The Return of the Prodigal: Bayesian Inference For Astrophysics

Thomas J. Loredo
Department of Astronomy, Space Sciences Building,
Cornell University, Ithaca, New York 14853

ABSTRACT

Astronomers are skeptical of statistical analyses, the more so the more sophisticated they are. This has been true especially of Bayesian methods, despite the fact that such methods largely originated in the astronomical analyses of Laplace and his contemporaries in the early 1800s. I argue here that astronomers hold statistics in low regard because many astronomers are poor statisticians. Further, I argue that astronomers are poor statisticians because the frequentist methods they use have characteristics that invite statistical sloppiness when they are used by nonexperts. The Bayesian approach to statistical inference does not share these characteristics; adoption of Bayesian methods by astronomers thus promises to improve statistical practice in astronomy. I present a simplified discussion of some of the issues arising in the recent analysis of an important astrophysical data set—that provided by the Cosmic Background Explorer satellite—to illustrate some of the practical advantages of a Bayesian outlook. I offer some advice on how to educate astronomers about Bayesian methods. I conclude with a brief survey of recent applications of Bayesian methods to the analysis of astrophysical data. The breadth and number of these applications may well indicate that the time for Bayesian methods to return to the field of their origin has arrived.

---

1. INTRODUCTION

One could claim without too much exaggeration that statistical inference was invented because of astronomy. As noted by Stigler (1986), problems associated with reconciling discrepant observations in astronomy and geodesy motivated such legendary mathematicians and astronomers as Legendre, Laplace, and Gauss to develop the foundations of statistical inference based on probability theory. Their analyses of astronomical and geodetic problems led to such notions as the use of means to reduce uncertainty, the method of least squares, the normal distribution, the central limit theorem, and the "method of inverse probability" (inference using Bayes's theorem). Their work was essentially Bayesian in outlook, and the first mature treatise on statistical inference—Laplace's *Theorie Analytique des Probabilités* (Laplace 1812)—could fairly be called a Bayesian monograph.

Viewed from the present, this aspect of the early history of statistical inference is doubly ironic. First, contemporary astronomers (and physical scientists more generally) seldom receive any formal training in statistics, and frequently display a skepticism of sophisticated statistical analysis that borders on disdain. Second, until very recently, Bayesian methods *in particular* have been poorly understood and unwelcome tools among physical scientists. This has been true despite the fact that the most influential and practical Bayesian text of the first half of this century was written by a geologist and astronomer, Sir Harold Jeffreys (Jeffreys 1939).

While physical scientists have been ignoring Bayesian methods, these methods have been receiving increasing attention from applied statisticians and from practicing scientists in various other scientific disciplines, most notably biometrics and econometrics. Although Lindley's prediction of a Bayesian twenty-first century (Lindley 19xx) now appears somewhat optimistic, it is nevertheless true that Bayesian methods are flourishing outside the physical sciences, and are now well understood and broadly accepted by statisticians.

In this paper, I will argue that it is time for "the Prodigal" to return home: for Bayesian inference to offer its insights to its home discipline of astronomy with the same prodigality with which it has recently offered them to other disciplines. Indeed, there is already evidence that the Prodigal is in sight, for a number of investigators in various areas of astronomy have independently begun introducing Bayesian methods into the astronomer's toolbox in recent years.

In the next section I will describe current attitudes of astronomers toward statistics. I will document the low regard astronomers have toward statistics, and offer my own opinions as to why this is true. In a word: astronomers have a low regard for statistics because many astronomers are *horrible* statisticians. I will argue that the poor statistical performance of astronomers arises because many characteristics of the traditional "frequentist" approach to statistics that dominates astronomical practice invite statistical sloppiness. I will briefly describe how Bayesian inference, by its very structure, prevents or discourages some of the bad practices of astronomers. In Section 3, I will describe a few of the statistical mistakes astronomers commonly make. In Section 4, I will present a simplified version of a recent, important analysis of astrophysical data that illustrates both the pitfalls of traditional thinking and the benefits of Bayesian thinking. Presuming that the first four sections motivate some readers to want to educate astronomers about Bayesian methods, Section 5 offers advice on how Bayesian statisticians might make their work more accessible to astronomers. Finally, in Section 6, I will briefly point out some recent applications of Bayesian inference in astronomy.

I am an astrophysicist, and know best the situation in my own discipline. But it is my perception that much of what I will say applies also to other disciplines in the physical sciences. Thus readers may not err too severely if they read "physical sciences" wherever they see "astronomy," or "physicist" wherever they see "astronomer." I offer this essay to this audience of Bayesian statisticians in the hope that I might convince some of you that astronomers—and physicists more generally—need your insights, and motivate some of you to work toward bridging the gulf that exists between our disciplines. It appears to me that the gulf is not very wide, and that a small amount of effort on both sides could yield large rewards.

2. FRETTING ABOUT STATISTICS

A recent issue of *Physics Today* began with a brief editorial column titled, "Fretting About Statistics." The author, one of the most highly regarded experimentalists in atomic physics, offers the following observations about the status of statistics in the physical sciences:

> A colleague . . . claims that if you need to rely on statistics to understand your experiment, you are in serious trouble. The claim is obviously exaggerated, but he has a point: If you need to rely on statistics, you need to worry . . . . The power of statistics to deceive is so well known that the title "statistician" is slightly suspect . . . . With today's cheap and powerful workstations you can accumulate vast piles of data, analyze them in a jiffy and apply sophisticated statistical tests to reassure yourself that the data are consistent and that all is well. What you have really achieved, however, is the ability to fool yourself in a highly sophisticated manner . . . . (Kleppner 1992)

I have taken Kleppner's comments out of context with the consequence that the skepticism toward statistics that he describes is exaggerated. However, my own experience is that the exaggerated

skepticism may be more typical of many colleagues who have not thought as carefully about statistics as Kleppner may have.

Kleppner goes on to note that "it is only fair . . . to point out that statistical analysis has been crucial to more than a few dazzling discoveries." The principal example he cites is an astrophysical one we will discuss further in Section 5. The point I wish to make here is that, despite their distrust of statistics, physical scientists are finding themselves more and more often in the position of having to devise and rely on sophisticated manipulations of their data in order to draw scientific inferences from them. This is a natural consequence of the drive to understand physical systems at more detailed levels or in more extreme regimes than have been accessible previously in the history of science, thus requiring more complicated preparation and observation of the systems under study. In astronomy, these problems are exacerbated further by the fact that astronomy is an observational, rather than an experimental, science. Astronomers almost always observe phenomena of interest from great distances. The systems under study are thus inaccessible to direct manipulation, and the physical quantities of interest usually can only be observed indirectly, through their influence on light. This necessarily complicates one's inferences.

Thus the scientific inferences of physicists, and particularly of astronomers, are fraught with uncertainty, uncertainty which must be described quantitatively. As statistics is the mathematical discipline whose goal is to quantify uncertainty, most physical scientists who handle data inevitably find themselves "fretting" about statistics at some time in their career.

But why are astronomers *fretting* about statistics, instead of merely thinking or learning about statistics? Look as hard as you want, and you will not find an editorial by a physicist or an astronomer titled "Fretting About Integration" or "Fretting About Tensor Algebra." What is it that leads astronomers to react differently to statistics than they do to other mathematical disciplines? I believe the answer to this question lies in two related characteristics of *frequentist* statistics, the type of statistics which has dominated modern astronomical statistics. First, frequentist statistics bears a different relationship to the problems it solves than do other mathematical disciplines. Second, a consequence of this difference is that good statistical practice requires a high level of mastery of arcane knowledge, a mastery which astronomers lack. Let me now elaborate on these assertions.

2.1. *What Is Different About Frequentist Statistics?*

In their training, astronomers master several mathematical disciplines at reasonably high levels of sophistication. Some examples would include algebra; real, complex, and vector analysis (differentiation and integration); tensor algebra; and differential equations. Each of these disciplines shares the characteristic that they consist of *tools for finding definitive solutions to well-posed problems*. Implicit in this characterization is the assumption that a definite (and often unique) solution exists.

Take integration as an example. I recently came across the following integral in my work:

$$I(\kappa, \phi) \equiv \int_0^{2\pi} \exp[\kappa \cos(\theta - \phi)] d\phi. \tag{2.1}$$

Fortunately, integral representations of some Bessel functions are part of my "integration toolbox," so I could show that $I(\kappa, \phi) = I_0(\kappa)$, where I_0 denotes the 0th order modified Bessel function. Later in the same project, I came across the superficially similar integral,

$$I(\boldsymbol{\kappa}, \boldsymbol{\phi}) = \int_0^{2\pi} d\theta \, \exp\left[\sum_{\alpha=1}^{N} \kappa_\alpha \cos(\alpha\theta - \phi_\alpha)\right]. \tag{2.2}$$

Here $\boldsymbol{\kappa}$ and $\boldsymbol{\phi}$ denote vectors of parameters, with N components. This time my toolbox was not up to the task. None of my favorite tricks—integration by parts, parametric integration, Fourier or Laplace transforms—gave me the answer. Queries of other astronomers and physicists have failed to provide a satisfactory answer (although a theoretical particle physicist found a nested infinite series solution that, although useless in practice, is pretty impressive!).

The point of these examples is that for integration (and most other mathematical disciplines), you either have the tools you need to solve the problem or you do not, and there is no question about whether you have them or not: either you can find the answer or you cannot. In the latter case, consulting an expert (or gaining more expertise yourself) is not optional; if you want the answer, you have no choice but to seek the expertise.

Not so with frequentist statistics. It is qualitatively different from other mathematical disciplines in that it consists of *tools for characterizing tentative solutions to problems*, rather than tools for finding definitive solutions. For example, if I need to estimate a parameter, frequentist statistics will not "hand" me an estimator. Rather, it requires that I provide it one or more estimators (perhaps as a parameterized class); it will then provide me with various characterizations of how well they perform (bias, efficiency, etc.). But it is up to me to provide the estimators, to specify the relevant characterizations of their performance, and to choose how to balance the various characterizations in order to settle on a particular estimator.

For brevity, I will use the term *frequentist distinction* to refer to this distinguishing aspect of frequentist statistics: that it characterizes tentative solutions, rather than provides definitive solutions.

To many astronomers, this poorly appreciated distinction of frequentist statistics gives statistics an air of imprecision. It appears that, no matter how carefully we pose our problems, there are no "right" answers. Indeed, experience shows that it is not unusual for investigators who choose different statistics to reach different conclusions, with no compelling criteria available to choose between them. No wonder, then, that astronomers so often and so publicly "fret about statistics," trusting statistical results only when the conclusions are evident without quantitative analysis.

2.2. *What We Don't Know Does Hurt Us*

A consequence of the frequentist distinction is that much of the discipline appears to an outsider to be a collection of a huge number of arcane studies of the behavior of *ad hoc* statistics in many problems. It is true that for simple problems (essentially those with minimal sufficient statistics equal in number to the number of parameters) the statistic of choice is unambiguous and reasonably well known, at least among statisticians. But such problems do not take us very far in the real world. For problems of realistic complexity, only a statistician is likely to know where to look in the huge and arcane literature of statistics in order to find a suitable, well-studied method. More likely than not, the precise problem of interest has not been previously studied in depth, necessitating use of expert judgement in adapting solutions of similar problems, or in devising and studying new tentative statistics customized to the problem at hand.

We astronomers do not possess this arcane expert knowledge, and most of us have enough trouble mastering the arcane knowledge of our own discipline that there is little possibility of us mastering that of another as well. Ideally, then, we would consult an expert when we need statistical wisdom that we do not possess ourselves. Realistically, this does not happen. I suppose this may be due in part to arrogance or stubbornness. But in all fairness, I think there is a deeper reason.

Since frequentist statistics does not actually solve problems, astronomers who use frequentist methods are not forced to confront their statistical ignorance. We can apply a simplistic method to a complicated problem and get *an* answer. It may be a poor answer, but it is an answer nonetheless. For this reason, statistics is a field where a little knowledge can hurt you, because there is no "alarm" automatically warning you when you are applying a method outside its realm of applicability, or alerting you to the fact that superior methods exist. When I try to evaluate an integral, and do not have the tools, *I do not get an answer*. If I want to estimate a parameter, and choose simply to use some moment of the data, I will get an answer, regardless of whether moments give poor estimates for the data under consideration, and regardless of whether some other procedure exists that gives a demonstrably better answer.

2.3. *The Bayesian Alternative*

The Bayesian approach to inference is in an important sense more like integration than like frequentist statistics. Like integration and other mathematical disciplines, it is a collection of tools for finding definitive answers to well-posed problems. Once a problem is posed, the rules of probability theory lead one directly to its solution, providing one can do the required mathematics. One simply computes the probabilities of all hypotheses being considered, using Bayes's theorem or other applications of the basic sum and product rules of probability theory. Once a few "tricks of the trade"—like marginalization, or "extending the question"—are mastered, the procedure of writing out the formal solution becomes almost mechanical for most problems.

This being said, there are a number of practical problems that arise in the application of Bayesian methods that I do not wish to sweep under the proverbial carpet. I would divide such problems into two classes: problem specification, and Bayesian calculation. Some brief comments about these problems may be in order here.

In many mathematical disciplines the conditions required for a problem to be well-posed are so clear that they need hardly be stated explicitly. That such conditions exist for problems of inference seems to have escaped notice by many scientists, however. We know we cannot solve a differential equation unless we specify not just the equation, but also boundary conditions. It should hardly come as a surprise, then, that there are certain conditions that must be satisfied to make an inference problem well-posed. The Bayesian "recipe" is simple: to assess a hypothesis, calculate its probability conditional on all the information at hand. The requirements for a problem to be well-posed are: (1) That we explicitly identify all the relevant hypotheses, and (2) That we specify whatever other information is needed to guarantee that all the direct probabilities we need in our calculation can be unambiguously evaluated (indeed, (1) is a special case of (2)). There is inevitable subjectivity in both requirements. For the most part, this subjectivity is shared by both Bayesian and frequentist methods. It is perhaps most prominent in assigning prior probabilities needed in Bayesian calculations, and I will say a little more about this later. But subjectivity also arises in assigning sampling distributions, and this subjectivity—shared by Bayesian and frequentist methods—may well be more worrisome than that associated with priors, even though a century of frequentist focus has conditioned us to quick acceptance of a few "stock" sampling distributions. My own view is that it is an important advantage of Bayesian methods that their very structure makes the inherent subjectivity of statistical inference explicit. I will remark on a few specific examples of subjectivity hidden in frequentist methods but made visible with Bayesian methods in the following section.

Once a problem is well-posed, the rules of probability theory lead us to a formal statement of its solution in the form of formulas for calculating or summarizing the needed probabilities. These formulas inevitably require integrals over the hypothesis space, integrals which can be *very* hard to evaluate accurately if the hypothesis space is of moderate or large dimension. Indeed, devising clever algorithms for performing such integrals is one of the main areas of current research in Bayesian inference. But it is worth emphasizing that the difficulties of Bayesian calculation are *practical* difficulties; they arise because we know what we need to calculate. I am sure I am not alone in feeling more comfortable having to devise clever approximations to the known definitive solution to a problem than I would having to devise clever tentative solutions to the problem directly.

Practical difficulties aside, the essential point is that a well-posed inference problem has a unique and definitive Bayesian solution, but usually no definitive frequentist solution. Cleverness is required in both Bayesian and frequentist approaches, but cleverness of quite different kinds. Bayesian methods may require mathematical cleverness in finding ways to evaluate the formal solution (perhaps involving clever restatement of the problem). This is a kind of cleverness that astronomers and physical scientists have some experience with (in fact, much recent Bayesian work on Monte Carlo methods for evaluating Bayesian integrals is built on ideas that were first developed in the physics literature; e.g., the Metropolis algorithm). Frequentist methods require cleverness in specifying a

solution. This cleverness can only come from experience with advanced statistics and familiarity with the statistics literature. And unfortunately, the frequentist approach has no built in "alarms" alerting us when we have not been clever enough.

3. SEVEN STATISTICAL SINS

By its nature, then, frequentist statistics lets astronomers avoid their statistical ignorance. In this section, I will describe seven common errors I have personally seen, repeatedly, in the astronomical literature: consequences of us trying to get by on what little we know. Some of these "sins" display misunderstandings so basic that they might well fail a beginning statistics student out of his or her first statistics course! Yet all of these blunders have been commited by astronomers who, in other respects, are quite excellent scientists.

