astrophysical data. Jeffreys advocated the use of Bayesian methods in geophysics and astronomy long ago, and developed Bayesian solutions for a wide variety of important problems (Jeffreys 1939); but his work was largely ignored, partly because of conceptual problems that were only resolved by others after his work. In recent years, several investigators have finally taken up Jeffrey's challenge, and have begun applying Bayesian methods to the analysis of a variety of astrophysical data. The work of Kraft, Burrows, and Nousek (1991) analyzing Poisson counting data, of Loredo and Lamb (1992) analyzing the neutrinos observed from supernova SN 1987A, and of Gregory and Loredo (1992) analyzing event arrival times for periodicity, has been mentioned above. Here we briefly mention other published analyses.

Bretthorst (1988) has developed a rich theory for the analysis of data sampled with Gaussian noise, extending earlier work of Jaynes (1987). Applied to periodic models, Bretthorst's algorithm can measure periodic signals with precision and sensitivity greater than that obtained with standard methods based on the discrete Fourier transform, particularly when the signal is more complicated than a single sinusoid. Bretthorst (1988) has presented a preliminary analysis of almost 300 years of sunspot data demonstrating the superiority of Bayesian methods for the analysis of such data.

In another preliminary study, Jaynes (1988) and Bretthorst and Smith (1989) have applied Bretthorst's methods to the problem of resolving closely spaced point sources with separations significantly smaller than the width of the imaging point spread function, demonstrating that Bayesian methods can easily resolve such objects under certain conditions.

Morrow and Brown (1988) have applied Bayesian methods to the analysis of helioseismology data. Their calculation uses prior information about the relationship between the frequencies and wavenumbers of solar oscillations to make an ill-posed fitting problem well-posed.

Goebel, *et al.* (1989) have applied the 'AutoClass II' Bayesian classification program developed by Cheeseman, *et al.* (1988) to the problem of identifying classes of objects in the Low Resolution Spectra (LRS) atlas of objects observed by the Infrared Astronomical Satelite (IRAS). AutoClass II applies Bayesian parameter estimation and model comparison principles to the spectra of over 5000 objects in the atlas to automatically classify the objects into a hierarchy of classes whose number and parameters are found automatically from the data. Many of the resulting classes are in concert with those previously identified by the IRAS Science team and other later investigators, but several new classes were also identified. A number of these have been verified to be distinct classes by independent observations of additional properties of the member objects, such as their spatial distribution. The models underlying AutoClass II are very simple; its great success in classifying IRAS LRS data should motivate its application to other astrophysical data, as well as the extension of the algorithm to more sophisticated models.

Bayesian notions have inspired the development of maximum entropy methods for the deconvolution of astrophysical images, though all published astrophysical applications of such methods have so far relied on frequentist statistical criteria. Recently Gull (1989) and Skilling (1990) have developed a fully Bayesian entropic deconvolution method that provides, not only a single 'best' deconvolution, but a probability distribution for the flux in various regions of the image, allowing calculation of 'error bars.' Sibisi (1990) has applied this method to the analysis of nuclear magnetic resonance data, but it has yet to be applied to astrophysical data.

The past several decades have seen a growth of interest in Bayesian methods in applied statistics, econometrics, and other fields so rapid that it has been termed a 'Bayesian revolution.' We look forward to a similar revolution occurring in astrophysics—the field for which Laplace first developed such methods—bringing with it new clarity and precision in the quantification of uncertainty, and better enabling astrophysics to fulfill its promise as the arena in which the unifying power of physics can be most spectacularly demonstrated.

ACKNOWLEDGMENTS. It is a pleasure to thank Don Lamb, Phil Gregory, Larry Bretthorst, Ed Jaynes, David Hestenes, and Arnold Zellner for many valuable conversations and for their encouragement. Larry Bretthorst also offered helpful comments on an earlier draft of this work. I am also grateful to Ed Jaynes for access to unpublished manuscripts. This work was supported by NASA grants NGT-50189, NAGW-830, and NAGW-1284 at the University of Chicago, and by a NASA GRO Fellowship, NASA grant NAGW-666, and NSF grants AST-87-14475 and AST-89-13112 at Cornell University.

REFERENCES

Aczel, J. (1966) *Lectures on Functional Equations and their Applications*, Academic Press, New York.

Annis, M., W. Cheston, and H. Primakoff (1953) 'On Statistical Estimation in Physics', *Rev. Mod. Phys.* **25**, 818–830.

Basu, D. (1975) 'Statistical Information and Likelihood', *Sankhyā* **37**, 1–71.

Basu, D. (1977) 'On the Elimination of Nuisance Parameters', *J. Amer. Stat. Assoc.* **72**, 355–366.

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., 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.

