SIAM/ASA J. Uncertainty Quantification
Vol. 3, pp. 586–598

Constraints versus Priors*

Philip B. Stark†

Abstract. There are deep and important philosophical differences between Bayesian and frequentist approaches to quantifying uncertainty. However, some practitioners choose between these approaches primarily on the basis of convenience. For instance, the ability to incorporate parameter constraints is sometimes cited as a reason to use Bayesian methods. This reflects two misunderstandings: First, frequentist methods can indeed incorporate constraints on parameter values. Second, it ignores the crucial question of what the result of the analysis will mean. Bayesian and frequentist measures of uncertainty have similar sounding names but quite different meanings. For instance, Bayesian uncertainties typically involve expectations with respect to the posterior distribution of the parameter, holding the data fixed; frequentist uncertainties typically involve expectations with respect to the distribution of the data, holding the parameter fixed. Bayesian methods, including methods incorporating parameter constraints, require supplementing the constraints with a prior probability distribution for parameter values. This can cause frequentist and Bayesian estimates and their nominal uncertainties to differ substantially, even when the prior is "uninformative." This paper gives simple examples where "uninformative" priors are, in fact, extremely informative, and sketches how to measure how much information the prior adds to the constraint. Bayesian methods can have good frequentist behavior, and a frequentist can use Bayesian methods and quantify the uncertainty by frequentist means—but absent a meaningful prior, Bayesian uncertainty measures lack meaning. The paper ends with brief reflections on practice.

Key words. frequentist methods, Bayesian methods, uncertainty quantification, constraints, priors, risk

AMS subject classifications. 62A01, 62C10, 62C20, 62G15, 62P99

DOI. 10.1137/130920721

1. Introduction.

This paper is primarily about philosophy and interpretation rather than computation. It sketches differences in the *assumptions* frequentist and Bayesian methods require and differences in *meaning* between frequentist and Bayesian measures of uncertainty, and recapitulates some well-known (within statistics) connections between frequentist and Bayesian methods that are less well known in the wider uncertainty quantification community. It also addresses the misperception that frequentist methods cannot incorporate constraints on unknown parameters, and flags some circumstances in which frequentists might choose to use Bayesian methods without using Bayesian interpretations of the results.

This paper focuses on making inferences about a (possibly infinite-dimensional) vector θ known to be in the subset Θ of a Hilbert space $\mathcal{H}$, from observations related to that vector. In examples below, the set Θ is a "ball" in a norm or seminorm. A one-dimensional special case is to estimate a single, real-valued parameter that is known to be between 0 and 1. Then, the Hilbert space $\mathcal{H}$ is the real line, the norm is the absolute value, and the set Θ is a ball of

*Received by the editors May 13, 2013; accepted for publication (in revised form) June 1, 2015; published electronically July 23, 2015.
http://www.siam.org/journals/juq/3/92072.html

†Department of Statistics, University of California, Berkeley, CA 94720 (stark@stat.berkeley.edu).

radius 1/2 centered at the point 1/2.

This one-dimensional problem does arise in practice: The parameter might be a mortality rate; the fraction of a healthcare invoice that is fraudulent; the percentage by which the margin of victory of an election is overstated; the fraction of incident energy absorbed by a material; the ratio of applied load to failure strength for a structure such as a bridge, dam, building, or engine mount (given that the structure did not fail); the ratio of an individual's exposure to a toxic substance to the exposure that would be lethal (given that the individual did not die of the exposure); or the ratio of the speed of a particle to the velocity of light.

Useful constraints on high-dimensional parameters, even infinite-dimensional parameters such as functions of time and space, often can be expressed as the assertion that a norm or seminorm is bounded by a constant. For example, we might be interested in the electric field generated by some device, as a function of position in space; we might know a priori that the energy of that field—a quadratic norm—is less than 1, in suitable units. Or we might be interested in the spatial variation of ground motion caused by a large earthquake; we might have an a priori upper bound on the total energy released by the earthquake. Constraints of this kind can reduce the uncertainty or parameter estimates, sometimes substantially; they can even render otherwise insoluble problems solvable [1, 2, 3, 4, 5, 6, 7].

The two most common approaches to dealing with problems of statistical estimation and inference are *frequentist* and *Bayesian*. The frequentist approach uses constraints directly: The estimation or inference problem is restricted to the set of parameters that satisfy the constraints. In the Bayesian or "subjective" approach, the analyst imposes the constraints by choosing a prior probability distribution that assigns probability 1 to the set of parameters that satisfy the constraint. These two approaches result in uncertainty estimates that differ substantially in definition and interpretation, and that can differ substantially numerically. This paper recounts some of the philosophical and definitional differences.

2. Notation.

The notation follows that in [8]. We want to use measurements Y to learn about the (unknown) state of the world, denoted by θ, in a mathematical representation of some physical system. In many interesting cases, θ is a function of position and/or time, which in principle would require infinitely many numbers to describe perfectly—it is an element of an infinite-dimensional space. We shall assume that the possible values of Y are n-vectors of real numbers, i.e., elements of $\mathbb{R}^n$. Mathematics, physics, or previous experiments tell us that θ satisfies a constraint like those described in the introduction. The set Θ expresses this constraint: We know a priori that $\theta \in \Theta$.