Before I completely alienate any astronomical colleagues who have stumbled upon this paper, I should point out that I have made every one of these mistakes myself. I have the advantage of having made most of them while still a graduate student, so that few of them have appeared in print; but I am as guilty as other astronomers of sloppy statistical thinking. It is my personal experience that a Bayesian outlook naturally guides one away from some blind alleys that otherwise may appear attractive. For each of the problems I mention, I will thus describe how Bayesian methodology discourages or prevents bad statistical practice.

3.1. *Confusing Parameter Estimation With Model Criticism*

Frequentist methodology does not consider probabilities for hypotheses conditional on the observed data, but considers only probabilities for the observed and hypothetical data conditional on an hypothesis. This encourages sloppy thinking regarding specification of the hypothesis space relevant to a problem. The first three "sins" I will discuss all originate from trying to solve a problem that is not well-posed because of incomplete or incorrect problem specification.

The first sin is the failure to distinguish between two different kinds of inference problems: parameter estimation and model criticism. With alarming frequency, astronomers use goodness-of-fit (model criticism) methods to find the boundary of a "confidence region," or we quote the covering probability of a confidence region as the "significance" of a parameter (astronomers' slang for one minus the Type I error probability for a model with the parameter fixed at a default value). This sin is usually commited only when the parameter must be nonnegative (which is true of most physical parameters).

I first committed this sin myself in an attempt to measure the mass of the fundamental particle called the electron antineutrino using arrival times of about two dozen of them detected from a supernova—a star that exploded—observed in a companion galaxy to our own Galaxy in 1987. The simplest theories of fundamental particles require neutrinos to have zero mass. If this is true, the equations of special relativity imply that all neutrinos must move with the velocity of light, regardless of how much energy they carry. More complicated theories, however, allow the neutrino to have nonzero mass, in which case more energetic neutrinos should move faster than less energetic ones. The neutrinos detected from that exploding star were the first to be detected from a known source outside our solar system. They traveled a distance of about 150,000 light years—vastly larger than any terrestrial length scale—giving us hope that even very small velocity differences could be detected at Earth as a detectable spread in arrival times of the neutrinos, the most energetic ones arriving first. Unfortunately, the source itself emitted neutrinos over a broad energy range and over a timescale similar in magnitude to the lag expected from interesting mass values, weakening our ability to measure the neutrino mass. Still, it remained interesting to specify quantitatively what upper limit the data implied on the neutrino mass.

To find a "95% confidence upper limit," I devised a goodness-of-fit (GOF) statistic, S, and plotted its value as a function of neutrino mass, m. For each mass value, I simulated many hypothetical

data sets, calculated their GOF statistics, and found the fraction of these, $F_<$, that had S values worse (larger) than that for the actual data. I identified the mass value, m_{95}, beyond which $F_<$ fell below 0.05 as the 95% confidence upper limit. Of course, this procedure is completely falacious. The 95% probability associated with m_{95} refers to a property of that mass value only (it is $1-\alpha$, where α is the Type I error probability for a hypothesis test that would just reject the m_{95} model). It does not refer to the interval of mass values below m_{95}, as would the covering probability for a confidence region. Since the best-fit point ($m = 0$) had $F_< \approx 0.3$, no "confidence regions" of size below 30% exist with this method, but surely there is an interval of m for which we can meaningfully assign 30% confidence.

Fortunately, I discovered the errors of my ways before attempting to publish my results. Unfortunately, many other astronomers committed the same blunder, and both they and their referees failed to notice it, so that the literature analyzing the supernova neutrinos is filled with several such erroneous confidence region calculations, for the neutrino mass and for other parameters. I have since seen the same mistake made in *many* other contexts, with distressing frequency.

The reverse mistake is also made. An example is given in Section 4, below. To determine whether data provide evidence that a parameter, Q, is nonzero, an estimate of the parameter was made, with an estimate of its uncertainty (a standard deviation). The measured Q value was three standard deviations away from zero, and this was deemed a positive detection with "3σ significance," astronomers' slang for a probability for falsely rejecting the $Q = 0$ hypothesis of 0.37% (the tail area outside 3σ for a Gaussian). Of course, the covering probability for the confidence region is not simply related to the Type I error for a GOF test of the $Q = 0$ hypothesis. A crude χ^2 test, in fact, indicates consistency with the $Q = 0$ hypothesis. The Bayesian analysis of Section 4 seems to agree with this conclusion (the full data are not publicly available, making a definitive statement impossible).

A point that I must emphasize is that these mistakes have been made by some of the most highly regarded astronomers in the world. I cannot emphasize this too strongly: Very basic mistakes are being made here, but they are being made by otherwise quite brilliant people. To me this is evidence, not merely of ignorance on the part of those making the mistakes, but of something inherently confusing about the methods.

The confusion arises because there are several qualitatively different probabilities in frequentist statistics. Covering probabilities for confidence regions, Type I error probabilities, Type II error probabilities—all of these are quantities that span $[0, 1]$ that scientists can use to assess the reasonableness of hypotheses. But none of them are probabilities *for hypotheses*, so it is easy for nonexperts to confuse which is most closely related to the question they are asking. This confusion is exacerbated by the fact that all frequentist probabilities must condition on a particular point hypothesis, even those that refer to an entire class of hypotheses. For some problems (particularly for confidence region calculations), the hope is that the final result is independent of the particular hypothesis used. But this is seldom true in real problems, so that one hypothesis must inevitably be chosen to "represent" a class of hypotheses (e.g., approximate confidence regions are found using calculations conditioning on the best-fit hypothesis).

This confusion simply cannot arise in the Bayesian approach. One always calculates probabilities for hypotheses, so there is never ambiguity over what kind of hypothesis one's probability is associated with: you have to explicitly state it in order even to start the calculation. If I want to know how sure I can be that a parameter is in some region, I simply calculate the probability that it is in that region (parameter estimation). If instead I want to know how sure I can be that the parameter is zero, I must calculate the probability for the hypothesis that the parameter is zero (model comparison). The formalism forces one to distinguish between these options. This is an example of the property I alluded to in the previous section; that Bayesian methods by their nature make explicit the things we must specify to make an inference problem well-posed.

3.2. *Failure to Specify Alternatives*

The Bayesian methodology will force one to distinguish between estimation and model criticism. An important characteristic of Bayesian methods is that model criticism must always take the form of model *comparison*. A Bayesian calculation cannot assess the viability of one hypothesis without explicit consideration of alternative hypotheses, unless one is satisfied with a model probability equal to one, regardless of the data (so long as the data are possible consequences of the model). The formalism *forces* you to specify alternatives. This is in stark contrast to the frequentist notion of a goodness-of-fit test, with which one can assess the viability of a hypothesis without specifying an alternative.

This is an important distinction between Bayesian and frequentist approaches. After averaging and least-squares (or minimum χ^2) parameter estimation, probably the most common statistical procedure in astronomy is the GOF test, usually based on the χ^2 statistic for measurements with "errors," or the Kolmogorov-Smirnov D statistic for samples from a point process. That there are no obvious Bayesian counterparts to GOF tests seems to some to be a serious defficiency of the Bayesian approach.

Here I will instead argue that the failure of GOF tests to explicitly consider alternatives is the problem, and that the second statistical sin of astronomers is their failure to recognize the dependence of their conclusions on unspecified alternatives.

We might summarize the difference between the Bayesian and GOF approaches for assessing models as follows. If the observed data are improbable presuming a model is true, then there are two possibilities: either a rare event has occurred, or another model is true. The reasoning of GOF tests is that we should always presume another model is true when a "rare" event occurs. The reasoning of Bayesian methods is that we should only presume another model is true if another model would make the data sufficiently more probable than they would be if the model under consideration is true.

Viewed in this somewhat abstract manner, it is hard not to accede superiority to the Bayesian approach. It quantitatively considers both of the possibilities underlying rare outcomes. The GOF approach categorically rejects one possibility without bothering to assess its viability. Further benefits of the Bayesian approach acrue when one considers how to distinguish "rare" events from other events. For most hypotheses, the sample space is so large that *all* events have low probability. In the continuum limit, any particular event has zero probability! Thus GOF tests must consider a variety of hypothetical data that are somehow "like" the observed data to develop a useful definition of "rare," introducing inevitable subjectivity into the result.

Of course, these problems with GOF tests are not news to frequentist statisticians, nor are they ignored by them. From the beginning, Neyman and Pearson recognized the contextual character of hypothesis testing, and specified that a test be characterized not only by the probability of falsely rejecting the null, but also by the probability of falsely accepting it. The latter probability, of course, explicitly depends on alternatives.

Unfortunately, astronomers seldom consider the power of a hypothesis test. When astronomers say a test is "powerful," they almost always mean it in a colloquial sense: the test usually gives the right answer, in their experience. As best as I can determine, the majority of astronomers are not even aware that "power" has a technical meaning, and that determining the power of a test requires explicit consideration of alternative models. Many seem to quote Type I error probabilities as if they completely described the viability of the hypothesis under study. Sometimes hypotheses are accepted by one investigator, only to be rejected by another who chose a different GOF statistic. That the choice of statistic might be related to an implicit choice of alternatives is seldom recognized.

In frequentist hypothesis testing, we can get away with ignoring alternatives because the frequentist approach characterizes a test with *two* probabilities: the false rejection and false acceptance probabilities. The former do not depend on alternatives, so despite the warnings of Neyman and Pearson, it remains *possible* to associate a probability with a model without considering alternatives,

encouraging one to ignore alternatives altogether. In a Bayesian calculation, on the other hand, the viability of a model is determined by *one* probability—that for the model itself—and it is *not* possible to calculate this probability without explicit consideration of alternatives. The formalism forces one to acknowledge the contextual nature of such inferences.

Quite frequently in preprints and talks (but less so in the published literature, presumably thanks to the care of referees and editors) astronomers will call the Type I error probability "the probability for the model." This is a very revealing error. It indicates quite clearly that the quantity astronomers seek is the quantity that Bayesian calculations alone can provide. We might rephrase this second sin, then, as confusing the Type I error probability with the probability for the model.

3.3. *Confusing Data with Hypotheses*

It is often easy to develop the perception that our data are very nearly direct measurements of what we want to infer. This seems to be true, after all, of some of the simplest data. This perception encourages one to commit a subtle statistical sin, and to confuse the sample space with the hypothesis space. As a result inferences are sometimes made without ever explicitly specifying the hypothesis space, with the consequence that important subjectivity in one's conclusions goes unrecognized.

Many astronomers seem to hold the view that the job of statistics is to "correct" the data somehow, in order to remove uncertainties and biases. Thus you will commonly see plots of data "corrected" for nonuniform detection efficiency, for example (even though no such correction may be possible without reference to a model). Or you will see tables or plots of "background-subtracted data," for since the total signal is the sum of background and signal contributions, the "signal data" must surely be the difference between the total data and a background estimate (often leading to negative signal estimates and confidence regions that extend to regions of negative signal). You will also regularly find plots of "data with their uncertainties." Finally, the data/hypothesis confusion seems to be most serious in the field of inverse problems, where data are regularly described as "blurred" or "convolved" versions of reality that must be mathematically "deconvolved" in order for us to learn the truth. It was in this field that I first faced my own confusion between data and hypotheses, when I realized that *inverse problems are inference problems*, not problems of inverting a mathematical operator.

Not all astronomers who follow these practices run astray in their conclusions. To some degree, these practices are a kind of generally accepted shorthand; and even when they might lead one astray, common sense and experience will help a good scientist make reasonable inferences, even if that scientist's concepts and tools are deficient. But too often such phrases as those just quoted are symptoms of a deeper confusion about the role of data in inference that leads to incorrect procedures and corrupted conclusions. This confusion, I believe, has its roots in two aspects of how frequentist statistics looks at data.

First, frequentist calculations explicitly average over hypothetical data. As is well known, this makes properties of the sampling distribution play a much more central role in frequentist statistics than they do in Bayesian inference. For example, the width of the sampling distribution *as a function of the data* is crucial, and it is the importance of this quantity that leads astronomers to talk routinely about the "uncertainty of the data" and to place "error bars" on plots of raw data, despite the fact that the values of the data are the only thing in the problem about which we have *no* uncertainty.

Second, the very title of the frequentist approach to inference—*statistics*—implies at the outset an emphasis on constructing and studying functions of the data (i.e., statistics), an emphasis that encourages the "corrected data" view of inference. For this reason, in my own writing I prefer the phrase "Bayesian Inference" to "Bayesian Statistics," because the emphasis of the Bayesian approach is on the logic and calculus of inference. One may end up studying functions of the data, but only as a consequence of a more fundamental analysis.

In the Bayesian approach, the quantities of interest are probabilities for hypotheses conditioned on the data. The hypotheses must be specified explicitly even just to write down the symbolic form of the quantity we want to calculate. The formalism itself, by making explicit the requirements for an inference problem to be well-posed, discourages the data/hypothesis confusion.

3.4. *Incorrect Treatment of Nuisance Parameters*

The fourth statistical sin often committed by astronomers is improper consideration of nuisance parameters. In fact, sometimes important nuisance parameters are not even explicitly introduced, a special case (in the continuum limit) of the failure to specify relevant alternatives, a sin already mentioned. For example, astronomers frequently search for counterparts to objects detected in one wavelength band in a picture taken in another band. This is often done by taking a "window" that is roughly the size of the expected image, sliding it around the second picture in the region where the image is expected, and seeing if there is significant excess intensity in the window. In such a procedure, the location of the window is a (vector-valued) nuisance parameter. Too often no consideration is made for this implicit parameter in assessing the significance of a candidate feature. I have already discussed the need to explicitly specify alternatives. Here, as in the model comparison case discussed above, the Bayesian approach forces one to recognize such parameters by forcing one to explicitly specify the hypothesis space. I will not further discuss the omission of nuisance parameters here.

Instead, I want to point out that nuisance parameters that have been explicitly recognized are often not treated correctly, despite the fact that they are straightforwardly dealt with in Bayesian calculations through marginalization. Of course, statisticians readily admit the lack of a general frequentist procedure to adequately handle nuisance parameters. This lack is why such parameters have earned the apellation, "nuisance." But I have yet to see an introductory statistics text that forthrightly admits this deficiency. Needless to say, it is thus unknown to most astronomers, who have thus been forced to invent methods of their own without realizing the limitations of those methods. Two methods are commonly used.

The simplest method sometimes used is to condition on best-fit values of the nuisance parameters. The weaknesses of this method are obvious even to astronomers, who consequently seldom use it. In particular, if the model leads to strong correlations between inferred values of the nuisance and interesting parameters, the uncertainty in the interesting parameters can be greatly underestimated by this procedure.

The more sophisticated method is the profile likelihood method, independently reinvented by astronomers aware of the problems of the conditional method just described (see, e.g., Lampton, Margon and Bowyer 1976). This method is certainly an improvement over the conditional method, in that it attempts to account for correlations. In cases where the likelihood function is a Gaussian (possibly correlated) with respect to the parameters, it gives results similar or identical to those obtained by marginalization in a Bayesian calculation. However, it maximizes rather than integrates over the nuisance parameters, and thus does not properly account for the volume of parameter space. With non-Gaussian likelihoods, profile likelihoods can easily and demonstrably lead one awry (I provided a discussion of this for astronomers in Loredo 1992).

Finally, it is worth pointing out that from the Bayesian point of view, a model comparison calculation is an inference problem in which *all* of the parameters for each model are nuisance parameters, in the sense that the global or marginal likelihood for a model (the prior predictive for the data) is obtained by integrating over the entire parameter space of each model. Such integrals give rise to the "Occam's Razor" behavior of Bayesian model comparison calculations (Jefferys and Berger 1992). There is no frequentist counterpart to this aspect of Bayesian inference.

I consider the inability of frequentist methods to handle nuisance parameters one of the most serious drawbacks of such methods. This is because it is my experience that *every real astronomical inference problem has nuisance parameters.* In every inference problem I have worked on, from the very first analysis of real data I performed as an undergraduate physics major to the analysis

I am currently working on a dozen years later, there has always been at least something like an uncertain background to contend with, and often one or more nuisance parameters considerably more complicated than an additive background. I believe many "several-sigma" detections that eventually proved spurious, and "accurate" parameter estimates that eventually proved inaccurate, owe their origin to incorrect handling of nuisance parameters. That Bayesian methods can so straightforwardly handle such realistic complication is a powerful advantage of such methods.

3.5. *Imprecise Specification of the Sample Space*

This audience need not be reminded of the disturbing sensitivity of some frequentist inferences to intuitively irrelevant features of the sample space (e.g., stopping rules). This is news to astronomers, however, many of whom are at least mildly disturbed on learning of the dependence of simple frequentist results on stopping rules (I tried to spread the news a little in Loredo 1992). For the most part, astronomers take a completely cavalier attitude toward specification of the sample space, yet another statistical sin.