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

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

Bretthorst, G.L. (1990a) 'Bayesian Analysis I: Parameter Estimation Using Quadrature NMR Models', *J. Magn. Reson.* **88**, 533–551.

Bretthorst, G.L. (1990b) 'Bayesian Analysis II: Signal Detection and Model Selection', *J. Magn. Reson.* **88**, 552–570.

Bretthorst, G.L. (1990c) 'Bayesian Analysis III: Applications to NMR Signal Detection, Model Selection and Parameter Estimation', *J. Magn. Reson.* **88**, 571–595.

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.

Buccheri, R., and O.C. DeJager (1989) 'Detection and Description of Periodicities in Sparse Data. Suggested Solutions to Some Basic Problems', in H. Ogelman and E.P.J. van den Heuvel (eds.), *Timing Neutron Stars*, Kluwer Academic Publishers, Dordrecht, pp. 95–111.

Cheeseman, P., J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman (1988) 'AutoClass: A Bayesian Classification System', in J. Laird (ed.), *Proceedings of the 5th International Conference on Machine Learning*, Morgan Kaufmann Publishers, Inc., San Mateo, CA, pp. 54–64.

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.

Cleveland, T. (1983) 'The Analysis of Radioactive Decay With a Small Number of Counts by the Method of Maximum Likelihood', *Nuc. Instr. and Meth.* **214**, 451–458.

Collura, A., A. Maggio, S. Sciortino, S. Serio, G.S. Vaiana, and R. Rosner (1987) 'Variability Analysis in Low Count Rate Sources', *Ap. J.* **315**, 340–348.

Cornfield, J. (1969) 'The Bayesian Outlook and its Application', *Biometrics* **25**, 617–642.

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 Psychological 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.

Efron, B. (1978) 'Controversies in the Foundations of Statistics', *Math. Monthly.?*, 231–246.

Feigelson, E.D. (1989) 'Statistics in Astronomy', in S. Kotz and N.L. Johnson (eds.), *Encyclopedia of Statistical Science, Supplement Volume*, John Wiley and Sons, New York, p. 7.

Garret, A.J.M. (1991) 'Ockham's Razor', in W.T. Grandy and L. Schick (eds.), *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, in press.

Gehrels, N. (1986) 'Confidence Limits for Small Numbers of Events in Astrophysical Data', *Ap. J.* **303**, 336–346.

Geisser, S. (1988) 'The Future of Statistics in Retrospect', in J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.), *Bayesian Statistics 3*, Oxford University Press, pp. 147–158.

Goebel, J., K. Volk, H. Walker, F. Gerbault, P. Cheeseman, M. Self, J. Stutz, and W. Taylor (1989) 'A Bayesian Classification of the IRAS LRS Atlas', *Astr. Ap.* **222**, L5–L8.

Good, I.J. (1950) *Probability and the Weighing of Evidence*, Griffin, London.

Gregory, P.C., and T.J. Loredo (1992) 'A New Method for the Detection of a Periodic Signal of Unknown Shape and Period', submitted to *Ap. J.*.

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.

Hearn, D. (1969) 'Consistent Analysis of Gamma-Ray Astronomy Experiments', *Nuc. Instr. and Meth.* **70**, 200.

Helene, O. (1983) 'Upper Limit of Peak Area', *Nuc. Instr. and Meth.* **212**, 319–322.

Hestenes, D. (1989) 'Toward a Modeling Theory of Physics Instruction', *Am. J. Phys.* **?**, ?.

Hillas, A.M. (1975) 'Note on the Probability of Observing an Excessive Number of Counts', *Proc. 14th International Cosmic Ray Conference*, **9**, 3439–3443.

Howson , C., and P. Urbach (1989) *Scientific Reasoning: The Bayesian Approach*, Open Court Press, LaSalle, IL.

Jaynes, E.T. (1968) 'Prior Probabilities', *IEEE Trans.* **SSC-4**, 227. Reprinted in Jaynes (1983).

Jaynes, E.T. (1973) 'The Well-Posed Problem', *Found. of Phys.* **3**, 477. Reprinted in Jaynes (1983).

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. Reprinted in Jaynes (1983).

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. Reprinted in Jaynes (1983).

Jaynes, E.T. (1980) 'Marginalization and Prior Probabilities', in A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics*, North-Holland, Amsterdam, p. 43. Reprinted in Jaynes (1983).

Jaynes, E.T. (1983) *Papers on Probability, Statistics, and Statistical Physics* (ed. R.D. Rosenkrantz), D. Reidel Pub. Co., Dordrecht.

Jaynes, E.T. (1986) '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. (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. (1988) '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. (1990) 'Probability Theory as Logic', in P. Fougere (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht.

Jaynes, E.T. (1993) *Probability Theory—The Logic of Science*, in preparation.

Jefferys, W., and J. Berger (1992) 'Ockham's Razor and Bayesian Analysis', *Am. Scientist* **80**, 64–72.