The measurements Y are related to the state of the world θ through a *measurement model*, which gives the probability distribution $\mathbb{P}_\eta$ of Y for each possible state of the world $\eta \in \Theta$. That is, if $\theta = \eta$, then $Y \sim \mathbb{P}_\eta$, where $\sim$ means "has probability distribution" or "is distributed as." Depending on θ, Y may be more likely to take values in some subsets of $\mathbb{R}^n$ than in others. For any particular set $A \subset \mathbb{R}^n$, $\mathbb{P}_\eta\{Y \in A\}$ generally will not be equal to $\mathbb{P}_{\eta'}\{Y \in A\}$ unless $\eta = \eta'$. However, it can be the case that $\eta \neq \eta'$ and yet $\mathbb{P}_\eta\{Y \in A\} = \mathbb{P}_{\eta'}\{Y \in A\}$ for all (measurable) subsets $A \subset \mathbb{R}^n$. (In that case, θ is not *identifiable*.)

We shall assume that there is a known measure μ that dominates the set of distributions $\mathcal{P} \equiv \{\mathbb{P}_\eta : \eta \in \Theta\}$, so we can use "densities" even if some members of $\mathcal{P}$ have atoms. This assumption makes it easier to define likelihoods, which are required for the Bayesian

framework; it also implies that all the measures $\mathcal{P}$ are defined on a common sigma-algebra $\mathcal{A}$ of subsets of $\Re^n$, which avoids some potential pathologies.

With respect to μ, the density of $\mathbb{P}_\eta$ at y is

$$p_\eta(y) \equiv d\mathbb{P}_\eta/d\mu|_y. \tag{2.1}$$

For any fixed y, the *likelihood of* η *given* $Y = y$ is $p_\eta(y)$, viewed as a function of η alone.

While we might like to know θ, it is often impossible to estimate it with any useful level of accuracy; in many problems, θ is not even identifiable. But it still may be possible and scientifically interesting to estimate a *parameter* $\lambda = \lambda[\theta]$, a property of θ. The parameter might be the average of θ over some volume of space or time, a norm or seminorm of θ, or the number of local maxima θ has, for instance. We shall assume that the possible values of $\lambda[\theta]$ are also elements of a Hilbert space.

3. The Bayesian approach. The Bayesian approach starts with a *prior probability distribution* π on Θ, and the likelihood function $p_\eta(y)$. To have a prior probability distribution involves some technical restrictions that will not be considered here.[1] Together, π and p_η imply a joint distribution of θ and Y. The *marginal distribution* or *predictive distribution* of Y is

$$m(y) = \int_\Theta p_\eta(y)\, \pi(d\eta). \tag{3.1}$$

On observing that $Y = y$, we assimilate the information by applying Bayes' rule to find the *posterior distribution of* θ *given* $Y = y$:

$$\pi(d\eta|Y = y) = \frac{p_\eta(y)\ \pi(d\eta)}{m(y)}. \tag{3.2}$$

(The marginal density $m(y)$ can vanish, but the probability that it does is zero.) In principle, this solves the problem: All the information in the prior and the data is now contained in the posterior distribution. The posterior distribution $\pi_\lambda(d\ell|Y = y)$ of $\lambda[\theta]$ is the distribution induced by the posterior distribution of θ: For any (measurable) set A of possible values of λ,

$$\Pr(\lambda[\theta] \in A|Y = y) = \int_{\ell \in A} \pi_\lambda(d\ell|Y = y) \equiv \int_{\eta:\lambda[\eta]\in A} \pi(d\eta|Y = y). \tag{3.3}$$

If we get additional data, we apply Bayes' rule again, using the current posterior as the new prior.

3.1. Arguments for the Bayesian approach. There are a number of arguments for the Bayesian approach. One is that people are in fact Bayesian: that the Bayesian approach is *descriptive*. There is considerable evidence that this is false. I do not know of anyone who uses Bayes' theorem to combine and update prior beliefs in ordinary life. Extensive empirical research, starting with the seminal work of [9], shows that even people with training in probability fail to incorporate Bayes' rule in their day-to-day reasoning.

[1] For instance, Θ must be a measurable subset of a measurable space, and the likelihood function must be jointly measurable with respect to η and y.

The second argument is that people should be Bayesian: that the Bayesian approach is *normative*. According to the argument, if people are not Bayesians, their probability assignments are "incoherent" and others can make "Dutch book" against them. ("Dutch book" is a combination of bets such that no matter what the outcome is, the bettor loses money.) The coherence argument depends in part on the assumption that all beliefs can be expressed as probability distributions, an assumption that many are unwilling to grant. The "Dutch book" argument depends on the non-Bayesian analyst's willingness to cover an unlimited number of bets, and on the assumption that the Bayesian analyst's prior is *proper*, that is, that the total mass of the prior is 1. In practice, improper priors are common; for instance, it is common to use a uniform prior for parameters on unbounded domains. See [10, 11].

A third argument is that the choice of the prior does not matter (much), because the data eventually overwhelm the prior: No matter what prior you start with, given enough data, you end up with essentially the same posterior distribution. This convergence occurs in some circumstances but not in others [12, 13]: The prior can bite you on the posterior. See [14] for further discussion of these and related issues.