In some cases, it is easy to be forgiving about this sin. After all, the stopping rule is irrelevant to most Bayesian calculations (and when it *is* relevant, it *should* be relevant!). In addition, it would often be impossible to specify precisely the sample space for an astronomical study. For example, the size of the sample can depend on such vagaries as the weather or the temperament of an unruly, complicated piece of equipment.

This being said, there are many cases in which the sin cannot be easily forgiven. These are cases where implicit nuisance parameters give rise to complications in the sample space. I earlier referred to the problem of detecting a counterpart by use of a sliding window. The window position is an unrecognized nuisance parameter. But it can be accounted for, at least to some extent, by considering each measurement taken with a different window location to be a different "sample," and adjusting the final result for the number of (possibly dependent) samples examined.

This problem arises most frequently in the search for periodic signals in astrophysical time series. One folds the data at a number of trial periods, and uses some statistic to decide whether to reject the hypothesis that the folded phases are from a uniform distribution. Most such studies (but unfortunately not all) take into account the number of periods examined: the sample space consists, not of the time series itself, but of sets of folded phases. Although one time series is examined, many sets of folded phases are examined, weakening the significance of any resulting "detection" (i.e., rejection of the uniform hypothesis). Unfortunately, other complications of the sample space sometimes go unaccounted for. For example, some models have additional implicit parameters, such as unknown offset phases or parameters specifying the statistic (such as the number of bins in a χ^2 test). These should be accounted for somehow, particularly if they have been set with reference to the data, in which case additional samples effectively have been examined. Bayesian methods bypass these complications in the sample space by dealing with these complications in the *hypothesis* space. Gregory and Loredo (1992) discuss these issues further.

3.6. *Use of Ad Hoc Statistics*

In the previous section I described the frequentist distinction. As noted there, this encourages nonexperts to take a method that they know works in one problem and apply it directly to another problem, possibly with some generalization. This takes us to the next sin: use of *ad hoc* statistics, chosen because of familiarity or intuitive appeal, but with no deeper justification.

The most common example of this sin is the use of moments of the data to estimate parameters that can themselves be interpreted as moments. This practice can also be thought of as another example of the failure to distinguish between data and hypotheses. Several examples of this sin arise in the study of gamma-ray bursters (GRBs). These are astrophysical sources of gamma rays that turn on suddenly and unpredictably for any amount of time from a few hundredths of a second to a few hundred seconds, and then fade again to invisibility. Perhaps 2000 such bursts have been observed since their discovery in 1967, but the physical nature of the sources of the bursts remains

enigmatic. Indeed, the distances to the burst sources is uncertain to many orders of magnitude! The principal information available for inferring the spatial distribution of burst sources is the observed distribution of the intensities of bursts (dim bursts are presumably farther away) and of the directions to bursts (which should mimic the distribution of stars in our Galaxy if burst sources are local to it, or be isotropic if bursts originate far from our Galaxy). One way one might parameterize the source distribution is by the dipole moment of the directions to the sources (the average of the cosines of the angles from the burst directions to some specified origin, such as the Galactic center). Some investigators estimate the source dipole moment by taking the dipole moment of the observed source directions. If bursts are sampled uniformly over the sky and if there is no uncertainty in their measured directions, there is some justification for such a procedure. But this is not the case. Similar methods have been used to estimate the angular correlation function for burst directions, and moments of the intensity distribution. Some weaknesses in these procedures have been recognized, but attempts to deal with them have themselves been *ad hoc*, involving "weighted" or "smeared" moments. Loredo and Wasserman (1994) provide a Bayesian look at the problem, derive the likelihood function for these data, and point out that moments are not sufficient statistics for these data.

Conditioned by too much emphasis on the Gaussian distribution, astronomers underestimate the possible dangers of naive use of moments for inference. The extreme case is the Cauchy distribution, where the mean of any number of samples is no better an estimator of the location parameter than any single sample. While more commonly occuring distributions may not suffer from so drastic a failure of moments or other ad hockery, the failures are still potentially rather dramatic. I have discussed this problem, and the problem of recognizable subsets that underlies it, in Loredo (1992) using a simple inference problem that arose in the analysis of the supernova neutrinos mentioned above.

In Bayesian inference, there is no freedom for ad hockery in the choice of what function of the data to use for inferences. The data enter inferences through the likelihood function, and the rules of probability theory dictate precisely how the likelihood function must be manipulated to make the desired inferences.

Ad hockery is fairly common, too, in the choice of GOF statistics for hypothesis testing. From the Bayesian point of view, the choice of statistic implicitly corresponds to a choice of relevant alternative models. As mentioned above, Bayesian model comparison calculations force one to explicitly identify relevant alternatives. The calculations then identify the functions of the data that optimally address the problem.

3.7. *Ignoring Simple Prior Information*

The final sin I will mention is the sin of ignoring simple prior information. Sometimes prior information plays a somewhat subtle role in inferences; an example is provided in the following section. Given the fact that there is no role for prior information in frequentist calculations, astronomers can perhaps be forgiven for neglecting these somewhat subtle effects.

But too often astronomers use methods that ignore prior information as simple as the requirement that an inferred intensity be nonnegative. The result is often intensity estimates or error bars that lie in regions of negative intensity. The remarkable thing about such absurd estimates is not so much that they occur, but that they are published. When the method being used gives negative estimates, this is accepted as a fact of life, rather than as evidence that the method is erroneous.

In my own field of high energy astrophysics, where light is detected by counting individual photons that are often few in number, this situation arises fairly frequently when a background rate is subtracted from a signal estimate. Negative estimates arise not only because of the neglect of prior information, but sometimes also due to to poor treatment of a nuisance parameter in cases when the background rate is uncertain. One collects photons from an "off-source" direction in order to estimate the background intensity, and then collects photons "on-source" to use to estimate the signal. The uncertain background rate is a nuisance parameter, and our prior information specifies

that neither the background nor the signal rate can be negative. The usual approach is based on subtraction of the best-fit background estimate from a signal+background estimate based on the on-source data. For weak signals, this can easily lead to negative signal estimates. I discuss this "on/off" problem in some detail in Loredo (1992). Its Bayesian solution is simple, instructive, and, once found, intuitively appealing: the signal posterior is a weighted sum of gamma distributions, with weights determined by the predictive probability for the number of background counts present in the on-source measurement, based on the information about the background provided by the off-source measurement. Cases that lead to negative signal estimates with the usual method are handled easily; the posterior usually peaks at zero signal intensity, but vanishes for negative intensities. Of course, in such cases the precise shape of the posterior can be somewhat sensitive to the shape of the prior. But this is simply the calculation's way of telling us that the data are uninformative.

This exhausts my list of sins. I will close this section by emphasizing again a point made above: these mistakes are often made by otherwise excellent astronomers. This pairing of great scientific and mathematical talent with statistical sloppiness is evidence of serious defficiencies, not only in the statistical education of astronomers, but also in the tools of frequentist statistics themselves. For only if the tools invite misuse can we explain their misuse by scientists who so successfully master other mathematical tools on their own, at reasonably high levels of sophistication.

4. AN EXAMPLE

I earlier quoted from Kleppner's *Physics Today* editorial. After "fretting" about statistics for several paragraphs, he continues as follows:

> Having raised some reservations . . . it is only fair for me to point out that statistical analysis has been crucial to more than a few dazzling discoveries. The anisotropy in the cosmic background radiation recently reported by the Cosmic Background Explorer [COBE] team is a case in point. . . . . (Kleppner 1992)

This case—actually, a simplified version of it—will be the topic of this section.

The COBE satellite is one of NASA's "Great Observatories;" two others whose names you might recognize from recent coverage in the press are the Hubble Space Telescope and the Compton Gamma Ray Observatory. The primary goal of the COBE mission is to make detailed measurements of properties of the Cosmic Background Radiation, the fading glow of the hot, early phase of the evolution of our universe known colloquially as the "big bang." The two most important properties COBE measures are the spectrum of the radiation, and its anisotropy (how its properties vary with direction across the sky). Before discussing the data and its analysis, let us briefly review the cosmic background radiation to establish the scientific motivation for studying its spectrum and anisotropy, and to make clear the historic importance of the COBE data.

4.1. *The Cosmic Background Radiation*

It is a well-established observational fact that distant galaxies appear to be moving away from us, with speeds proportional to the distances between the galaxies and our own Milky Way Galaxy. The universe is expanding, in the sense that the distances between galaxies are growing. Turning the clock backward, the average density in the universe at earlier times must have been larger than it is now. Together with the laws of gravity and of thermodynamics, the observed expansion implies that the universe has a finite age—between 10 and 20 billion years—and that at earlier times it was not only more dense, but also hotter than it is at the present time.

Running the equations that describe our evolving universe back to within a few centuries of its birth, we find conditions that were so hot and dense that matter was broken up into its constituent parts. At these early epochs, the universe was almost entirely a plasma of electrons and protons exchanging energy with electromagnetic radiation. The energy exchange was very efficient because electrons and protons are electrically charged particles, and thus easily interact with electromagnetic

radiation. But eventually the universe expanded and cooled enough that electrons and protons could bind to form atomic hydrogen and atoms of other elements. When this happened, matter and radiation decoupled from each other, for although atoms are made up of charged particles, as a system an atom is electrically neutral and interacts much more weakly with radiation than do free charged particles.

As the universe continued to expand, it cooled further, and atoms eventually combined to form molecules, stars, galaxies—and us. All this while, the radiation from the hot, early epochs of the universe's history propagated with very little interaction with matter, uniformly cooling (growing in wavelength) with the expansion of the universe but otherwise not changing. This radiation, then, provides us with a "picture" of what the universe looked like a few hundred thousand years after its origin billions of years ago. It should be visible from all directions as a diffuse background against which the stars and galaxies appear, and is thus called the Cosmic Background Radiation (CBR).

This radiation was expected to have a spectrum (distribution in energy) like that of a "blackbody"—a perfect absorber and emitter of radiation—because of the efficiency of exchange of energy before the epoch of decoupling. Before the COBE observations, there were tantalizing hints that the CBR spectrum was *not* that of a blackbody, suggesting that there might have been a time after decoupling when the universe was reionized (perhaps by an exploding early population of stars), so that matter and radiation once again interacted strongly, changing the blackbody spectrum that was produced just before the earlier decoupling. The first significant finding of the COBE mission was that the CBR spectrum *is* that of a blackbody to an extraordinary degree of precision (approximately 0.03%).

As already noted, the CBR provides us with a picture of the universe at the time of decoupling, long before galaxies formed. This picture cannot be featureless, because we know that galaxies *did* form. Some parts of the early universe must have been denser or hotter or otherwise different from their surroundings in a manner that would distinguish them as future sites for clusters of galaxies, and this difference must have left some mark on the CBR which should be visible today as anisotropy in its appearance. For over two decades, this mark of the inhomogeneity of the early universe had been sought by observers, to no avail. Simultaneously, theorists tried to predict the appearance of the features. As more and more sensitive observations set tighter and tighter limits on the strength of the features, it became increasingly difficult to construct theoretical models for the growth of structure in the universe that could account for the presently observed structure from perturbations small enough to escape detection.

At the time of the launch of the COBE mission, cosmology was on the brink of a crisis. Observations had pushed the limit on the amplitude of perturbations of the CBR down to less than one hundredth of one percent. Increasingly careful and clever theoretical calculations had pushed the predicted perturbation amplitude down to a few *thousandths* of one percent, below the observed limit, but within the sensitivity of the COBE detectors. It seemed essentially impossible that any theory could account for presently observed structure with smaller perturbations. If the COBE experiment did not see the long sought for perturbations, it would force major changes in our understanding of the early universe.

It took a year for the COBE experiment to map the CBR over the entire sky, but the results were worth waiting for. The perturbations were detected, at roughly the level predicted by current theories of the formation of large scale structure. Observers and theorists alike hailed the observations as among the most important in decades.

It is the analysis of the CBR anisotropy data that we will discuss in some detail here. In the next subsection, I will describe the salient features of the COBE anisotropy data. For the sake of brevity and clarity, I am forced to omit many details that are very important in the analysis of the actual data. Thus let me emphasize at the outset that all numbers and conclusions I am presenting are merely illustrative of the actual results one might obtain from a full analysis of the actual data. If and when more data become public, more complicated and accurate analyses may become possible. This idealized analysis is adequate, however, to raise some methodological issues, and to illustrate some of the statistical malpractice alluded to in earlier sections.

4.2. *Measuring the CBR Anisotropy*

Theoretical calculations predict much more than simply a typical amplitude (in a root-mean-square sense) for the CBR anisotropy. They predict how the amplitudes of perturbations should depend on their angular sizes. Competing theories may predict similar *rms* perturbation amplitudes, but very different distributions of amplitude with angular size. The COBE experiment provides information on perturbations of all sizes from $\sim 7^\circ$ to the size of the entire sky. It thus provides us with an opportunity to compare rival theories quantitatively, and to infer the values of free parameters that are present in most theories. Such inferences must be statistical in nature, not only because there is noise present in the data, but because the predictions themselves are statistical. Theories of the evolution of large scale structure are deterministic, but they require that one specify initial conditions to be evolved by the deterministic equations of motion. We do not know the initial conditions; the best we can do is assign a probability distribution over the possible initial conditions. Hence the dual role of statistics in the study of CBR anisotropy.

As noted above, the CBR has a blackbody spectrum. The shape of this spectrum (as a function of the frequency, energy, or wavelength of the radiation) is completely specified by one parameter, the CBR temperature, T. The sky-averaged temperature of the CBR is $\bar{T} \approx 2.7$ K. The anisotropy of the CBR can be described simply by considering the CBR temperature to be a function of direction, $\mathbf{n}$. Our task is to make inferences about $T(\mathbf{n})$ based on the observational data.

A blackbody spectrum with a temperature of 2.7 K peaks at wavelengths of about 1 mm, in the microwave region of the electromagnetic spectrum. The instrument used to detect such radiation is called a *radiometer*. It consists of a horn-shaped antenna that collects radiation incident from a small patch of the sky and funnels it to a detector that measures the amount of energy incident on it within a narrow wavelength band, and within a small integration time, τ. Electronics then convert the deposited energy to a voltage and eventually to a digital signal.

One can model the relationship between the signal and the sky temperature as follows. Let $F(T)$ denote the power (energy per unit time) that would flow perpendicularly through a unit area whose normal pointed to a blackbody of temperature T, in the wavelength band of the detector. This is a nonlinear function of T that one can calculate from the known form of the blackbody spectrum. Let $A(\mathbf{n})$ denote the area presented by the antenna to radiation from direction $\mathbf{n}$. Then, in the absence of noise, the amount of energy we expect the detector to observe is,

$$\langle E \rangle = \tau \int d\mathbf{n}\, A(\mathbf{n}) F[T(\mathbf{n})]. \tag{4.1}$$

I have written this as an expectation, because the power observed from a blackbody is actually "noise power:" even in the absence of other noise sources, the observed power will fluctuate on repeated measurement. These "thermal fluctuations" can be described by a Gaussian distribution, with the standard deviation for the fluctuations given by a simple function of the temperature, wavelength range, and integration time.

The detector electronics are designed to produce a signal that, in the absence of noise, is proportional to the energy deposited in the detector. However, the physics of the detection process introduces an inevitable large offset (essentially, every detector introduces thermal power into the measurement because it has a nonzero temperature of its own). Thus the expected signal can be written,

$$\langle S \rangle = O + g\tau \int d\mathbf{n}\, A(\mathbf{n}) F[T(\mathbf{n})], \tag{4.2}$$

where O is the offset, and g is a factor giving the conversion from deposited energy to output signal amplitude; g is called the detector gain.

Now recall that the relative size of the perturbations we are measuring is of the order of 0.001%. That is, if we write

$$T(\mathbf{n}) = \bar{T} + \delta T(\mathbf{n}), \tag{4.3}$$

then $\delta T/T \lesssim 10^{-5}$. We can take advantage of this to linearize equation (4.2) in terms of δT. To $\mathcal{O}(\delta T^2)$, we can write

$$F(T) = F(\bar{T}) + \delta T F'(\bar{T}). \tag{4.4}$$

Let $\bar{A}$ denote the sky-averaged area presented by the horn. Then equation (4.2) can be written,

$$\langle S \rangle = O + 4\pi g\tau \bar{A} F(\bar{T}) + g\tau F'(\bar{T}) \int d\mathbf{n}\, A(\mathbf{n}) \delta T(\mathbf{n}). \tag{4.5}$$

Thus the expected signal is linear in the temperature perturbation, $\delta T(\mathbf{n})$.