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

Kendall, M., and A. Stuart (1979) *The Advanced Theory of Statistics, Vol. 2: Inference and Relationships*, Charles Griffin & Company Limited, London.

Kraft, R.P., D.N. Burrows, and J.A. Nousek (1991) 'Determination of Confidence Limits for Experiments with Low Numbers of Counts', *Ap. J.*, in press.

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

Leahy, D.A., W. Darbro, R.F. Elsner, M.C. Weisskopf, P.G. Sutherland, S. Kahn, and J.E. Grindlay (1983) 'On Searches for Pulsed Emission with Application to Four Globular Cluster X-Ray Sources: NGC 1851, 6441, 6624, and 6712', *Ap. J.* **266**, 160–170.

Leahy, D.A., R.F. Elsner, and M.C. Weisskopf (1983) 'On Searches for Periodic Pulsed Emission: The Rayleigh Test Compared to Epoch Folding', *Ap. J.* **272**, 256–258.

Lee, Peter M. (1989) *Bayesian Statistics: An Introduction*, Oxford University Press, New York.

Li, T.-P., and Y.-Q. Ma (1983) 'Analysis Methods for Results in Gamma-Ray Astronomy', *Ap. J.* **272**, 317–324.

Lindley, D.V. (1990) 'The 1988 Wald Memorial Lectures: The Present Position in Bayesian Statistics', *Statistical Science* **5**, 44–89.

Lindley, D.V., and L.D. Phillips (1976) 'Inference for a Bernoulli Process (a Bayesian View)', *Am. Statistician* **30**, 112–119.

Lobo, J.A. (1990) 'Estimation of the Arrival Times of Gravitational Waves From Coalescing Binaries: The Performance of a Long-Baseline Interferometric Gravitational Wave Antenna', *Mon. Not. Roy. Astr. Soc.* **247**, 573–583.

Loredo, T.J. (1990) 'From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics', in P. Fougère (ed.) *Maximum-Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, pp. 81–142.

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 (1992) 'Bayesian Analysis of Neutrinos from SN 1987A: Implications for Cooling of the Nascent Neutron Star and for the Mass of the Electron Antineutrino', submitted to *Phys. Rev. D.*

MacKay, D. (1992) 'Bayesian Interpolation', submitted to *Neural Computation.*

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

Morrow, C.A., and T.M. Brown (1988) 'A Bayesian Approach to Ridge Fitting in the $\omega - k$ Diagram of the Solar Five-Minute Oscillations', in J. Christensen-Dalsgaard and S. Frandsen (eds.), *Advances in Helio- and Asteroseismology*, International Astronomical Union, pp. 485–489.

Nicholson, W.L. (1966) 'Statistics of Net-Counting-Rate Estimation with Dominant Background

Corrections', *Nucleonics* **24**, #8, 118–121.

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

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling (1986) *Numerical Recipes*, Cambridge University Press, Cambridge.

Press, W.H., and P. Schechter (1974) 'Remark on the Statistical Significance of Flares in Poisson Count Data', *Ap. J.* **193**, 437–442.

Protheroe, R.J. (1985) 'A New Statistic for the Analysis of Circular Data in Gamma-Ray Astronomy', *Proc. 19th International Cosmic Ray Conference (LaJolla)*, **3**, 485–488.

Protheroe, R.J. (1987) 'Periodic Analysis of Gamma-Ray Data', *Proc. Astron. Soc. Austr.* **7**, 167–172.

Rainwater, L.J., and C.S. Wu (1947) 'Applications of Probability Theory to Nuclear Particle Detection', *Nucleonics* **1**, #2, 60–69.

Sard, A., and R.D. Sard (1949) 'Some Statistical Considerations on Coincidence Counting', *Rev. Sci. Instr.* **20**, 526.

Scargle, J.D. (1982) 'Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data', *Ap. J.* **263**, 835–853.

Sibisi, S. (1990) 'Quantified MAXENT: An NMR Application', in P. Fougere (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht.

Skilling, J. (1990) 'Quantified Maximum Entropy', in P. Fougere (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht.

Stigler, S.M. (1982) 'Thomas Bayes's Bayesian Inference', *J. Roy. Stat. Soc.* **A145**, 250–258.

Sturrock, P.A. (1973) 'Evaluation of Astrophysical Hypotheses', *Ap. J.* **182**, 569–580.

West, M., and J. Harrison (1989) *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York.

Zech, G. (1989) 'Upper Limits in Experiments with Background or Measurement Errors', *Nucl. Inst. and Meth. in Phys. Res.* **A277**, 608–610.

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

Zhang, S.N., and D. Ramsden (1990) 'Statistical Data Analysis for Gamma-Ray Astronomy', *Exp. Astron.* **1**, 145–163.