I think Bayesian methods are popular in uncertainty quantification in part because (depending on the prior) they generally make the uncertainty appear smaller than a frequentist analysis would show, which makes the results more optimistic, and in part because they give a (numerically computable) recipe that can be applied to essentially any problem—if you have a prior, a fast enough computer, and a good algorithm. (The development of Markov chain Monte Carlo (MCMC) made it possible to compute posterior distributions in a far larger class of problems, greatly increasing the appeal of Bayesian methods.) Of course, the fact that you can compute something does not automatically make the answer meaningful, relevant, or useful: The world is full of "quantifauxcation," which consists of assigning a meaningless number to something, then concluding that because the result is quantitative, it must be meaningful.

A Bayesian can give a number for the probability of an act of nuclear terrorism in the year 2025; a frequentist cannot make sense of what "probability" could mean in that context. A Bayesian can give a number for the probability that there are civilizations of intelligent beings in other galaxies; a frequentist cannot make sense of the question. A Bayesian can give a number for the probability that a particular coin in your pocket is fair, sight unseen; a frequentist cannot make sense of the question. A Bayesian can give a number for the probability that a particular hypothesis is true; a frequentist thinks hypotheses are either true or false (*evidence* often has a random component—noise or measurement error—but a well-formed hypothesis itself is either true or false). In situations where both frequentist and Bayesian methods can be applied, probability and uncertainty mean quite different things to frequentists and to Bayesians, as elaborated below.

3.2. Priors. In the Bayesian approach, all information about θ is expressed as a probability distribution, and all probabilities quantify degree of belief. (In the frequentist framework, probability is defined in terms of long-run relative frequency.) If the analyst is certain that A is true, A has probability 1. If she is certain that A is false, A has probability 0. If she believes that A is true (or that A will occur) with the same strength that she believes that A is false (or that A will not occur), A has probability 1/2. If she believes twice as strongly that A is true as she believes A is false, A has probability 2/3. More generally, if A and B

are two statements and the analyst believes r times as strongly that A is true as she believes that B is true, then the probability of A is r times the probability of B.

To use the Bayesian framework to quantify uncertainty, one *must* quantify beliefs and constraints by means of probability distributions. (As discussed below, there can be good reasons to use Bayesian estimators without necessarily using the Bayesian framework to quantify uncertainty.) The prior probability distribution quantifies the analyst's beliefs about θ before data are collected: The constraint $\theta \in \Theta$ means the prior probability distribution π (defined on a suitable sigma-algebra of subsets of Θ) must assign probability 1 to Θ.

Even in the simplest nondegenerate case, there are infinitely many probability distributions that assign probability 1 to Θ. Which should be the prior? In principle, it is up to the analyst to introspect to find the unique prior that reflects her beliefs about θ; in turn those beliefs should be constructed from previous experience and previous beliefs through the repeated application of Bayes' theorem.

Although there is a literature on the elicitation of prior probabilities [15], I have never seen a Bayesian analysis of real data in which the data analyst made a serious attempt to quantify her beliefs using a prior. (Nor, to my knowledge, have I met anyone who uses Bayes' theorem in real life to update her beliefs.) Instead, in my experience, priors are generally taken as given and appear to be selected or justified in five ways: (i) to make the calculations simple (e.g., closed form), (ii) because the particular prior is conventional, (iii) so that the prior satisfies some invariance principle, (iv) with the assertion that the prior is "uninformative," or (v) because the prior roughly matches the relative frequencies of values in some population. Computational tractability has become less of a desideratum as simulation algorithms and computing power have advanced. Convention wields considerable force: The conventional tends to be treated as "correct" or "obvious." See [16, 8] for discussions of (iii) and (iv).

Some researchers use Laplace's principle of insufficient reason to select an "uninformative" prior: If there is no reason to believe that outcomes are not equally likely, assume that they are equally likely. Of course, the outcomes considered may depend on the parametrization, among other things. Generally, however, the principle leads to a prior π that is uniform on Θ. That is, the probability of any subset of Θ is assumed to be proportional to its Lebesgue measure.

For instance, the "uninformative" prior for a real parameter known to be in $\Theta \equiv [-1, 1]$ is the uniform distribution on $[-1, 1]$, which has density $f(\eta) = \{1/2,\ \eta \in [-1, 1];\ 0 \text{ otherwise}\}$. This prior captures the constraint $\theta \in [-1, 1]$, but it does far more than that: It assigns probabilities to all measurable subsets of $[-1, 1]$. For instance, it says that there is a 50% chance that θ is positive, a 50% chance that the absolute value of θ is greater than $1/2$, and a 90% chance that the absolute value of θ is greater than $1/10$. This is not information that came from the constraint: It is information added by the prior. The constraint $\theta \in \Theta$ requires π to assign probability 1 to Θ, but it does not restrict the probabilities π assigns to subsets of Θ. Any choice of π, "uninformative" or not, says more about θ than the original constraint did.

This problem—that turning constraints into priors adds information—grows worse as the dimension of the parameter θ grows. For instance, suppose the unknown θ is a vector in n-dimensional Euclidean space $\mathbb{R}^n$ and we know that $\|\theta\| \leq 1$—that is, Θ is the unit ball in $\mathbb{R}^n$. The volume of a spherical shell from radius $1 - \epsilon$ to 1 is a larger and larger fraction of the

volume of the unit sphere as the dimension n grows. For any $\alpha \in (0,1)$ and $\epsilon \in (0,1)$, there is a dimension n so that the (uniform) probability of $\{\eta : \|\eta\| \in [1-\epsilon, 1]\}$, the spherical shell from radius $1-\epsilon$ to 1, is at least α.