The actually observed signal includes noise, and the feasibility of using a radiometer to measure very small δT values depends on the size of the noise. The thermal fluctuations inherent in blackbody radiation are the dominant noise source in radiometers, allowing straightforward calculation of the expected size of the noise. For the COBE parameters, it takes roughly an hour or two of integration time to make the thermal noise contribution small enough to ensure sky temperature measurements accurate to 0.001%. To adequately sample the entire sky, such observations must be performed in $\sim$ 6000 directions, requiring on the order of a year of observing time. Due to the motion of the spacecraft (and other factors), data for one direction must be obtained by accumulating many short (0.5 s) observations spread out over several months of observations.

Unfortunately, the offset O in typical radiometers is as large or much larger than the part of the signal proportional to the deposited energy, and it cannot be maintained constant to a level of 0.001% of the sky contribution over such a long duration. A standard physicist's trick for measuring small perturbations in a signal contaminated by such offset drift is to make *differential* measurements: record the difference between the signal of interest and a reference signal using measurements taken on a short enough timescale that gains and offsets are effectively constant. The difference signal has the offset removed. The situation is not much improved if a constant reference is not available. In the COBE experiment, this problem is alleviated by using the *sky* itself as a "reference." The *Differential Microwave Radiometer* (DMR) on the COBE spacecraft consists of *two* identical horns pointing 60° away from each other. Both horns feed the same detector through a switch that switches from one horn to the other many times a second. In this way a signal proportional to the temperature difference of two directions separated by 60° can be produced. A particular direction is measured many times during a year's observation, with all other directions 60° away used as"reference" signals. In this way substantial information about δT can be compiled, although information about the actual temperature, $\bar{T} + \delta T$, is destroyed by the differential measurement.

Thus the DMR data is a set of numbers, d_i, each of which we can model by taking the difference of two equations like equation (4.5), with different area functions but with the same offset and gain, and with added noise, n_i;

$$d_i = G \int d\mathbf{n}\, [A_i^+(\mathbf{n}) - A_i^-(\mathbf{n})] \delta T(\mathbf{n}) + n_i. \tag{4.6}$$

Here we have defined a new detector gain by $G = g\tau F'(\bar{T})$, and A_i^+ and A_i^- denote the area functions for the two horns contributing positively and negatively to measurement number i. They are identical, up to a 60° rotation; we used this fact to cancel two terms from the contribution of $\bar{T}$ in each horn.

To make inferences about $\delta T(\mathbf{n})$, we must parameterize it somehow. A useful parameterization is in terms of the coefficients of an expansion in terms of spherical harmonics. One reason this expansion is useful is that the horns are rotationally symmetric, and thus the area functions can be written as a simple sum of Legendre polynomials. Combined with a spherical harmonic expansion of $\delta T(\mathbf{n})$, this allows us to simplify the integral in equation (4.6). Also, there is some theoretical motivation for adopting such a parameterization. The simplest and most popular theories of the formation of large scale structure postulate "Gaussian initial conditions," by which is meant harmonic coefficients for which *a priori* probabilities are independent and Gaussian, with zero mean and "cosmic variances"

that vary with harmonic index in a manner specified by the theory. A harmonic expansion is thus the natural parameterization for studying such theories, as it allows one to specify priors in a simple manner.

Thus we write

$$\delta T(\mathbf{n}) = \sum_{l=1}^{\infty} \sum_{m=-l}^{l} a_{lm} R_{lm}(\theta, \phi). \tag{4.7}$$

Here θ and ϕ are the polar angle and azimuth in some fixed coordinate system, and the a_{lm} are the expansion coefficients. The R_{lm} are *real* spherical harmonics. Just as the familiar complex spherical harmonics (usually denoted Y_{lm}) are products of modified Legendre polynomials in $\mu = \cos\theta$ and a complex azimuth factor, $e^{im\phi}$, the R_{lm} are products of modified Legendre polynomials and a *real* azimuthal factor,

$$R_{lm}(\theta, \phi) = K_{lm} P_{lm}(\mu) \operatorname{cas}(m\phi), \tag{4.8}$$

where K_{lm} is a normalization constant given by

$$K_{lm} = \left[\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!} \right]^{1/2}, \tag{4.9}$$

and $\operatorname{cas}\phi$ denotes the "cosine and sine" function,

$$\operatorname{cas}\phi = \cos\phi + \sin\phi. \tag{4.10}$$

The $\operatorname{cas}\phi$ function is the basis of the Hartley transform (Bracewell 1986), a real, orthogonal transform similar in many respects to the Fourier transform, but more appropriate for the analysis of real quantities. Thus real spherical harmonics merely substitute the Hartley basis for the Fourier basis used to describe the azimuthal behavior of complex spherical harmonics. It is easy to show that each R_{lm} function is a linear combination of $Y_{lm'}$ functions with the same l and with $m' = \pm m$, and vice versa. This implies that the R_{lm} form a complete orthonormal basis with properties similar to those of the Y_{lm}. Since $\delta T(\mathbf{n})$ is a real function, there are practical advantages in using the R_{lm}. In particular, there are no constraints on the coefficients (apart from those arising from positivity of T); in a Y_{lm} basis, coefficients with the same l and $|m|$ must be related to guarantee that δT is real.

As previously noted, the area functions are azimuthally symmetric. Thus the A^+ function for a particular sample can be written in terms of the cosine of the angle from the positive horn's symmetry axis, γ^+, as follows:

$$A^+(\mathbf{n}) = \sum_l \frac{2l+1}{4\pi} w_l P_l(\cos\gamma^+). \tag{4.11}$$

To simplify later equations, we have taken a $(2l+1)/4\pi$ factor out of the expansion coefficients, w_l. The exact same equation, with the same coefficients, holds for A^-, except that it is in terms of Legendre polynomials in the cosine of the angle, γ^-, from the negative horn's axis.

The angle addition theorem lets us write the expansion for each horn in terms of angles in any other chosen coordinate system using spherical harmonics. Let $\mathbf{n}^+$ denote the colatitude and azimuth of the positive horn's axis in any chosen spherical coordinate system. Then the theorem lets us rewrite equation (4.11) as,

$$A^+(\mathbf{n}) = \sum_l w_l \sum_{m=-l}^{l} R_{lm}(\mathbf{n}) R_{lm}(\mathbf{n}^+), \tag{4.12}$$

with a similar equation holding for A^-. Using these expansions, the δT expansion of equation (4.7), and the orthonormality of the real spherical harmonics, equation (4.6) can be written,

$$d_i = G \sum_l w_l \sum_m a_{lm} [R_{lm}(\mathbf{n}_i^+) - R_{lm}(\mathbf{n}_i^-)] + n_i. \tag{4.13}$$

This is our key equation; we want to make inferences about the a_{lm} (or about parameters specifying the a_{lm}) using this model equation. We will first discuss some Bayesian inferences, and then discuss some published frequentist inferences.

4.3. *Some Bayesian Inferences*

The data will enter Bayesian inferences about the a_{lm} through the likelihood function, which is simply a product of independent Gaussians for each datum. The standard deviations of these Gaussians will be nearly equal, to the extent that the sky temperature and offset are constant; for simplicity we presume the standard devations to be identical, and denote their value by σ. Then the likelihood function for the a_{lm} is the exponential of a quadratic form,

$$\mathcal{L}(a) \propto \exp\left(-\frac{s}{2\sigma^2}\right), \tag{4.14}$$

with

$$s = \sum_{i=1}^{N}\left(d_i - G\sum_{lm} a_{lm} w_l [R_{lm}(\mathbf{n}_i^+) - R_{lm}(\mathbf{n}_i^-)]\right)^2. \tag{4.15}$$

This is a likelihood function for parameters of a linear model, and thus varies with the a_{lm} as a correlated Gaussian. The formal manipulations with which we can rewrite this to make its Gaussian form (as a function of a_{lm}) explicit are probably too familiar to this audience to deserve much comment. I will partly write them out, however, to facilitate some comments on the practical feasibility of carrying out the necessary linear algebra.

To simplify the appearance of the equations, I will adopt three abbreviations. First, I will use vector symbols to denote all N components of quantities with a data index, i; for example $\mathbf{d} \equiv \{d_i\}$. Second, I will use a single greek letter index, α, to denote jointly the parameter indices l and m (note, then, that w_α will be the same for all α corresponding to the same l). Third, I will combine the factors multiplying the coefficients into the symbol,

$$\Delta R_{\alpha i} = G w_\alpha [R_\alpha(\mathbf{n}_i^+) - R_\alpha(\mathbf{n}_i^-)]. \tag{4.16}$$

With these abbreviations, the quadratic form can be written,

$$\begin{aligned} s &= \left(\mathbf{d} - \sum_\alpha a_\alpha \Delta\mathbf{R}_\alpha\right)^2 \\ &= \mathbf{d}\cdot\mathbf{d} + \sum_\alpha\sum_\beta a_\alpha a_\beta \eta_{\alpha\beta} - 2\sum_\alpha a_\alpha \mathbf{d}\cdot\Delta\mathbf{R}_\alpha, \end{aligned} \tag{4.17}$$

where the "model metric" $\eta_{\alpha\beta}$ is given by,

$$\eta_{\alpha\beta} = \Delta\mathbf{R}_\alpha \cdot \Delta\mathbf{R}_\beta. \tag{4.18}$$

We can now "complete the square" to write s as the sum of a "perfect square" in the a_α and a residual term that is independent of a_α. We can do this by taking the "square root" of the metric, either by finding its Cholesky decomposition, or more usefully by diagonalizing it. This procedure will automatically identify sufficient statistics for estimating the a_α coefficients. I will not bother with the details, which are well-known in both the Bayesian and frequentist literature (a good Bayesian treatment of similar problems, written by a physical scientist, is that of Bretthorst 1988). But it is worth thinking about the practical problems that will arise in the calculations.

We must calculate the model metric, and diagonalize it. Calculation of a single $\eta_{\alpha\beta}$ element involes a dot product in the data space (i.e., a sum over the N data). For the COBE data, $N \sim 10^6$. In principle, we need an infinite number of such sums, since the harmonic expansion formally contains an infinite number of terms. But the DMR horns have finite resolution, and thus provide little information about harmonics with large l values. Such a large l cutoff arises in our equations through the w_l factors, which vanish exponentially at large l. We can estimate the largest accessible order as follows. Harmonics of order l have a characteristic angular scale of approximately π/l. The

horns have an angular resolution of 7° or 0.12 radian. Thus harmonics with orders larger than $l \approx 25$ are not well-resolved by the DMR. We can therefore cut our harmonic sum at $l \approx 25$, corresponding to $M \approx 700$ coefficients (since there are $2l + 1$ coefficients for each l). Although not infinite, this remains a distressingly large number, as there are $M^2/2$ elements in η, each requiring calculation of an N-fold sum. The total operation count is thus $\sim 10^{11}$. Once calculated, we need to diagonalize η; but even with an algorithm whose work scales like M^3, this calculation is dwarfed by the actual calculation of η itself. Thus some sophistication is needed to perform the calculations. To the extent that the observations evenly cover the sky, the orthogonality of the harmonics comprising the $\mathbf{\Delta R}_\alpha$ may make the η matrix quite sparse, enabling calculation with sparse matrix techniques. The extent to which this is true has not yet been studied (the data do not evenly cover the sky, complicating such a study). Monte Carlo methods may offer an alternative, approximate approach for performing the needed calculations. We will skip these issues here, noting that similar issues have arisen in frequentist analyses of these data and have been somewhat cleverly dealt with, giving promise that headway can be made with a Bayesian calculation (if perhaps only an approximate one).

We will now presume that we can manipulate $\mathcal{L}(a)$—either analytically or through clever computation—to make its dependence on the coefficients accessible to us, so we can proceed with useful inferences. Here we will constrain ourselves to inferences about the $l = 2$ (quadrupole) coefficients only. Our motivation is partly one of simplicity; this will be enough to illustrate some important points of methodology. But the quadrupole coefficients are interesting scientifically as well, and measurements of the quadrupole moment of δT were among the most publicized early results of analyses of the DMR data. This is because an important nuisance parameter—Earth's velocity with respect to the CBR, which induces a large CBR dipole moment through the Doppler effect—renders the cosmological contribution to the measured dipole moment uncertain, so the quadrupole is the lowest order, largest scale cosmological perturbation that can be reliably measured.

We further presume that the likelihood function's dependence on each a_{2m} is independent of the values of coefficients of other orders and of other m values. We do this, not only for simplicity, but because it has been implicitly or explicitly presumed in virtually all of the frequentist studies of the DMR data. To simplify our notation, we use the symbols $q_m = a_{2m}$ to denote the quadrupole coefficients in the remainder of this section, and an unadorned q to denote the set of all five coefficients.

Given all of these idealizations, I will henceforth refer to the data as originating from an *Idealized* DMR (IDMR) instrument, to distinguish it from the actual DMR data (which are not publicly available in raw form). Our idealizations imply that the quadrupole factor in the IDMR likelihood can be written as a function of q as follows:

$$\mathcal{L}(q) = \prod_{m=-2}^{2} \frac{1}{\sigma_m \sqrt{2\pi}} \exp\left[-\frac{(q_m - \hat{q}_m)^2}{2\sigma_m^2}\right], \tag{4.19}$$

where $\hat{q}_m$ and σ_m are sufficient statistics that arise from the linear algebra described above. They contain all the information the data provide about the quadrupole moment, so at times I may refer to them as "the quadrupole data" (I will always distinguish them from the values of the actual quadrupole components, however!). For the analysis presented here, I will set the IDMR $\hat{q}_m$ and σ_m equal to values corresponding to those reported for the actual DMR (Smoot et al. 1992; note that this work uses a different—and less convenient—harmonic basis than that used here). These values are listed in Table 1; Figure 1 provides a graphical display of these sufficient statistics. The normalization of equation (4.19) is arbitrary; I have chosen it to simplify some later calculations. So long as our prior does not couple coefficients of different l, all other coefficients can be marginalized, leaving a marginal likelihood for q given by equation (4.19). (It is perhaps worth emphasizing that theories couple moments of different l by predicting relationships between their cosmic variances; but we will focus on the $l = 2$ coefficients by themselves here, without explicitly considering a particular theory.)

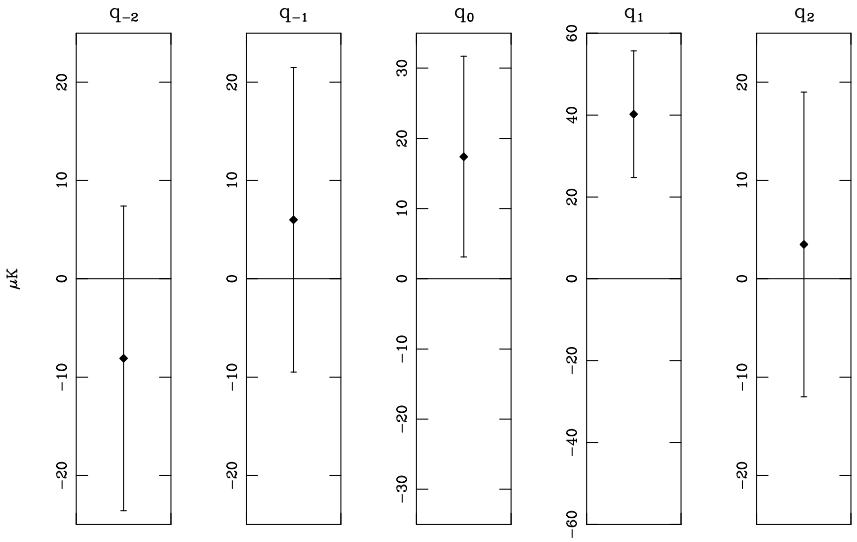

Figure 1. Graphical display of the IDMR sufficient statistics for the quadrupole moment listed in Table 1.

TABLE 1

IDMR Sufficient Statistics For Quadrupole Moment

m	$\hat{q}_m$	σ_m
-2	-8.1	15.5
-1	6.0	15.5
0	17.4	14.3
1	40.2	15.5
2	3.5	15.5

We will ask three questions: (1) Do these data indicate the presence of a quadrupole anisotropy in the CBR? (2) Presuming such an anisotropy, what is its magnitude? (3) Presuming Gaussian initial conditions, what is the cosmic variance associated with quadrupole perturbations? It is easy to distinguish these questions in the Bayesian framework; indeed, the framework forces one to distinguish them. As we will see, this is not so in frequentist calculations.