What does this mean? Starting with the constraint that $\|\theta\| \leq 1$—and without collecting any data—we end up with arbitrarily high certainty that in fact $\|\theta\| \geq 1-\epsilon$. It is the prior that gives us this certainty, not the constraint. The prior is not "uninformative" about the norm.

Conversely, suppose we put a rotationally invariant prior on the unit ball in such a way that the marginal distribution of the norm is uniform. Consider the ball of radius $1-\epsilon$. It has probability $1-\epsilon$ regardless of the dimension of the space, even though its volume is a negligible fraction of the volume of the unit ball if the dimension of the space is large. This prior is not "uninformative" with respect to volume: It says that the model is extremely likely to be in a subset of Θ that has very small volume.

The problem reaches a head in infinite-dimensional spaces. For instance, suppose θ is an element of an infinite-dimensional separable Hilbert space, and that the constraint set Θ is rotationally invariant (an example would be $\Theta \equiv \{\eta : \|\eta\| \leq 1\}$). If we want the prior to respect that rotational invariance, it is a theorem that the prior either assigns probability 1 to the event that $\theta = 0$ or it assigns probability 1 to the event that the norm of θ is infinite—contradicting the constraint the prior was intended to capture [17].

3.3. Interpretation of Bayesian priors and estimates. Prior probability distributions—and hence posterior distributions—are quantifications of the analyst's degree of belief. As such, they change the subject from the experiment and the external world to the analyst's state of mind. Suppose I claim that my prior probability distribution for the load on a structure as a fraction of its breaking strength is the uniform distribution on $[0,1]$. I am right if that accurately reflects what I believe. I am wrong if it does not accurately reflect what I believe. The relationship between the prior and the world has no bearing on whether I am right or wrong. Experiments that could show I am wrong involve checking what I actually believe—for instance, psychological testing or determining what bets I would take at what odds—rather than measurements of the structure or similar structures.

Two analysts can have very different priors and both be right, because what makes a prior right is that it correctly quantifies the analyst's belief. If I do not share your prior beliefs in detail, then even if we agree on the likelihood function and the data, we will have different posterior distributions for θ. Why should your posterior distribution matter to me? If a Bayesian analysis results in the statement, "there is a 99.9% chance that the applied load will be less than 10% of the breaking strength," it means that the analyst is quite sure that the load will be low, but it is not at all clear what it means about safety. For a different prior, an equally correct analysis might find that there is a 99.9% chance that the applied load will exceed 90% of the breaking strength. If so, a Bayesian analysis might appropriately be viewed with skepticism. On the other hand, if one could show that no matter what prior is used, there is at least a 99.9% chance that the applied load will be less than 10% of the breaking strength, the Bayesian position would seem much more persuasive. While the Bayesian approach has its merits, the utility and persuasiveness of Bayesian analyses may hinge on the sensitivity of the conclusions to the choice of prior.

4. The frequentist approach. The main difference between Bayesian and frequentist approaches to constraints boils down to the difference between believing that θ is drawn at random from Θ according to the known distribution π and believing that θ is simply an unknown element of Θ. (The interpretation of probability also differs substantially between the two points of view.)

As discussed above, for Bayesians, probability quantifies degree of belief. For frequentists, probability has to do with long-term regularities in repeated trials. The probability of an event is defined to be the long-run limiting relative frequency with which the event occurs in independent trials under "essentially identical" conditions. "Essentially identical" is in quotation marks because if the conditions were exactly identical, then (within classical physics, at least), the outcome would be identical. The canonical random experiment, tossing a fair coin, will give heads every time or tails every time if the coin is tossed with initial conditions that are similar enough.[2] Defining "essentially identical" is a serious problem for the frequentist approach. Another is the assumption that repeated trials result in relative frequencies that converge to a limit. This assumption is an assertion about how the world works, an assertion that cannot be verified empirically, since it does not posit any particular rate of convergence. As mentioned above, the frequentist approach severely restricts the kinds of things one can make probability statements about: Only trials that, in principle, can be repeated indefinitely lead to probabilities. For instance, a conventional frequentist approach cannot make sense of questions like "what is the chance of an act of nuclear terrorism in the year 2025?" or "what is the chance of an earthquake with magnitude 8.0 or above in the San Francisco Bay Area in the next 20 years?," much less supply numerical values for those chances.

In the frequentist approach, probability generally resides in the measurement process or the experiment, not in the parameter. There is statistical uncertainty because there is sampling variability or measurement error or random assignment of subjects to treatments, not because the underlying parameter is random. Since it is not necessary to assume that θ is random to use the frequentist approach, it is not necessary to augment the constraint $\theta \in \Theta$ with a prior probability distribution.

5. Summarizing uncertainty. There are many ways to quantify uncertainty. We shall consider two, each of which has a Bayesian and a frequentist variant: mean squared error (a frequentist measure) and posterior mean squared error (the related Bayesian measure); and confidence sets (a frequentist construct) and credible regions (the related Bayesian construct).

5.1. Mean squared error. Recall that we have assumed that $\lambda[\theta]$ takes values in a Hilbert space. Suppose we choose to estimate $\lambda[\theta]$ by the estimator $\widehat{\lambda}(Y)$, a (measurable) map from possible data values y into possible values of $\lambda[\eta]$, $\eta \in \Theta$. The mean squared error (MSE) of $\widehat{\lambda}$ when $\theta = \eta$ is

$$\mathrm{MSE}(\widehat{\lambda}(Y), \eta) \equiv \mathbb{E}_\eta \|\widehat{\lambda}(Y) - \lambda[\eta]\|^2. \tag{5.1}$$

The MSE depends on η. The expectation is with respect to $\mathbb{P}_\eta$, the distribution of the data Y on the assumption that $\theta = \eta$. If we get to select the estimator $\widehat{\lambda}$, we might seek an estimator

[2] For instance, Diaconis [18] notes that he and other magicians can toss a coin to land heads or tails at will and describes an apparatus built by physicists that tosses a coin reproducibly.

that makes $\mathrm{MSE}(\widehat{\lambda}(Y), \theta)$ small. But since the true value of θ is unknown, in general we cannot select the estimator $\widehat{\lambda}$ to make the actual MSE as small as possible. Instead, we might choose $\widehat{\lambda}$ to make the largest MSE as η ranges over Θ as small as possible. That choice is the *minimax MSE estimator*.

A related Bayesian measure of uncertainty is the posterior mean squared error (PMSE),

$$\mathrm{PMSE}(\widehat{\lambda}(y), \pi) \equiv \mathbb{E}_\pi \|\widehat{\lambda}(y) - \lambda[\eta]\|^2. \tag{5.2}$$

The PMSE depends on π and the observed value of y. The expectation is with respect to the posterior distribution of θ given $Y = y$. Since π is known, we can select (for each y) the estimator that has the smallest possible PMSE. That estimator, the Bayes estimator for PMSE, is the *marginal posterior mean*, the mean of $\pi_\lambda(d\ell|Y = y)$, the marginal posterior distribution of $\lambda[\theta]$ given Y [19, 20]:

$$\widehat{\lambda}_\pi(y) \equiv \int \ell \pi_\lambda(d\ell|Y = y). \tag{5.3}$$

Even though the MSE and PMSE both involve expectations of the squared norm of the difference between the parameter estimate and the true value of the parameter, they are conceptually quite different: The MSE is an expectation with respect to the distribution of the data Y, holding the parameter $\theta = \eta$ fixed, while the PMSE is an expectation with respect to the posterior distribution of θ, holding the data $Y = y$ fixed.

5.2. Confidence sets and credible regions. Suppose $\alpha \in (0, 1)$. A random set $\mathcal{I}(Y)$ of possible values of λ is a $1 - \alpha$ *confidence set* for $\lambda[\theta]$ if

$$\mathbb{P}_\eta\{\mathcal{I}(Y) \ni \lambda[\eta]\} \geq 1 - \alpha \quad \forall \eta \in \Theta. \tag{5.4}$$

The probability on the left is with respect to the distribution of the data Y, holding η fixed. In the frequentist view, once the data are collected and we know that $Y = y$, there is no longer any probability: The set $\mathcal{I}(y)$ is some particular set, and the value $\lambda[\theta]$ is some particular (but unknown) vector, so either $\mathcal{I}(y)$ contains $\lambda[\theta]$ or it does not. The "coverage probability" of the rule $\mathcal{I}$ is the (smallest) chance that $\mathcal{I}(Y)$ will include $\lambda[\eta]$ as η ranges over Θ, with Y generated from $\mathbb{P}_\eta(y)$.

A related Bayesian construct is a *posterior credible region*. A set $\mathcal{I}(y)$ of possible values of λ is a $1 - \alpha$ posterior credible region for $\lambda[\theta]$ if

$$\mathbb{P}_{\pi(d\theta|Y=y)}(\lambda[\theta] \in \mathcal{I}(y)) \equiv \int_{\mathcal{I}(y)} \pi_\lambda(d\ell|Y = y) \geq 1 - \alpha. \tag{5.5}$$

The probability on the left is with respect to the marginal posterior distribution of $\lambda[\theta]$, holding the data fixed: It is the posterior probability that $\mathcal{I}(y)$ contains $\lambda[\theta]$ given that $Y = y$. In the Bayesian view, once the data are collected and we know that $Y = y$, there is still probability, because the value of θ itself remains random: Its value is uncertain, and all uncertainty is represented as probability.

There are countless ways of constructing confidence sets and credible regions. Since the volume of the set or region is a measure of precision—the uncertainty after the data have been collected—it can be desirable to choose confidence sets and credible regions to minimize their (expected) volumes. See, for instance, [7, 21, 22].