To determine whether a quadrupole anisotropy is detected, we must perform a model comparison calculation, comparing the no-quadrupole ($q_m = 0$) model, M_0, with an alternative model M_Q allowing a quadrupole. The rules of Bayesian inference dictate precisely how this comparison must be done: we must calculate the probabilities of the competing models. Since there are two models, it is easiest to summarize the results in terms of the odds in favor of one model over the other. Denoting the data by D, and our background information specifying the models under consideration by I, the odds in favor of the quadrupole model over the no-quadrupole model is,

$$\begin{aligned} O &= \frac{p(M_Q \mid D, I)}{p(M_0 \mid D, I)} \\ &= \frac{p(M_Q \mid I)}{p(M_0 \mid I)} \, \frac{p(D \mid M_Q)}{p(D \mid M_0)}. \end{aligned} \tag{4.20}$$

The first factor is the prior odds ratio, a measure of our subjective preference for one model over

another. The second factor is the Bayes factor, B, specifying the implications of the data and the description of the models. It is a ratio of the "global likelihoods" for the models. The global likelihood for M_0, in the denominator, is the probability for the data presuming all q_m are actually zero, so that the measured values are due to noise. This probability is simply equal to $\mathcal{L}(0)$. The global likelihood for M_Q is given by,

$$p(D \mid M_Q) = \int dq\, p(q \mid M_Q)\, \mathcal{L}(q). \tag{4.21}$$

Its value necessarily depends on what this model asserts about the quadrupole amplitudes through the prior placed on q. This feature of Bayesian model comparisons is sometimes troubling to astronomers, since results of frequentist model comparison methods (such as those based on likelihood ratios) typically do not depend on prior information. It is useful, however, to recall an important exception widely known to astronomers. In assessing the significance of a periodic signal in an astrophysical time series, it is widely appreciated that a correction must be made for the number of periods searched. This is the frequentist counterpart for the effect the prior range for the period would have in a Bayesian model comparison. The Bayesian contention is that such "corrections" for the size of the parameter space are necessary *whenever* there are unknown parameters in one or more of the models being compared, and which we must infer from the data.

We will use a prior over a range of q_m values constrained from above by previous observations and from below by theory. The prior should take into account the fact that there is no preferred orientation for the coordinate system we are using to study the CBR. Thus, for example, the joint prior for the five q_m coefficients should not be the product of five independent bounded priors; the resulting hypercubical prior volume is not invariant with respect to rotations. An appropriate invariant to use to bound the prior volume can be found from examining the root mean square temperature fluctuation, whose value is independent of the coordinate system orientation:

$$\begin{aligned} \delta T^2_{\rm rms} &= \frac{1}{4\pi} \int d\mathbf{n}\, \delta T^2(\mathbf{n}) \\ &= \frac{1}{4\pi} \sum_{l=1}^{\infty} \sum_{m=-l}^{l} a^2_{lm}. \end{aligned} \tag{4.22}$$

The $l = 2$ contribution to this sum defines the rms quadrupole fluctuation,

$$Q^2_{\rm rms} = \frac{1}{4\pi} \sum_{m=-2}^{2} q^2_m. \tag{4.23}$$

Observers with different coordinate systems will agree on $Q_{\rm rms}$, even though they will assign different q_m values. We can define a radius coordinate in the five dimensional space spanned by the quadrupole coefficients as follows;

$$r = \left[\sum_{m=-2}^{2} q^2_m \right]^{1/2}. \tag{4.24}$$

The invariant quadrupole moment is simply proportional to this "quadrupole radius," with $Q_{\rm rms} = r/\sqrt{4\pi}$. To be consistent with the symmetry of the problem, the prior must depend on the q_m components only through the combination r: $p(q \mid M_Q) = f(r)$.

We will assign a prior that is flat with respect to $Q^2_{\rm rms}$. This prior will *not* be flat with respect to the q_m. We can find the relationship between the $Q^2_{\rm rms}$ and q priors as follows:

$$\begin{aligned} p(Q^2_{\rm rms} \mid M_Q) &= \int dq\, p(Q^2_{\rm rms}, q \mid M_Q) \\ &= \int dq\, p(Q^2_{\rm rms} \mid q, M_Q)\, p(q \mid M_Q) \\ &= \int dq\, \delta \left(Q^2_{\rm rms} - \frac{r^2}{4\pi} \right) f(r). \end{aligned} \tag{4.25}$$

The volume element in the 5-dimensional q space can be written in terms of dr and a 4-dimensional solid angle element, $d^4\Omega$, as $dq = r^4 dr\, d^4\Omega$. The integrand of equation (4.25) depends only on r. Thus the integral over $d^4\Omega$ is trivial; it is the 4-area of a unit 4-sphere, equal to $8\pi^2/3$. Using this result, and rewriting the δ-function in terms of r, we find,

$$\begin{aligned} p(Q_{\rm rms}^2 \mid M_Q) &= \frac{8\pi^2}{3} 2\pi \int dr\, \delta\left(r - \sqrt{4\pi} Q_{\rm rms}\right)\, r^3 f(r) \\ &= \frac{128\pi^{9/2}}{3}\, Q_{\rm rms}^3\, f(\sqrt{4\pi} Q_{\rm rms}). \end{aligned} \tag{4.26}$$

Thus for the prior to be flat with respect to $Q_{\rm rms}^2$, we require a nonuniform q prior with $f(r) \propto r^{-3}$ to cancel the $Q_{\rm rms}^3$ factor arising from the volume element. Normalized over the $Q_{\rm rms}^2$ interval $[Q_{\rm lo}^2, Q_{\rm hi}^2]$, the prior is,

$$p(q \mid M_Q) = \frac{3}{16\pi^3 (Q_{\rm hi}^2 - Q_{\rm lo}^2)}\, \frac{1}{r^3} \tag{4.27}$$

when $Q_{\rm rms}^2$ is in the prior range, and $p(q \mid M_Q) = 0$ otherwise.

Note that a prior that is flat with respect to $Q_{\rm rms}^2$ is not flat with respect to $Q_{\rm rms}$. They are related by

$$\begin{aligned} p(Q_{\rm rms} \mid M_Q) &= p(Q_{\rm rms}^2 \mid M_Q) \frac{dQ_{\rm rms}^2}{dQ_{\rm rms}} \\ &= 2 Q_{\rm rms}\, p(Q_{\rm rms}^2 \mid M_Q) \\ &= \frac{256\pi^{9/2}}{3}\, Q_{\rm rms}^4\, f(\sqrt{4\pi} Q_{\rm rms}). \end{aligned} \tag{4.28}$$

Thus the flat $Q_{\rm rms}^2$ prior implies a linear $Q_{\rm rms}$ prior; alternatively, a flat $Q_{\rm rms}$ prior implies a $1/Q_{\rm rms}$ behavior for the $Q_{\rm rms}^2$ prior. There is ambiguity over which choice is most natural. I have chosen the flat $Q_{\rm rms}^2$ prior because this prior implies a form for the marginal priors for the individual q_m coefficients that is more nearly flat than that implied by a prior flat with respect to $Q_{\rm rms}$. This way our prior is roughly uniform both for the invariant quantity, $Q_{\rm rms}^2$, and for the directly measured quantities, q_m.

For a sample calculation, let us set $Q_{\rm hi} = 300\,\mu$K, corresponding to to the previous observational upper limit of $Q_{\rm rms}/T \approx 10^{-4}$, and $Q_{\rm lo} = 3\,\mu$K, corresponding to a theoretical lower limit of $Q_{\rm rms}/T \approx 10^{-6}$ (we can set this lower limit to zero with a negligible change in our results). The integral in equation (4.21) can be performed very simply with Monte Carlo integration, using just a few lines of `FORTRAN` and a few minutes of CPU time. The resulting Bayes factor is $B \approx 0.003$. The data favor a model with *no* quadrupole fluctuation; only a prior odds ratio greater than about 300:1 in favor of M_Q will make the posterior odds exceed unity. Given the data of Figure 1, it is not very surprising that a zero quadrupole model is favored, but the smallness of B may be somewhat surprising. The Bayes factor is roughly inversely proportional to $Q_{\rm hi}^2$. As we will see shortly, the data imply that $Q_{\rm rms} \lesssim 20\,\mu$K. Even if we were to "cheat" and use this *a posteriori* knowledge to assign a prior upper limit of $Q_{\rm hi} = 20\,\mu$K, the Bayes factor increases only to $B \approx 0.7$, corresponding to ambivalence between the quadrupole and no-quadrupole models. The IDMR data simply do not provide compelling evidence for a nonzero quadrupole moment.

It is instructive to consider an alternative quadrupole model, M_Q', that implies a flat prior over q. Normalized, this prior is

$$p(q \mid M_Q') = \frac{15}{256\pi^{9/2}}\, \frac{1}{Q_{\rm hi}^5 - Q_{\rm lo}^5}. \tag{4.29}$$

If we again set $Q_{\rm lo} = 3\,\mu$K and $Q_{\rm hi} = 300\,\mu$K, we find the Bayes factor in favor of this model over M_0 is extremely small, $\approx 6 \times 10^{-7}$. The reason for this has to do with the geometry of the 5-dimensional

q space. The volume in a shell of fixed thickness grows like $Q_{\rm rms}^4$. Thus a flat q prior, assigning probability proportional to volume, places almost all of its probability near $Q_{\rm hi}$, which we shall see is quite far from the value preferred by the data. For example, the flat prior assigns nearly 97% probability to the region beyond $Q_{\rm rms} = 150\,\mu$K; but the posterior assigns similar probability to the region below $20\,\mu$K. This geometric effect makes the Bayes factor quite sensitive to the upper limit for the M_Q' model; it is roughly proportional to $Q_{\rm hi}^{-5}$. These results emphasize the importance of taking into account the physical meaning of the parameters in assigning a prior.

Our conclusion that the IDMR data do not provide compelling evidence for a nonzero quadrupole moment is otherwise not too sensitive to the shape of the prior, provided that the volume factors are accounted for. For example, the information provided by previous experiments indicates a preference for small $Q_{\rm rms}^2$ values but does not really impose a sharp cutoff at $Q_{\rm hi}^2$. Let us instead use a smoothly decaying prior with, say, 95% of its probability below $Q_{\rm hi}^2$,

$$p(Q_{\rm rms}^2 \mid M_Q) = \frac{\log(20)}{Q_{\rm hi}^2} \exp\left(-\log(20)\frac{Q_{\rm rms}^2}{Q_{\rm hi}^2}\right). \tag{4.30}$$

This is quite a different prior from a flat prior over $Q_{\rm rms}^2$. However, when $Q_{\rm hi} = 300\,\mu$K, it changes the Bayes factor only from 0.003 to 0.009; and when $Q_{\rm hi} = 20\,\mu$K, it changes it from 0.7 to 0.95. Our conclusions are thus robust with respect to such changes in the prior.

Regardless of the data, most astronomers believe there must be *some* quadrupole anisotropy, albeit a small one. In essence, most of us have a prior odds in favor of M_Q that is large, based on physical reasoning. Thus it is useful to presume a (possibly vanishing) quadrupole moment is present, and infer its magnitude. This is a parameter estimation problem, but it is not well-posed until we specify the parameterization of interest. If we wish to infer the amplitudes of the quadrupole components, we can simply multiply $\mathcal{L}(q)$ by a prior to find the joint posterior. Using the flat prior on $Q_{\rm rms}^2$, the posterior for the individual quadrupole components is,

$$p(q \mid D, M_Q) \propto \frac{\mathcal{L}(q)}{[\sum_m q_m^2]^{3/2}}. \tag{4.31}$$

The prior modulates the Gaussian shape of the likelihood and introduces a weak correlation between the components. But the posterior is not drastically different from the likelihood. For example, the solid curve in Figure 2*a* shows the marginal posterior for q_1, and the dashed curve shows its Gaussian factor in $\mathcal{L}(q)$ (i.e., the inference we would make with a flat prior on the q_m). They differ noticably, but not drastically, because the Gaussian varies rapidly enough with q_m that multiplication by any function with characteristic scale larger than σ_m has little effect on the shape of the posterior. It is interesting to note that this change of prior, which has little affect on parameter estimates, had a large effect on the model comparison calculation, changing the Bayes factor by several orders of magnitude. This is a common characteristic of Bayesian calculations: parameter estimates are usually much less sensitive to changes in the prior than are model comparisons.

Figure 2*b* shows the marginal posterior for the largest detected moment, q_1. This posterior is more significantly shifted by the prior, which somewhat favors small values for the q_m. Still, the shift is of the order of a standard deviation; large enough to notice, but not large enough to qualitatively change one's scientific conclusions.

It is more useful to consider a different parameterization, and to infer the value of $Q_{\rm rms}^2$, since this provides a measure of the quadrupole amplitude that is independent of orientation. After all, the naturalness of this parameter is the reason we are using a nonuniform prior for the q_m. The easiest way to calculate the posterior for $Q_{\rm rms}^2$ is to "extend the question" as we did in equation (4.25): introduce the q_m as auxiliary parameters and integrate them out. This lets us write

$$p(Q_{\rm rms}^2 \mid D, M_Q) = \int dq\, p(Q_{\rm rms}^2, q \mid D, M_Q)$$

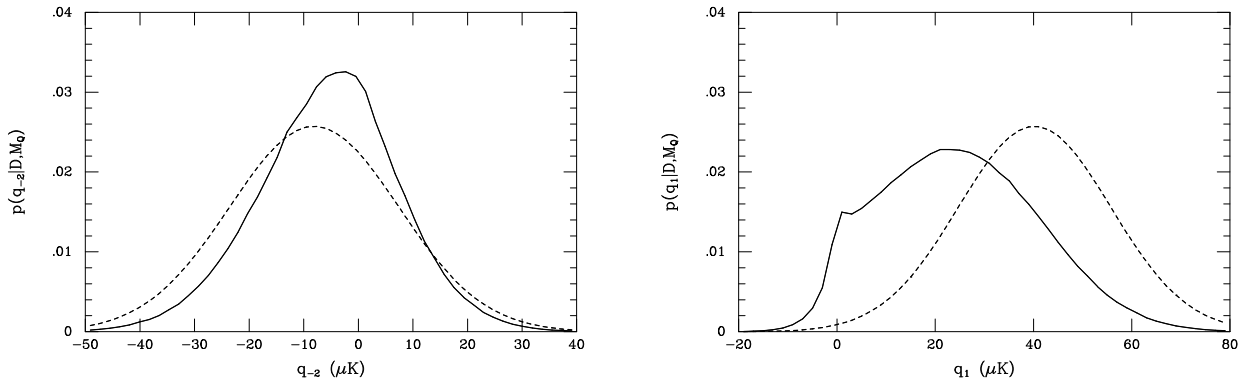

Figure 2. (*a*) Marginal posterior for q_{-2} based on a prior flat with respect to $Q_{\rm rms}$ (solid curve), and on a prior flat with respect to all q_m (dashed curve). (*b*) As in (*a*), but for q_1.

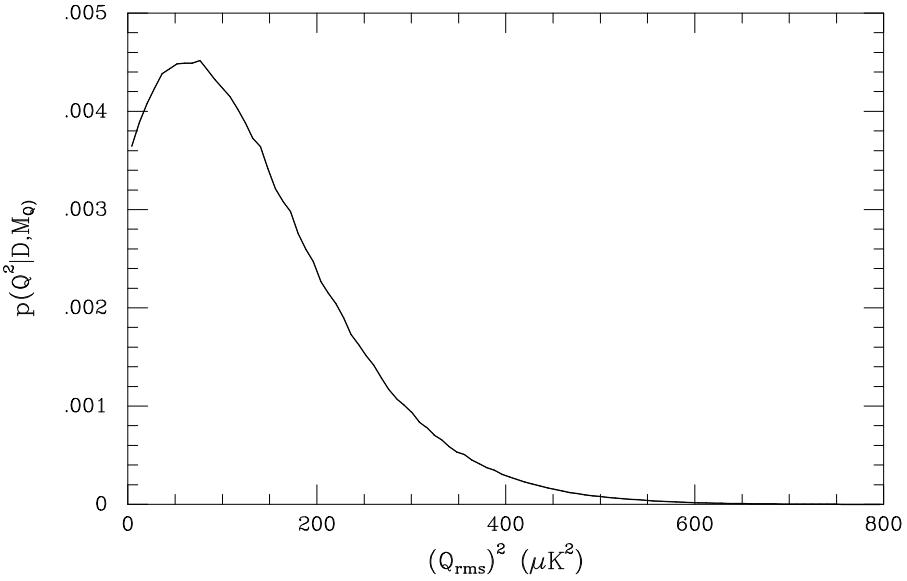

Figure 3. Marginal posterior for $Q^2_{\rm rms}$ using a prior flat with respect to $Q^2_{\rm rms}$.

$$= \int dq\, p(Q^2_{\rm rms} \mid q, D, M_Q)\, p(q \mid D, M_Q)$$
$$= \frac{3}{16\pi^3 (Q^2_{\rm hi} - Q^2_{\rm lo})} \int dq\, \delta\left(Q^2_{\rm rms} - \frac{r^2}{4\pi}\right) \frac{\mathcal{L}(q)}{r^3}. \qquad (4.32)$$

This is vaguely like the integral defining the noncentral χ^2 distribution, and someone more familiar with that distribution than I am may be able to make some analytical headway with this integral. It is easy enough to evaluate it with Monte Carlo methods, however. The result is shown in Figure 3. The posterior mean is $\langle Q^2_{\rm rms}\rangle = 138\,\mu{\rm K}^2$, and the posterior standard deviation is $\sigma_{Q^2} = 103\,\mu{\rm K}^2$. All highest posterior density credible regions (hereafter simply "credible regions") containing 60% or more of the posterior probability contain the point $Q^2_{\rm rms} = 0$. This certainly seems intuitively consistent with our earlier finding that no quadrupole moment is necessary to account for these data.