6. Decision theory. This section gives a brief sketch of decision theory, a framework for comparing estimators. For a bit more detail, see [8]; for a rigorous treatment, see [23]. Decision theory treats estimation as a two-player game: Nature versus analyst. The game frequentists play has slightly different rules from the game Bayesians play. According to both sets of rules, Nature and the analyst know Θ, $\mathbb{P}_\eta$, for all $\eta \in \Theta$, λ, and the payoff rule (loss function) $\text{loss}(\ell, \lambda[\eta])$, the amount of money the analyst loses if she guesses that $\lambda[\eta] = \ell$ when in fact $\theta = \eta$. Nature selects an element θ of Θ. The analyst selects an estimator $\widehat{\lambda}$. The analyst does not know the value of θ, and Nature does not know what estimator the analyst plans to use. Data Y are generated using the value of θ that Nature selected; the data are plugged into $\widehat{\lambda}$, and $\text{loss}(\widehat{\lambda}(Y), \lambda[\theta])$ is calculated. Holding θ constant, a new value of Y is generated, and $\text{loss}(\widehat{\lambda}(Y), \lambda[\theta])$ is calculated again. This is repeated many times. The analyst has to pay the average value of $\text{loss}(\widehat{\lambda}(Y), \lambda[\theta])$ over all those values of Y, the *risk of* $\widehat{\lambda}$ *at* θ, denoted $\rho_\theta(\widehat{\lambda}, \lambda[\theta])$. The analyst's goal is to lose as little as possible in repeated play.

In the Bayesian version of the game, Nature agrees to select θ at random according to the prior distribution π, and the analyst knows π. In the frequentist version of the game, the analyst does not know how Nature will select θ from Θ. This is perhaps the most important difference between the frequentist and Bayesian viewpoints: Bayesians claim to know more about how Nature generates the data.

A cautious frequentist might wish to select $\widehat{\lambda}$ to minimize her worst-case risk, on the assumption that Nature might play deliberately to win as much as possible. An estimator that minimizes the worst-case risk over $\eta \in \Theta$ (for some specified class of estimators) is called a *minimax estimator*; its maximum risk is the *minimax risk*. Minimax estimates are not the only option for frequentists (indeed, in many problems the minimax estimator is not known, and frequentists rely on estimators that have simple recipes and generally good asymptotic properties, e.g., maximum likelihood), but minimaxity is a common principle for optimality, as is *minimax regret*.

A Bayesian might instead select the estimator that minimizes the average risk on the assumption that Nature selects θ at random following the prior probability distribution π. An estimator that minimizes the average risk when θ is selected from π (for some specified class of estimators) is called a *Bayes estimator*; its average risk for prior π is the *Bayes risk*.

6.1. Duality between Bayes risk and minimax risk. The Bayes risk depends not only on Θ, the distributions $\{\mathbb{P}_\eta : \eta \in \Theta\}$, the parameter λ, and the loss function; it also depends on π. Consider allowing π to vary over a (suitably) rich set of possible priors. The prior π^* for which the Bayes risk is largest is the *least favorable* prior. The least favorable prior typically is not the "uninformative" prior. The works [21, 22] give numerical methods for approximating least favorable priors. Under some technical conditions, the Bayes risk for the least favorable prior is equal to the minimax risk [19, 20]. If the Bayes risk is much smaller than the minimax risk, it is because the prior added information not present in the constraint itself. When that occurs, one might reasonably wonder *why* the analyst believes Nature selected θ from the prior distribution π, and might ask whether one shares the analyst's beliefs about θ.

For example, [8] compare the minimax risk and the Bayes risk for estimating $\theta \in \Theta \equiv [-\tau, \tau]$ from data $Y \sim N(\tau, 1)$, the bounded normal mean problem, and evaluate the frequentist coverage probability of Bayes posterior credible regions. They find that when $\tau = 1/2$,

the Bayes risk for MSE is less than half the minimax risk, while for $\tau = 3$, the Bayes risk is much closer to the minimax risk. When $\tau = 1/2$, the expected length of a 95% Bayes credible region is at least as big as the expected length of a particular 95% confidence region (the truncated Pratt interval) for all $\eta \in \Theta$ and values of η for which its expected length is a couple of percent larger, but the frequentist coverage probability of the Bayes credible region is as low as 68% for some values of η. In contrast, when $\tau = 3$, there are values of η for which the expected length of the 95% Bayes credible region is about 33% less than the expected length of the truncated Pratt interval and values of η for which it is about 10% longer, while the coverage probability of a 95% credible region stays above 90% for all $\eta \in \Theta$.

7. Frequentist properties of Bayes estimates. It may behoove a frequentist analyst to consider Bayesian estimation methods but assess those methods using frequentist criteria. The frequentist need not believe the prior or even think that the parameter is random. The point is that the Bayesian approach gives a recipe for calculating an estimate, just like maximum likelihood gives a recipe for calculating an estimate. The performance of that estimate can be measured using frequentist constructs without relying on the posterior distribution or the interpretation of the prior or the posterior.

Indeed, Bayesian estimates may have good frequentist properties. For example, an estimator $\widehat{\lambda}$ is *admissible* for a given loss function if no other estimator has a risk that is never larger than the risk of $\widehat{\lambda}$, whatever $\theta \in \Theta$ happens to be, and has risk smaller than the risk of $\widehat{\lambda}$ for at least one $\eta \in \Theta$. Admissibility is a desirable frequentist property for an estimator: It means no estimator strictly dominates it. There are typically infinitely many admissible estimators in a given problem. In general, if an estimator is the Bayes estimator for some proper prior on Θ and has finite Bayes risk, then it is admissible. (Bayes estimators for improper priors typically are not admissible.)

Similarly, as illustrated in section 6.1, there are problems in which the Bayes estimator for some particular prior has low MSE, no matter what the true value of $\theta \in \Theta$ happens to be. The frequentist justification for using a Bayes estimator does not lie in whether the prior makes sense or whether the posterior properly measures the uncertainty, but in whether the frequentist performance of the estimator is adequate. For instance, [21, 22] use Bayes estimators for approximately least-favorable priors because they are approximately minimax estimators.

8. "The difference between theory and practice . . . is smaller in theory than it is in practice." In complex scientific applications, there are very few "orthodox" frequentist *or* Bayesian analyses. Rather, models and priors tend to be chosen for convenience or tractability—or to respond to criticisms by adding layers of complexity. To paraphrase David Freedman, frequentist analyses tend to make up models, and Bayesian analyses tend to make up priors [14]. I would argue that *both* camps tend to make up models, by which I mean the mapping $\eta \to \Pr_\eta$ and the set Θ. And both groups tend to *invent* constraints far more stringent than the constraints that actually come from the underlying scientific problem, artificially reducing the apparent uncertainty.

If you point out a missing source of uncertainty or variability, a frequentist may model it; a Bayesian may model it and put a prior on any new parameters. The result—in both camps—is a tendency toward rococo recursion in which the embellishments have embellishments, and

the weakness of the foundation is obscured by the complexity of the edifice. In the end, one still has a model or a prior, but an incomprehensible one that can't possibly correspond to anyone's true beliefs, or to Nature.

The verse by Augustus de Morgan (known for de Morgan's rules, which both frequentists and Bayesians rely on) describes multilevel modeling and hierarchical priors well:

> Great fleas have little fleas upon their backs to bite 'em,
> And little fleas have lesser fleas, and so ad infinitum.
> And the great fleas themselves, in turn, have greater fleas to go on,
> While these again have greater still, and greater still, and so on.

Frequentist analyses of complex problems often have model-selection phases (e.g., deciding which variables to use in a regression model) that are not accounted for properly in the quantification of uncertainty through confidence sets and so on. This problem has been recognized for decades, but only recently have rigorous methods to deal with it been proposed, and only in quite limited contexts [24, 25, 26, 27].

The best frequentist analyses tend to be bespoke: tailored to the scientific details of the problem. That requires substantive and statistical knowledge. Calculating frequentist estimates may require solving difficult—and sometimes numerically intractable—constrained numerical optimization problems. The set of tractable problems will grow over time as algorithms improve and computational power increases.

The advent of fast MCMC codes makes it possible to compute Bayesian estimates in a broad variety of applications, sometimes without much scientific thought: Insert a prior, a likelihood, and data, and run MCMC to sample the posterior; then out come an estimate and an uncertainty appraisal. What that appraisal means is generally not examined, much less questioned. The warning "if all you have is a hammer, everything looks like a nail" is apropos. This is especially true if you love your hammer; it is especially pernicious if true nails are rare.

In much the same way, frequentists often hurry to apply the latest modeling technique to every set of data with the right "signature," with little regard to how the data were collected or the underlying science. The result is unlikely to advance scientific knowledge: Current popularity is not a reliable measure of validity, and CPU cycles, while useful, are not a good substitute for thought.

9. Conclusion.

In many applications, prior information often comes in the form of constraints. For instance, masses and energies are nonnegative and finite, velocities do not exceed the speed of light, and the fraudulent portion of a healthcare invoice does not exceed the total amount of the invoice. Frequentist methods can use such constraints directly. Bayesian methods require augmenting the constraints with prior probability distributions. The difference between the frequentist and Bayesian viewpoints is that Bayesians claim to know more about how the data are generated: Frequentists claim to know that the parameter $\theta \in \Theta$, but not how θ is to be selected from Θ. Bayesians claim to know that the parameter θ is selected at random from Θ according to a prior probability distribution π known to them. Both claim to know $\mathbb{P}_\eta$, the probability distribution that the data would have if the value of θ were η, for each $\eta \in \Theta$.

In Bayesian analysis, the prior probability distribution captures the analyst's beliefs about the parameter before the data are collected. The prior is updated using the data to construct

the posterior distribution via Bayes' rule. The posterior combines the analyst's prior beliefs with information from the data. An analyst with different prior beliefs will in general arrive at a different posterior distribution.

The Bayesian framework changes the subject in a subtle way. To measure a probability in the Bayesian framework is to discover what the analyst thinks, while to measure a probability in the frequentist framework is to discover empirical regularities in the system under study. Because of the difference in interpretations of probability, the Bayesian framework allows probability statements to be made about a much larger range of phenomena. However, the relevance of an analyst's statements for others may be limited, since it hinges on the analyst's prior, which is claimed to quantify her degree of belief—but perhaps no one else's.

Bayesian and frequentist measures of uncertainty differ. For instance, mean squared error and posterior mean squared error are expectations of the same quantity, but with respect to different distributions: MSE is an expectation with respect to the distribution of the data, holding the parameter fixed; while PMSE is an expectation with respect to the posterior distribution of the parameter, holding the data fixed. Similarly, coverage probability and credible level are the chance that a set contains the parameter, but coverage probability is computed with respect to the distribution of the data, holding the parameter fixed and allowing the set to vary randomly; while credible level is computed with respect to posterior distribution of the parameter, holding the data and the set fixed and allowing the parameter to vary randomly.