Again it is instructive to see what would have happened had we used the flat prior over the q_m. The solid curve in Figure 4 shows the resulting posterior for $Q^2_{\rm rms}$. It vanishes at $Q^2_{\rm rms} = 0$, and gives a mean $\pm$ standard deviation estimate for $Q^2_{\rm rms}$ of $93 \pm 59\,\mu{\rm K}^2$. It may come as a surprise

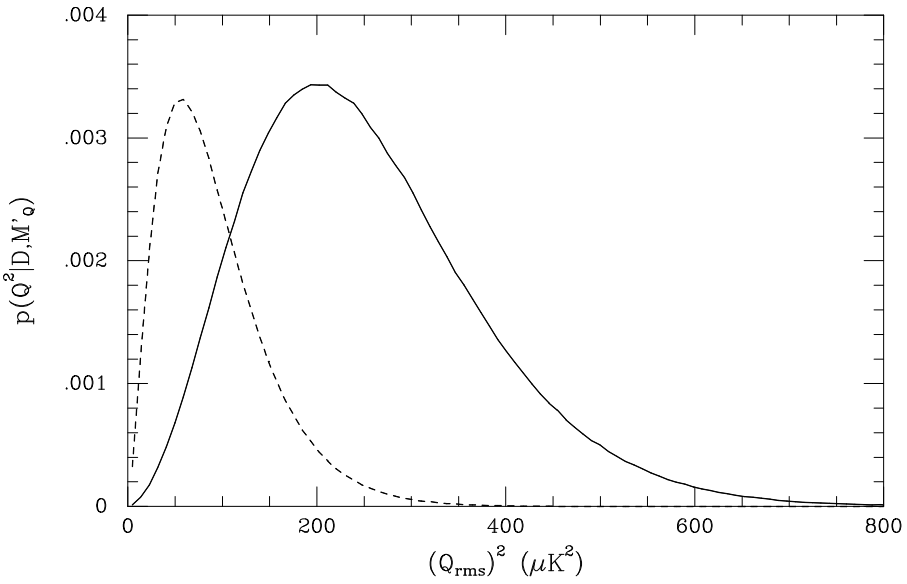

Figure 4. Posterior for $Q^2_{\rm rms}$ using a prior flat with respect to q_m. Solid curve is based on the IDMR data of Table 1; dashed curve is based on data with the same σ_m, but with all $\hat{q}_m = 0$, and is multiplied by 0.4.

to find the posterior vanish at $Q^2_{\rm rms} = 0$ and to find the inferred $Q^2_{\rm rms}$ value well over a standard deviation away from zero, given that we have just shown that the data favor the no-quadrupole model over the M'_Q model we are considering here with odds greater than 10^6:1. Again, geometry is to blame. As already noted, the volume of a shell of thickness $dQ_{\rm rms}$ grows very quickly, as the fourth power of $Q_{\rm rms}$. There is thus vanishingly little volume near the origin; one can show the posterior *must* vanish at $Q^2_{\rm rms} = 0$ simply because of these volume effects, even if the best-fit values of the q_m are all zero, so long as their uncertainties do not vanish. To make the point, the dashed curve in Figure 4 shows the posterior found by replacing the measured q_m values with zero, but keeping the uncertainties given in Table 1. We comment further on how the geometric effects can help us reconcile the estimation and model comparison calculations below.

Finally, we can ask the question that is perhaps the most interesting to a cosmologist: What do the data tell us about theories for the intial conditions? Recall that most theories postulate Gaussian initial conditions, meaning that we can model the a_{lm} as being drawn from independent Gaussian distributions with zero means and "cosmic variances," C_{lm}, specified by theory. The absence of a preferred coordinate system implies that, for all spherically symmetric theories, the cosmic variances cannot depend on m: $C_{lm} = C_l$. The quadrupole components thus represent five samples from a zero mean Gaussian with variance C_2, so we might hope that our measurements of them will allow us to make useful inferences about the magnitude of C_2. By Bayes's theorem, the posterior for C_2 given the data, D, and prior information I is,

$$p(C_2 \mid D, I) \propto p(C_2 \mid I)\, p(D \mid C_2, I). \tag{4.33}$$

We can calculate the likelihood conditional on C_2 by "extending the question" again, as follows:

$$\begin{aligned} p(D \mid C_2, I) &= \int dq\, p(D, q \mid C_2, I) \\ &= \int dq\, p(D \mid q, C_2, I)\, p(q \mid C_2, I) \\ &= \int dq\, \mathcal{L}(q) \prod_{m=-2}^{2} \frac{1}{\sqrt{2\pi C_2}} \exp\left(-\frac{q_m^2}{2C_2}\right). \end{aligned} \tag{4.34}$$

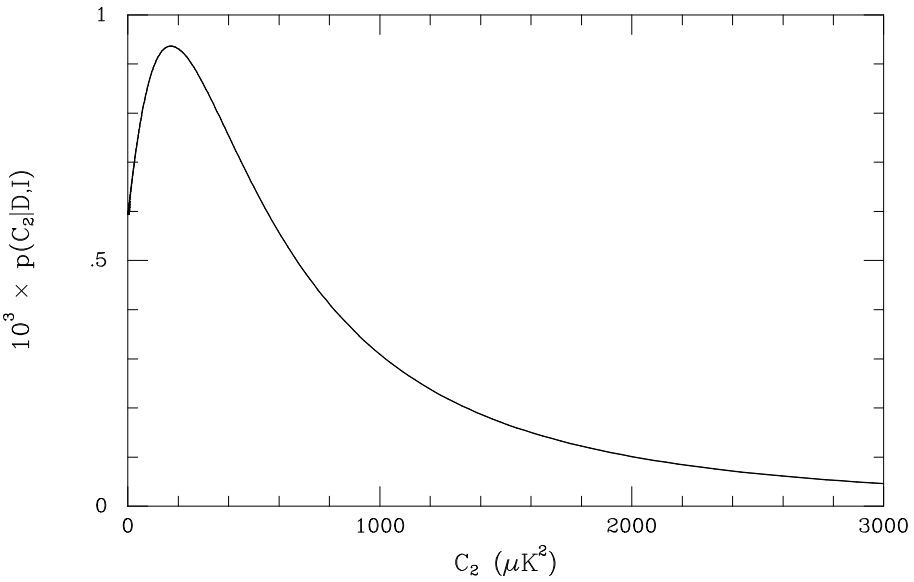

Figure 5. Posterior for quadrupole cosmic variance C_2.

The integral can easily be done analytically once we complete the squares in the q_m. The result is,

$$p(D \mid C_2, I) \propto \prod_{m=-2}^{2} \frac{1}{\sqrt{C_2 + \sigma_m^2}} \exp\left[-\frac{\hat{q}_m^2}{2(C_2 + \sigma_m^2)}\right]. \tag{4.35}$$

That is, the likelihood for C_2 is what we would get by treating the measured q_m values as being samples from zero mean Gaussians with variances equal to the *sum* of the cosmic variance and the noise variance. Multiplying by a flat prior and normalizing gives the posterior plotted in Figure 5. Note that it does *not* vanish at $C_2 = 0$; in fact, the point at $C_2 = 0$ is inside credible regions as small as 50%. Examining equation (4.35), we see the posterior has a power-law tail proportional to $C_2^{-5/2}$. This is reflected in the large disparity between the locations of the posterior mode at $C_2 = 169\,\mu\mathrm{K}^2$ and the posterior mean, $\langle C_2 \rangle = 1860\,\mu\mathrm{K}^2$. The 68% credible region extends from 0 to 1200 $\mu\mathrm{K}^2$; the 95% credible region extends to 6100 $\mu\mathrm{K}^2$. These results are consistent with our model comparison results in that they imply that the data are consistent with zero cosmic variance, i.e., no actual quadrupole anisotropy. While the data offer possibly useful upper limits on C_2, the power-law behavior of the tails probably makes the precise value of upper limits somewhat sensitive to the prior. This sensitivity, and the consistency with $C_2 = 0$, are indications that the data tell us little about the quadrupole moment of the CBR.

4.4. *Some Frequentist Analyses*

Now let me describe some frequentist analyses of these IDMR data, of the kind reported in analyses of the actual DMR data (Janssen and Gulkis 1992; Smoot et al. 1992). Let me emphasize at the outset that I am not offering these results as the best that frequentist statistics has to offer, but rather as typical of what highly talented astronomers do based on their limited familiarity with frequentist methods.

All published analyses of the COBE anisotropy data are based on "sky maps" of $\delta T(\mathbf{n})$ constructed as follows. The sky is descretized into 6144 direction pixels; at the time of each DMR measurement, each horn is pointing in one of these pixels. We label the pixels with a single index; the pixel to which a horn points is a function of the datum index, i. The model equation (4.6) is rewritten as,

$$d_i = G(\tilde{T}_a - \tilde{T}_b) + n_i. \tag{4.36}$$

Here $\tilde{T}_a$ is the *antenna temperature* in the direction of pixel $a = a(i)$. It is defined by

$$\tilde{T}_a = \frac{\int d\mathbf{n} A_a(\mathbf{n}) T(\mathbf{n})}{\int d\mathbf{n} A_a(\mathbf{n})}, \tag{4.37}$$

where $A_a(\mathbf{n})$ denotes the area function of a DMR horn when its axis is in the direction indexed by a. A χ^2 statistic is formed from the data,

$$\chi^2 = \sum_i \frac{[d_i - G(\tilde{T}_a - \tilde{T}_b)]^2}{\sigma^2}. \tag{4.38}$$

Minimizing this with respect to the 6144 antenna temperatures leads to a normal equation for the antenna temperatures with a very sparse design matrix. A sparse matrix algorithm is used to solve for the antenna temperatures, up to an arbitrary offset. Each pixel is assigned a Gaussian uncertainty with variance given by the diagonal component of the covariance matrix of the solution. All subsequent analyses are performed using these maps as the "data." Coefficients in a spherical harmonic expansion are found by fitting the 6144 antenna temperatures to a sum of spherical harmonics, and dividing the resulting a_{lm} coefficients by the w_l coefficients describing the horn area functions.

The original reports of a detection of a CBR quadrupole component with the real DMR (Smoot et al. 1992) reported $\hat{q}_m$ and σ_m values corresponding to the IDMR values given in Table 1. The reported detection was based on the following calculation. An estimate of $Q_{\rm rms}$ was found simply by setting $q_m = \hat{q}_m$ in equation (4.23). A standard deviation was found using a method that was not described; but the results are consistent with the "propagation of errors" procedure well-known to physicists (see Bevington 1969):

$$\sigma_Q^2 = \sum_m \left(\frac{\partial Q_{\rm rms}}{\partial q_m} \right)^2 \sigma_m^2. \tag{4.39}$$

These calculations give $Q_{\rm rms} = 13 \pm 4\,\mu$K. Since the estimate is over three times the size of its uncertainty, this was deemed evidence that a significant quadrupole component was detected. No attempt to infer the cosmic quadrupole variance independent of higher order moments was made (theories predict relationships between the C_l coefficients, so it is possible to infer C_2 from fits to all *higher order* multipoles, which is what was done).

4.5. *Discussion*

It is hard for someone who spends lots of time thinking carefully about data analysis to resist flying off the handle at least a little bit in response to the data mutilations just described. So before I lose control, let me make some positive remarks. The COBE team consists of some of the most talented experimental physicists in the world. The care they have taken in the design and construction of their experiment, and in the analysis of systematic sources of uncertainty (none of which I have discussed here) sets a high standard that more astronomers should aspire to. There is little doubt in my mind that the DMR instruments detected significant anisotropy in the CBR, as is evidenced by the high significance of the DMR measurements of higher order moments than the quadrupole. Thus my following criticisms, calling into question the DMR quadrupole detection, detract only slightly (if at all) from the great historical significance of this experiment.

In addition, the DMR team has now analyzed a second year of data, and reports increased significance for their results. New methods were used in the analysis of the second year's worth of data, and although I do not like some of what I have seen of these methods, I have not studied them in any detail and cannot intelligently comment on them.

Finally, as noted above, my IDMR analysis is merely illustrative of what *might* be true of a careful analysis of the first-year DMR data. I do not have access to the raw data, and so cannot

say with certainty what the actual data imply for the CBR quadrupole (and I may not be able to do the required calculations even if I had the data!). Nevertheless, I believe the above analysis calls into question the significance of the reported quadrupole detection, and may point to problems that could affect other inferences based on the DMR data as well. Given the decades of effort spent constructing the COBE instruments and the years spent calibrating them and removing systematic effects, it seems a shame to compromise the valuable information provided by the data by using faulty methodology. Hopefully this discussion will motivate further consideration of the methodology used to analyze this data.

[I cannot resist noting here that, after this paper was prepared, the COBE team published an analysis of the second-year DMR data, including a reanalysis of the first-year data. New methods were used for inferring the quadrupole, although considerable ad hockery remains. However, they conclude that there is no significant evidence for a nonzero quadrupole in the second-year data, and that a joint analysis of both years' worth of data shows only a marginally significant detection (90% significance).]

As noted above, all analyses of the DMR data are based on sky maps produced by solving normal equations for the antenna temperatures in various directions. This is an excellent example of the confusion of data with hypotheses. The object of the DMR measurements is to measure the sky temperature, so it is natural for an astronomer to try to convert the data into something that looks like what we want to infer—a map of the sky temperature. As natural as this procedure may appear to some astronomers, it is far from obvious that fits to the resulting map, even after correction for the area functions of the horns, give results corresponding to what one would find from rigorous modeling of the raw data. To the extent that correlations between the inferred antenna temperatures comprising the sky map can be ignored, the linearity of the procedure may lead to an approximate correspondence between rigorous fitting and fitting to sky maps with subsequent beam corrections. But this has not been demonstrated; indeed, the issue has not even been recognized.

The estimation of $Q_{\rm rms}$ provides another example of the confusion of data with inferences: Harmonic fits provide estimates of q_m, and $Q_{\rm rms}$ can be written in terms of q_m, so it seems natural to estimate $Q_{\rm rms}$ simply by plugging in the q_m estimates. Inference is more subtle than this, however, particularly when nonlinear equations (like that relating $Q_{\rm rms}$ and the q_m) are involved, and when there are numerous nuisance parameters (for the $Q_{\rm rms}$ inference, the q_m are essentially nuisance parameters). The need to integrate over several nuisance parameters makes the consideration of prior information about the geometry of the q_m space quite important for this problem; it has been ignored in all published analyses. The nonlinearity of the $Q_{\rm rms}$ equation makes estimation of $Q_{\rm rms}$ more complicated than simply plugging in $\hat{q}_m$ estimates, even when we ignore volume effects: the estimate using the flat q prior is $Q_{\rm rms} = 15.5 \pm 3.9\,\mu$K, about 0.6 standard deviations larger than the frequentist estimate that in some sense implicitly uses this prior. It is difficult to be forgiving about this error, because most physicists encounter estimation using variance-weighted measurements in undergraduate lab courses. Surely one of the lessons of this well-known frequentist technique is the need to distinguish between estimation of terms in an equation like equation (4.23) and estimation of the value of the equation itself. Presuming the $\hat{q}_m$ estimates to be accurate, accounting for the nonlinearity and prior information replaces the $13 \pm 4\,\mu$K frequentist estimate with the posterior shown in Figure 3. It implies $Q_{\rm rms}$ estimates lower than the frequentist estimate, with a standard deviation wider than the frequentist σ_Q. More importantly, this posterior is not Gaussian and is obviously not well summarized by its mean and standard deviation; in particular, $Q^2_{\rm rms} = 0$ is inside the 68% credible region, even though the posterior mean for $Q^2_{\rm rms}$ is well over one standard deviation away from zero.

Gould (1993) noted the intuitive discrepancy between the claim of a very significant "3σ" quadrupole detection and the fact that the $\hat{q}_m$ values in Figure 1 do not appear too significantly different from zero. He outlined a crude χ^2 calculation that seemed to verify the intuitive consistency with no quadrupole, causing him to examine the $Q_{\rm rms}$ estimation procedure. Gould, too, found fault with the original $Q_{\rm rms}$ estimation procedure. Unfortunately, he noticed only that the original estimator

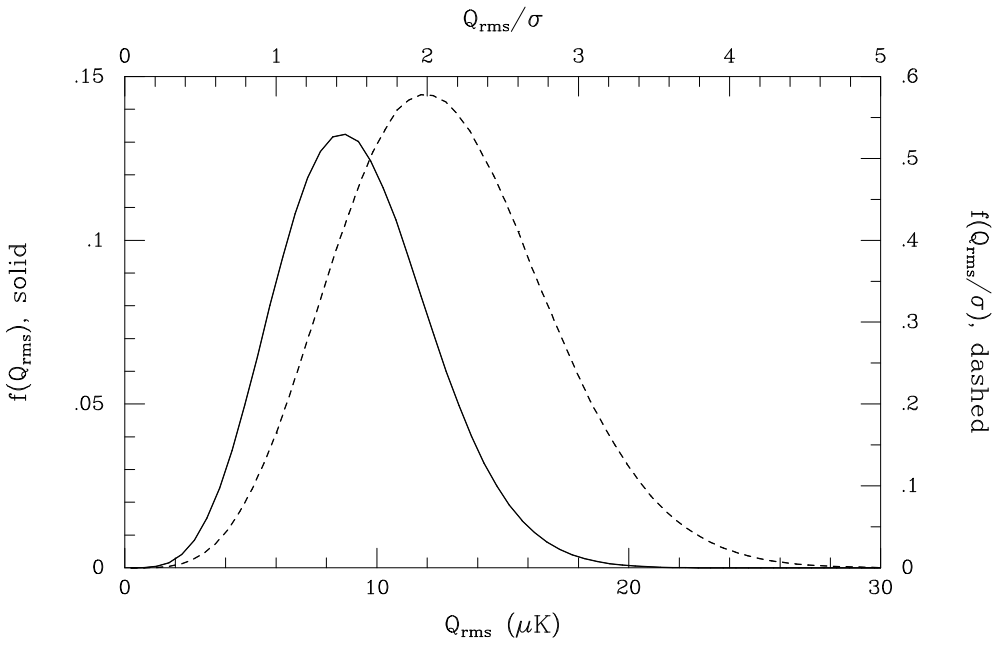

Figure 6. Frequency distribution of frequentist estimates of $Q_{\rm rms}$ based on data with $\hat{q}_m = 0$. Solid curve shows distribution of $Q_{\rm rms}$ estimates (left and bottom axes); dashed curve shows distribution of $Q_{\rm rms}$ divided by its propagated error (right and top axes).

was biased, and replaced it with an *ad hoc* unbiased estimator. This new estimator decreased the estimate to $10.5\,\mu$K, but it remained over "2σ" larger than zero, essentially because the volume effects are still unaccounted for by this estimator. But this estimator ignores even more basic prior information than the volume of q space: as Gould himself noted, it will produce imaginary $Q_{\rm rms}$ estimates for some samples, ignoring our prior knowledge that temperatures are real-valued (i.e., that $Q^2_{\rm rms} \geq 0$).

All of these investigators failed to distinguish between estimation and model comparison, yet another of the "sins" alluded to in the previous section. They use a parameter estimate to decide whether the estimated parameter is required by the data by seeing whether it is at least a couple standard deviations from its null value. Although it is true that some linear Bayesian model comparison calculations are amenable to an interpretation in terms of finding a new parameter a significant distance away from a null value (see Loredo 1989), this is merely an *interpretation* of a calculation that is fundamentally very different from a parameter estimation calculation. That a simple interpretation is not always possible is borne out by the Bayesian calculations that use the M'_Q model assigning a flat q prior. The resulting posterior distribution for $Q^2_{\rm rms}$ *vanishes* at the origin, yet a model comparison calculation very strongly favors the model with *no* quadrupole moment over this model. We described the geometrical origins of this apparent contradiction above; similar effects arise in the frequentist calculation. Figure 6 demonstrates this. I drew many samples of measured q_m coefficients from Gaussians with *zero means* and standard deviations given by σ_m. Thus all of these samples are from a parent population with *no* actual quadrupole. The curve in Figure 6 shows the distribution of $Q_{\rm rms}$ estimates obtained using the frequentist procedure of plugging the $\hat{q}_m$ values into the $Q_{\rm rms}$ equation (the unbiased procedure would often yield imaginary estimates for these data). The dashed curve shows the distribution of the number of standard deviations the estimate is from zero. The estimate is always nonzero; very often (over 50% of the time) it is more than two standard deviations away from zero. Again, this is simply a consequence of the geometry of the five-dimensional q_m space. It emphasizes the need to distinguish model comparison and parameter estimation.

Finally, there are many nuisance parameters involved in the CBR analysis beyond the few I have already mentioned. There are other cosmic backgrounds that contaminate the measurements, as do various instrumental and environmental effects (such as Earth's magnetic field). These are incorporated into the DMR analysis through linear modeling in the creation of the sky map. Essentially, best-fit amplitudes for a number of functions are found, and the amplitudes are fixed at these best-fit values for all subsequent analysis. I doubt that a proper accounting of these nuisance parameters would change the results substantially (it is likely only to slightly broaden the likelihood), and I doubt that the required marginalizations are possible (although the linearity may prove me wrong). But it is worth pointing out that no mention has been made that the procedure of conditioning on best-fit values is approximate. And more importantly, this is further evidence for my hypothesis, stated earlier, that no real astronomical inference problem lacks nuisance parameters!

5. TEACHING ASTRONOMERS BAYESIAN INFERENCE

Hopefully the preceding sections have convinced readers with a Bayesian bent that they should share their expertise with astronomers. In this section, I would like to offer some brief words of advice on how best to reach us. I must emphasize that this is advice for Bayesian statisticians *who wish to reach astronomers*. I will criticize some aspects of the Bayesian literature here. My advice and criticism should not be construed as directed toward all of the Bayesian literature. What and how you write to communicate amongst yourselves is your own business. But an astronomer dipping into the Bayesian literature encounters several stumbling blocks. Some of these criticisms apply to literature of previous decades more than to the present, but I offer them lest some be tempted to backslide!

5.1. *Avoid the Philosophical Aura*

Perhaps the toughest obstacle to acceptance of Bayesian methods by astronomers is the philosophical—indeed, almost religious—aura that too often accompanied Bayesian polemics in the past. Those advocating Bayesian methods spoke of coherence and consistency and axiomatics. They called themselves "Bayesians," as if they were the members of a religion founded by the Reverend Thomas Bayes. They even spoke of being "converted" from their frequentist ways.

I must confess to being guilty of all of these behaviors. My first paper on the application of Bayesian methods to astrophysical problems (Loredo 1990) began with a lengthy discussion of axiomatics, consistency, and rationality. It is true that some astronomers (myself among them) find such an approach to the foundations of Bayesian inference not only persuasive, but beautiful, in much the same manner that the most profound physical theories are beautiful. However, these astronomers comprise a small minority. The majority of the astronomical community could care less about such arguments. They want to see *calculations*, and see how they differ from what they already use; interest in the "why" behind the calculations will come later (if at all). I must admit that I find the disregard that astronomers, and physicists more generally, have for conceptual arguments distressing. Physics, after all, was originally a branch of philosophy: *natural* philosophy. But this disregard for conceptual foundations is a fact of life.

Thus you should avoid axiomatics and discussion of such concepts as coherency when trying to explain Bayesian methods to astronomical colleagues. They may well soon develop an interest in such things, but only after your facility with equations convinces them that you can not only argue, but actually solve problems. If you do end up discussing foundational matters, you would do well to become familiar with the work of Cox (1946, 1961) and Jaynes (1957, 1995)—both physicists—on the foundations of Bayesian inference. Their emphasis on internal consistency, and on limiting consistency with Boolean algebra, has greater appeal to physicists than do arguments based on coherency and betting, in my experience. Their work deserves more attention from statisticians.

Finally, I think we would be wise to stop referring to ourselves as "Bayesians." The term almost inevitably draws a curious and skeptical grin from physicists who are not themselves experienced with Bayesian methods. After all, people who use Lax-Wendorf methods to numerically solve differential equations do not call themselves "Lax-Wendorfians." Methods are Bayesian, not people. On the one hand, not one of us is perfectly Bayesian in all our inferences. On the other hand, we all know that good frequentists are really in-the-closet Bayesians, in a state of denial! Thus statements like, "I am a Bayesian," not only polarize and lend an almost religious aura to Bayesian statistics, but also are factually inaccurate. We develop and use *methods* that are Bayesian. Perhaps the word "Bayesian" is already so loaded that a new term is required, even in reference to methods rather than worldviews. Jaynes (1990, 1995) has suggested "probability theory as logic," which besides being new, has the additional virtues of being both precise and accurate. The word "logic" does lend the phrase an unfortunately philosophical tone, however.

5.2. *Emphasize Inference, Not Decision*

Physicists report evidence, and let readers make their own decisions. Of course, we often report our own conclusions and interpretations, but we must always provide the evidence that led us to the conclusions, and not merely the conclusions themselves. Thus we are more interested in the calculus of inference than in that of decision. Much early Bayesian literature emphasized decision too strongly (particularly texts), although it is my perception that this emphasis has declined in recent years. Astronomers have enough trouble accepting the subjectivity of inference that is made explicit by priors. Decision introduces additional subjectivity in the assignment of loss functions or utilities. This argues against an emphasis on decision over inference.

5.3. *Avoid Measure Theory*

If we are to benefit from your knowledge, you must communicate it to us in a language we can understand. Most of us have a reasonable level of mathematical sophistication. But almost none of us knows what a Borel subset or a sigma algebra is! Following Jaynes (1984), I would argue that we would be wasting our time and effort were we to learn the language of measure theory. We eventually perform our calculations in the comfortably discrete and finite domain of a computer; the generality of measure theory is irrelevant there. This is true even of so-called "nonparametric" Bayesian statistics (which is really "mega-parametric" statistics), where measure theory is most often used. In cases of practical interest, measure-theoretic language does not add rigor to a description of a method, only generality; a generality of little or no use to physical scientists.

When I read (or try to read) some of the Bayesian literature couched in measure-theoretic language, I feel like a `FORTRAN` or `C++` programer who has been handed a page of code in assembly language to debug. Sure, there are things you can do in assembly language that you can't do with `C++`. But if you never want to do those things, there is no advantage to be gained by learning and programming in assembly language. And there is much to be lost in terms of compatibility and accessibility.

5.4. *Emphasize Methodological Distinctions*

Say the word "Bayesian" to an astronomer, and it is likely that the first (and only) remarks you will get with any technical content will refer to priors. The common perception is that the essential distinguishing feature of Bayesian inference is the use—indeed, the *requirement*—of priors in Bayesian calculations. This perceived primacy of priors is strongly encouraged by the Bayesian literature, which so often and so strongly emphasizes priors.

Achilles might as well boast about his heel!

This emphasis encourages those with no deep understanding of Bayesian methods to believe that Bayesian results will differ from frequentist results only when there is strong prior information, and that otherwise the distinction between the approaches is merely a philosophical one involving the interpretation of probability. That such a belief is false needs no demonstration to this audience.

I believe the emphasis on priors is misplaced. It is true that the presence of priors in Bayesian calculations often provides them with substantial advantages over their frequentist counterparts which too often ignore even the simplest prior information, to their peril. But it is also true that priors introduce subjectivity into Bayesian calculations. This subjectivity is real and *should* be explicitly accounted for; but it is still a complication, and raises some practical difficulties. But more to the point, priors are *not* the essential distinguishing feature of Bayesian methods.

The essential distinguishing feature of Bayesian methods is that they average or integrate over the hypothesis space, in contrast to the averaging or integrating over the *sample* space that is the basis of frequentist methods. Priors are in some ways merely part of the "price" we have to pay in order to reap the benefits to be had from replacing sample averages with hypothesis averages. These benefits are important and numerous, and include the ability to handle nuisance parameters, the "Occam's Razor" effect in model comparisons, the ability to ignore stopping rules, safety from recognizable subsets, etc.. I have attempted to emphasize this aspect of Bayesian calculations in my more recent papers (see, e.g., Loredo 1992).

The nice thing about an emphasis on this distinction (besides it being the correct emphasis from a fundamental viewpoint!) is that it is *methodological*. It tells astronomers that they actually have to write different computer codes to do Bayesian calculations, not merely multiply existing results by a prior and reinterpret them. It also gives them a clear idea about how they must change their calculations. Finally, it makes it more obvious that Bayesian and frequentist results can differ even when there is no important prior information, since the two approaches actually perform different calculations. Realizing there *will* be a difference encourages us to *do the calculation*, and judge the methods by their results, rather than by argument.

The recent emphasis on Monte Carlo methods in the Bayesian computation literature facilitates explanation of the methodological distinction to astronomers. Many astronomers have expertise in performing frequentist calculations using Monte Carlo methods to draw samples from the sample space. The new Monte Carlo and Markov chain Monte Carlo methods for Bayesian calculation thus bear some similarity to methods already familiar to astronomers, the primary distintion being that the relevant samples are drawn from the hypothesis space rather than the sample space. Use and discussion of these methods thus may help clarify the distinction between Bayesian and frequentist calculations.

5.5. *Emphasize Robustness*

Granting that the role of priors should not be too heavily emphasized, neither should we merely brush aside discussion of priors. But such discussion should be placed in its proper context. The subjectivity arising from ambiguity in the prior is only one aspect of the inherent subjectivity of inference, a subjectivity that manifests itself in several places in both frequentist and Bayesian calculations, from the specification of the hypothesis and sample spaces to the assignment of a sampling distribution. Indeed, the results of a calculation are often much more sensitive to changes in the likelihood function than to changes in the prior. Familiarity has conditioned us to quick acceptance of "standard" sampling distibutions. This is not entirely bad. But a balanced mixture of skepticism and comfort born of familiarity should underly our attitudes towards both sampling distributions *and* priors.

Rather than seek the "correct" prior or the "correct" likelihood, we should seek to formulate a *well-posed problem* that we can solve, and to study the robustness of the solution to ambiguity in the problem specification. The robustness study need not be at the level of formality of Berger (1984). Informal "twidling" with the prior and likelihood (of the kind performed in the previous section) is usually sufficient, and a little experience goes a long way in giving one an intuitive sense of the robustness of a result. Finally, if a result is *not* robust to reasonable changes in the problem specification, this itself should be recognized as important information. If the answer depends sensitively on something we do not know, then we do not know the answer. This hardly appears

to me to be a deep insight. That Bayesian calculations can identify such states of ignorance is a benefit, not a drawback, of the Bayesian approach.

6. BAYESIAN INFERENCE IN ASTROPHYSICS

I am not alone in feeling optimism that adoption of Bayesian methods can help cure astronomy and other physical sciences of some of the statistical sloppiness that is too prevalent in these disciplines. In the same forum in which Kleppner discussed scientists' skepticism of sophisticated statistics, the reknowned condensed matter theorist Philip Anderson recently extoled the virtues of Bayesian inference:

> These statistics are the correct way to do inductive reasoning from necessarily imperfect experimental data. What Bayesianism does is to focus one's attention on the question one wants to ask of the data: It says, in effect, How do these data affect my previous knowledge of the situation? (Anderson 1992)

Anderson goes on to discuss Bayesian model comparison calculations, describing how integration over the parameter space gives rise to the "Occam's Razor" effect mentioned above, penalizing models for the size of their parameter space. Anderson sees such Bayesian benefits as an antidote for "misuse" of phrases like "significant at the 0.05% level" in the scientific literature.

But even more encouraging than Anderson's optimism is the fact that a growing number of astronomers are using Bayesian methods in their work. Accordingly, I will end this paper with a brief description of some recent applications of Bayesian inference to astronomical data analysis, as evidence of the interest of astronomers in an alternative to the frequentist methods familiar to them.