When choosing between Bayesian and frequentist approaches, an analyst is advised to consider more than computational convenience: The interpretation and intended use of the results matter, and these may depend on the application. Is the parameter in question actually random? If so, is its prior distribution known? Which is the more interesting question: what would happen if Nature generated a new value of the parameter and the data happened to remain the same, or what would happen for the same value of the parameter if the measurement were repeated?

Under relatively mild conditions, the largest Bayes risk as the prior is allowed to vary is equal to the smallest maximum risk (the minimax risk) of any estimator as the parameter is allowed to vary. When the Bayes risk for a given prior is less than the minimax risk, the choice of the prior added information not present in the constraint, thereby reducing the apparent uncertainty of the Bayes estimate. When the difference between the Bayes and the minimax risk is large, one is advised to pay particularly close attention to the prior to check whether one's prior beliefs closely match the analyst's prior beliefs.

REFERENCES

[1] G.E. Backus, *Comparing hard and soft prior bounds in geophysical inverse problems*, Geophys. J., 94 (1988), pp. 249–261.
[2] G.E. Backus, *Bayesian inference in geomagnetism*, Geophys. J., 92 (1988), pp. 125–142.
[3] G.E. Backus, *Confidence set inference with a prior quadratic bound*, Geophys. J., 97 (1989), pp. 119–150.
[4] P.B. Stark, *Minimax confidence intervals in geomagnetism*, Geophys. J. Int., 108 (1992), pp. 329–338.
[5] P.B. Stark, *Inference in infinite-dimensional inverse problems: Discretization and duality*, J. Geophys. Res., 97 (1992), pp. 14055–14082.
[6] S.N. Evans and P.B. Stark, *Inverse problems as statistics*, Inverse Problems, 18 (2002), pp. R55–R97.

[7] S.N. Evans, B. Hansen, and P.B. Stark, *Minimax expected measure confidence sets for restricted location parameters*, Bernoulli, 11 (2005), pp. 571–590.
[8] P.B. Stark and L. Tenorio, *A primer of frequentist and Bayesian inference in inverse problems*, in Large-Scale Inverse Problems and Quantification of Uncertainty, L. Biegler, G. Biros, O. Ghattas, Y. Marzouk, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, eds., John Wiley & Sons, Chichester, UK, 2010, pp. 9–32.
[9] A. Tversky and D. Kahneman, *Judgment under uncertainty: Heuristics and biases*, Science, 185 (1974), pp. 1124–1131.
[10] M.L. Eaton and D.A. Freedman, *Dutch book against some 'objective priors*,' Bernoulli, 10 (2004), pp. 861–872.
[11] M.L. Eaton, *Dutch book in simple multivariate normal prediction: Another look*, in Probability and Statistics: Essays in Honor of David A. Freedman, D. Nolan and T. Speed, eds., Institute of Mathematical Statistics, Beachwood, OH, 2008, pp. 12–23.
[12] P. Diaconis and D.A. Freedman, *On the consistency of Bayes estimates*, Ann. Statist., 14 (1986), pp. 1-26.
[13] P.W. Diaconis and D. Freedman, *Consistency of Bayes estimates for nonparametric regression: Normal theory*, Bernoulli, 4 (1998), pp. 411–444.
[14] D.A. Freedman, *Some issues in the foundations of statistics*, Found. Sci., 1 (1995), pp. 19–39.
[15] P.H. Garthwaite, J.B. Kadane, and A. O'Hagan, *Statistical methods for eliciting probability distributions*, J. Amer. Statist. Assoc., 100 (2005), pp. 680–701.
[16] P.B. Stark and D.A. Freedman, *What is the chance of an earthquake?*, in Earthquake Science and Seismic Risk Reduction, NATO Sci. Ser. IV: Earth and Environmental Sciences 32, Kluwer, Dordrecht, The Netherlands, 2003, pp. 201–216.
[17] G.E. Backus, *Isotropic probability measures in infinite-dimensional spaces*, Proc. Natl. Acad. Sci. USA, 84 (1987), pp. 8755–8757.
[18] P.W. Diaconis, *The problem of thinking too much*, Bull. Amer. Acad. Arts Sci., 56 (2003), no. 3, pp. 26–38.
[19] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag, New York, 1985.
[20] E.L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed., Springer-Verlag, New York, 1998.
[21] C.M. Schafer and P.B. Stark, *Using what we know: Inference with physical constraints*, in Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology, PHYSTAT2003, Menlo Park, CA, L. Lyons, R. Mount, and R. Reitmeyer, eds., 2003, pp. 25–34.
[22] C.M. Schafer and P.B. Stark, *Constructing confidence sets of optimal expected size*, J. Amer. Statist. Assoc., 104 (2009), pp. 1080–1089.
[23] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York, 1986.
[24] R. Berk, L. Brown, and L. Zhao, *Statistical inference after model selection*, J. Quant. Criminol., 26 (2010), pp. 217–236.
[25] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao, *Valid post-selection inference*, Ann. Statist., 41 (2013), pp. 802–837.
[26] Y. Benjamini and Y. Yekutieli, *False discovery rate controlling confidence intervals for selected parameters*, J. Amer. Statist. Assoc., 100 (2005), pp. 71–80.
[27] A. Weinstein, W. Fithian, and Y. Benjamini, *Selection adjusted confidence intervals with more power to determine the sign*, J. Amer. Statist. Assoc., 108 (2013), pp. 165–176.