Bayesian thinking was first explicitly introduced into modern astronomical data analysis in the context of inverse problems. In particular, the maximum entropy method for "deconvolving" images was founded on Bayesian principles (see, e.g., Gull and Daniell 1979 for an early "unashamedly" Bayesian introduction to these methods). In retrospect, such problems were probably not the best to use to introduce Bayesian thinking to astronomers. The parameter spaces in these problems are huge, and their size precluded a fully Bayesian treatment. Integrals over the parameter space were impossible to perform, so only the MAP (Maximum A Posteriori) estimator could be found. As a result, such methods fostered the misconception that Bayesian inference was simply frequentist statistics, with an additional prior factor. A dogmatic attitude toward the choice of prior further fostered misconceptions about the role of priors in Bayesian calculations. Still, maximum entropy methods produce visually impressive deconvolutions, and they have become a mainstay in the analysis of astronomical images. Although they did not use the entropy prior, Morrow and Brown (1988) applied similar ideas to the inversion of helioseismology data.

More recently, the growth of computing power and of familiarity with Bayesian methods is helping the Bayesian approach to inverse problems finally begin to live up to its full potential. Investigators now have a more flexible attitude towards priors (Gull 1989; Molina et al. 1992a,b; Piña and Puetter 1993), and emphasize the great importance of integrals over the parameter space, both for treating nuisance parameters, such as regularization constants, and for comparing rival image models (Gull 1989; Skilling 1990). Truly ingenious methods have been devised for approximating the needed integrals (see, e.g., Skilling 1989, 1993), although this remains an open area of research. The MAP estimator has now been replaced with summaries of the posterior that describe the uncertainty of features in the image, either with error bars on functionals of the image (Skilling 1990), or with movies that wander through the posterior "bubble" of probable images (Skilling, Robinson, and Gull 1991). The large computational demands of these methods have somewhat limited their impact; but this constraint should weaken as computer power and algorithmic ingenuity grow with time.

At the opposite end of the spectrum in terms of the size of parameter space, several investigators independently realized the advantages to be gained by Bayesian analyses of photon counting data based on the Poisson sampling distribution. Gull (1988) discusses the estimation of a signal rate in

the absence of background as a simple example of basic Bayesian concepts. He also presents one of the earliest explicit analyses of the "Occam's Razor" effect in Bayesian model comparison. Kraft, Burrows, and Nousek (1991) discuss estimation of a signal from counting data in the presence of a known background. Loredo (1990, 1992) discusses signal estimation with and without a background, including the "on/off" case were the background itself must be estimated from off-source observations. Graziani et al. (1993) apply these ideas to estimation and model comparison problems arising in the analysis of the spectra of gamma-ray bursts.

Goebel, et al. (1989) 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 Satellite (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.

Bretthorst (1988) has developed a rich theory for the analysis of data modeled with a Gaussian distribution for added 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 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) apply the Bretthorst algorithm 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. Most of Bretthorst's work has focused on applications in chemistry and in radar target identification, with the unfortunate consequence that his methods have so far had less impact on astronomy than they should.

Independently, Finn and Chernoff (1993) applied very similar ideas to the analysis of gravitational radiation data like that expected from the Laser Interferometer Gravitational-wave Observatory (LIGO), currently under construction; Cutler and Flanagan (1994) have clarified and extended this work. The most promising sources of radiation detectable by LIGO are neutron star binaries, whose orbits decay due to gravitational radiation, resulting in the two stars spiraling into each other. General relativity predicts the shape of the resulting radiation waveform with high accuracy. Finn and Chernoff, and Cutler and Flanagan, show how Bayesian calculations can be used to infer physically interesting parameters of the decaying binary system. The ability to integrate over nuisance parameters plays an important role in their analysis. They use their results to discuss issues of experimental design, and to determine how accurately theoretical calculations must be performed in order to provide models of sufficient accuracy to model the data. Interestingly, the data may be so informative that they will tax current ability to perform the calculations of the waveform shape to the needed accuracy.

The methods of Bretthorst, Finn and Chernoff, and Cutler and Flanagan model the time series as a deterministic, parameterized signal with additive noise. However, simple parameterized models do not exist for many astrophysical time series. Rybicki and Press (1992) discuss Bayesian and frequentist methods for a time series modeled as a correlated Gaussian process, with added uncorrelated noise. The problem motivating their work is the analysis of radio waves detected from gravitationally lensed quasars. For these objects, the bending of spacetime by a massive foreground galaxy allows light from a more distant quasar (a galaxy with an unusually bright central core) to reach Earth along two or more paths, producing two or more images of the same quasar. The paths have different lengths, so light from the quasar reaches Earth at different times from the various

images. The time lag depends on properties of the intervening galaxy and on the density and expansion rate of the universe, so its measurement could in principle provide important cosmological information. Unfortunately, the intensity of quasars varies unpredictably with time. The work of Rybicki and Press represents an attempt to model such behavior and measure lag times. To date, however, no precise measurements have been possible.

Gregory and Loredo (1992) treat the complementary case of a time series consisting of arrival times of individual events modeled as a Poisson point process. They discuss both detection of periodic and nonperiodic signals using model comparison calculations, and estimation of signal parameters such as the signal period or waveform shape. Searches over wide ranges of periods can easily require millions of integrals over phase and shape nuisance parameters. Gregory and Loredo adopt a simple, piecewise-constant model for the waveform shape that facilitates the calculations by allowing *analytical* integration over the numerous waveform shape parameters.

Loredo and Wasserman (1994) use Bayesian methods to analyze data describing the distribution of gamma-ray burst sources; this work was briefly mentioned in Section 3. They use a Poisson point process to model the occurence of a burst, and in addition use the Poisson counting distribution to model the detection of photons from a burst. The analysis displays several weaknesses of currently used methods. A set of data analysis tools is being developed in collaboration with a burst observing team to facilitate use of Bayesian methods to model burst data. In forthcoming work, Loredo and Lamb (1994) use similar Bayesian methods to analyze the supernova neutrinos mentioned in Section 3.

Finally, Bayesian methods are beginning to find acceptance in studies of the large scale structure of the universe. In Section 4, I outlined a Bayesian approach to the analysis of CBR data to elucidate the structure present at the time of decoupling of radiation and matter. Lawrence, Readhead, and Myers (1988) and Bond et al. (1991) earlier applied Bayesian parameter estimation methods to similar data, in an effort to quantify the constraints placed by null detections on theories. The COBE detections have renewed interest in these methods, and Bond and others are developing more sophisticated Bayesian algorithms for the analysis of CBR data.

There is also a wealth of data on the character of the *present* large scale structure in the universe. However, its analysis is hampered by complicated selection effects. Many of these go under the name of "Malmquist bias." Malmquist biases arise whenever we attempt to infer a distribution with respect to some parameter, θ, from a sample of objects that have uncertain measurements of θ, and which may be incomplete, with sample membership determined by an uncertain measurement of θ. The effects of uncertainty and selection distort the observed distribution in a manner which depends on the shape of the underlying distribution we want to infer. As a result, priors play an important role in such analyses. Landy and Szalay (1992) explicitly introduced Bayesian ideas into these analyses. Their approach is essentially an "empirical Bayes" analysis, where the prior for the underlying distribution of distances to galaxies is parameterized as a histogram, with histogram levels inferred from the data. Their analysis has raised great interest in (and some controversy over) the application of Bayesian methods to the analysis of large scale structure, as evidenced in a number of preprints describing work in progress on such methods.

In the biblical story, the prodigal son was immediately welcomed home with a feast and much rejoicing. The welcome astronomers have extended to the Bayesian Prodigal has been more reluctant. Nevertheless, there has been a clearly preceivable change in the openness of astronomers to Bayesian methods within the last five years. This openness may well lead to a "feast" of new and better statistical practice among astronomers in the coming years, a feast of the Prodigal's own making.

ACKNOWLEDGEMENTS

This work was supported by a NASA GRO Fellowship (NAG 5-1758), NASA grant NAGW-666 and NSF grants AST-91-19475 and AST-93-15375 at Cornell University.

REFERENCES

Anderson, P. W. (1992). The Reverend Thomas Bayes, Needles in Haystacks, and the Fifth Force. *Physics Today* **45**(1), 9–11.

Berger, J.O. (1984). The Robust Bayesian Viewpoint. *Robustness of Bayesian Analyses* (J.B. Kadane ed.), B.V.: Elsevier Science Publishers, 63–124.

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

Bond, J. R., G. Efstathiou, P. M. Lubin, and P. R. Meinhold (1991). Cosmic-Structure Constraints from a ONe-Degree Microwave-Background Anisotropy Experiment. *Physical Review Letters* **66**, 2179–2182.

Bracewell, R. N. (1986). *The Hartley Transform*, Oxford University Press, New York.

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

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

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

Cox, R.T. (1946). Probability, Frequency, and Reasonable Expectation. *Am. J. Phys.* **14**, 1–13.

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

Cutler, C., and E. Flanagan (1994) ???. *Phys. Rev.* **D49**, 2658.

Finn, L. S., and D. F. Chernoff (1993). Observing Binary Inspiral in Gravitational Radiation: One interferometer. *The Physical Review* **D47**, 2198–2219.

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. *Astronomy and Astrophysics* **222**, L5–L8.

Gould, A. (1993). An Estimate of the COBE Quadrupole. *The Astrophysical Journal* **403**, L51–L54.

Graziani, C., D. Q. Lamb, T. J. Loredo, E. E. Fenimore, T. Murakami, and A. Yoshida (1993). Establishing the Existence of Harmonically-Spaced Lines in Gamma-Ray Burst Spectra Using Bayesian Inference. *Compton Gamma Ray Observatory, St. Lous, MO 1992* (M. Friedlander, N. Gehrels, and D.J. Macomb, eds.), New York: American Institute of Physics, 897–901.

Gregory, P. C., and T. J. Loredo (1992). A New Method for the Detection of a Periodic Signal of Unknown Shape and Period. *The Astrophysical Journal*, **398**, 146–168.

Gull, S. F., and G. J. Daniell (1979). The Maximum Entropy Method. *Image Formation from Coherence Functions in Astronomy* (C. van Schooneveld, ed.), Dordrecht: D. Reidel Publishing Company, 219–225.

Gull, S. F. (1988). Bayesian Inductive Inference and Maximum Entropy. *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1* (G. J. Erickson and C. R. Smith, eds.), Dordrecht: Kluwer Academic Publishers, 53–74.

Gull, S.F. (1989). Developments in Maximum Entropy Data Analysis. *Maximum-Entropy and Bayesian Methods* (J. Skilling, ed.), Dordrecht: Kluwer Academic Publishers, 53–71.

Janssen, M. A., and S. Gulkis (1992). Mapping the Sky With the COBE Differential Microwave Radiometers. *The Infrared and Submillimeter Sky After COBE* (M. Signore and C. Dupraz, eds.), Dordrecht: Kluwer Academic Publishers, 391–408.

Jaynes, E.T. (1957) How Does the Brain Do Plausible Reasoning?. Stanford Univ. Microwave Laboratory Report No. 421, reprinted in *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1* (G.J. Erickson and C.R. Smith, eds.), Dordrecht: Kluwer Academic Publishers 1–24 (1988).

Jaynes, E.T. (1984). The Intuitive Inadequacy of Classical Statistics. *Epistemologia* **VII**, 43–74.

Jaynes, E. T. (1987). Bayesian Spectrum and Chirp Analysis. *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems* (C. R. Smith and G. J. Erickson, eds.), Dordrecht: D. Reidel Publishing Company, 1–37, 1987.

Jaynes, E. T. (1988). Detection of Extra-Solar System Planets. *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1* (G. J. Erickson and C. R. Smith, eds.), Dordrecht: Kluwer Academic Publishers, 147.

Jaynes, E.T. (1990). Probability Theory as Logic. *Maximum Entropy and Bayesian Methods* (P. Fougère, ed.), Dordrecht: Kluwer Academic Publishers, 1–16.

Jaynes, E.T. (1995) *Probability Theory—The Logic of Science*, in preparation. The latest version is available over the internet by anonymous `ftp` from `bayes.wustl.edu` in the directory `Jaynes.book`.

Jefferys, W. H., and J. O. Berger (1992). Occam's Razor and Bayesian Inference. *American Scientist* **80**, 64–72.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press, Oxford (1st edition 1939).

Kleppner, D. K. (1992). Fretting About Statistics. *Physics Today* **45**(7), 9–11.

Kraft, R. P., D. N. Burrows, and J. A. Nousek (1991). Determination of Confidence Limits for Experiments with Low Numbers of Counts. *The Astrophysical Journal* **374**, 344–355.

Lampton, M., B. Margon, and S. Bowyer (1976). Parameter Estimation in X-Ray Astronomy. *The Astrophysical Journal* **208**, 177–190.

Landy, S. D., and A. S. Szalay (1992). A General Analytical Solution to the Problem of Malmquist Bias Due to Lognormal Distance Errors. *The Astrophysical Journal* **391**, 494–501.

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

Lawrence, C. R., A. C. S. Readhead, and S. T. Myers (1988). Microwave Background Radiation Observations at the Owens Valley Radio Observatory: Analysis and Results. *The Post-Recombination Universe* (N. Kaiser and A. N. Lasenby, eds.), 173–181.

Lindley, D. (19xx). A Bayesian Twenty-First Century. ???

Loredo, T. J. (1990). From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics. *Maximum-Entropy and Bayesian Methods, Dartmouth, 1989*, (P. Fougère, ed.), Dordrecht: Kluwer Academic Publishers, 81–142.

Loredo, T. J. (1992). The Promise of Bayesian Inference for Astrophysics (with Discussion). *Statistical Challenges in Modern Astronomy*, (E. D. Feigelson and G. J. Babu, eds.), New York: Springer-Verlag, 275–297.

Loredo, T. J., and D. Q. Lamb (1994). Bayesian Analysis of Neutrinos from SN 1987A: Implications for Cooling of the Nascent Neutron Star and for the Rest Mass of the Electron Antineutrino. To be submitted to *Phys. Rev. D*.

Loredo, T. J., and I. M. Wasserman (1994). Inferring the Spatial and Energy Distribution of Gamma-Ray Burst Sources. I. Methodology. *The Astrophysical Journal (Supplement)*, in press.

Molina, R., B. D. Ripley, A. Molina, F. Moreno, and J. L. Ortiz (1992a). Bayesian Deconvolution with Prior Knowledge of Object Location: Applications to Ground-Based Panetary Images. *The Astronomical Journal* **104**, 1662–1668.

Molina, R., A. Del Olmo, J. Perea, and B. D. Ripley (1992b). Bayesian Deconvolution in Optical Astronomy. *The Astronomical Journal* **103**, 666–675.

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. *Advances in Helio- and Asteroseismology* (J. Christensen-Dalsgaard and S. Frandsen, eds.), International Astronomical Union, 485–489.

Piña, R. K., and R. C. Puetter (1993). Bayesian Image Reconstruction: The Pixon and Optimal Image Modeling. *Pub. of the Astron. Soc. of the Pacific* **105**, 630–637.

Rybicki, G. B., and W. H. Press (1992). Interpolation, Realization, and Reconstruction of Noisy, Irregularly Sampled Data. *The Astrophysical Journal* **398**, 169–176.

Skilling, J. (1989). The Eigenvalues of Mega-Dimensional Matrices. *Maximum-Entropy and Bayesian Methods* (J. Skilling, ed.), Dordrecht: Kluwer Academic Publishers, 455-466.

Skilling, J. (1990). Quantified Maximum Entropy. *Maximum-Entropy and Bayesian Methods, Dartmouth, 1989*, (P. Fougère, ed.), Dordrecht: Kluwer Academic Publishers, 341–350.

Skilling, J. (1993). Bayesian Numerical Analysis. *Physics and Probability: Essays in Honor of E. T. Jaynes* (W.T. Grandy, Jr. and P.W. Milonni, eds.), Cambridge: Cambridge University Press.

Skilling, J., D. R. T. Robinson, and S. F. Gull (1991). Probabilistic Displays. *Maximum Entropy and Bayesian Methods, Laramie, Wyoming, 1990* (W. T. Grandy, Jr. and L. H. Schick, eds.), Dordrecht: Kluwer Academic Publishers, 365–368.

Smoot, G. F. et al. (1992). Structure in the COBE Differential Microwave Radiometer First-Year Maps. *The Astrophysical Journal* **396**, L1–L5.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge: Harvard University Press.