LA-13412-MS

Plausible Inference and the Interpretation of Quantitative Data

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

Edited by Maco Stewart, Group CIC-1
Prepared by Sharon Hurdle, Group NIS-7

This work was supported by the U.S. Department of Energy, Office of Nonproliferation and National Security, Office of Research and Development.

An Affirmative Action/Equal Opportunity Employer

LA-13412-MS

UC-700 and UC-705
Issued: February 1998

Plausible Inference and the Interpretation of Quantitative Data

Charles W. Nakhleh

Los Alamos
NATIONAL LABORATORY
Los Alamos, New Mexico 87545

Contents

Plausible Inference and the Interpretation of Quantitative Data

by

Charles W. Nakhleh

Abstract

The analysis of quantitative data is central to scientific investigation. Probability theory, which is founded on two rules, the sum and product rules, provides the unique, logically consistent method for drawing valid inferences from quantitative data. This primer on the use of probability theory is meant to fulfill a pedagogical purpose. The discussion begins at the foundation of scientific inference by showing how the sum and product rules of probability theory follow from some very basic considerations of logical consistency. We then develop general methods of probability theory that are essential to the analysis and interpretation of data. We discuss how to assign probability distributions using the principle of maximum entropy, how to estimate parameters from data, how to handle nuisance parameters whose values are of little interest, and how to determine which of a set of models is most justified by a data set. All these methods are used together in most realistic data analyses. Examples are given throughout to illustrate the basic points.

I. Introduction

Whether we are interested in monitoring nuclear facilities, studying the dynamics of stellar interiors, or understanding the complicated chemical processes in the living cell, the ultimate objective of scientific study is to increase our quantitative knowledge of a physical phenomenon or set of phenomena. The cornerstone of this method is the collection and interpretation of quantitative data.

The acquisition of numerical data is the province of the experimental scientist and instrument developer. In our age of rapidly changing technologies, this job requires special competencies and full-time attention to technological innovations. But the development of new instruments

and technologies addresses only half the problem. The goal of the entire enterprise is not simply to acquire data but rather to gain knowledge. Poorly understood data are of little value; moreover, in practice it is common to meet a situation where the available data are related only indirectly to the phenomenon of interest. The problem of interpreting uncertain or incomplete data so as to draw valid conclusions is one of *logical inference*, and the concepts, methods, and algorithms developed to address this problem lie in the domain of the theoretical scientist.

The purpose of this primer is to develop as concisely as possible the essential theoretical ideas necessary for extracting valid inferences from numerical data. The emphasis throughout is on fundamental ideas and logical structure, and the examples given are chosen to illuminate the simplicity and cogency of the most important ideas. As will be seen below, consistent inference requires nothing more than the rules of probability theory—correctly understood and consistently applied.[1]

These ideas are not themselves new but they are not as thoroughly used by scientists as they should be. It was for this reason that this primer was felt to be pedagogically useful. Much of the recent impetus to apply probability theory to the analysis and interpretation of data (as well as to many other interesting and important areas) is due to E. T. Jaynes,[2] whose works are well worth the time and effort invested in reading them. An excellent account of the issues discussed here, and many other useful and interesting aspects of data analysis, is the book by Sivia.[3]

II. Scientific Inference: Foundations

Let us describe intuitively the problem we are facing when dealing with a situation of scientific inference. There is a phenomenon about which we would like to make some statements *A, B, C,* etc. For example, we might be interested in understanding the effect of a certain drug therapy on a given illness. We would like to determine whether the therapy has a positive effect, a negative effect, or no effect at all; and we would like to attach to each statement a measure of our estimate of its validity. We base our inferences on any evidence *E* that we have: we will assume for discussion that *E* consists of a data set *D* and any relevant background information *I* (which could include other data sets, theoretical ideas, conjectures, etc.).

---

[1] The methods to be developed below are widely called "Bayesian methods" in the statistical literature. But this terminology is misleading for, as will be amply clear, all we are actually doing is applying the rules of probability theory. For these reasons, I will try to avoid the term "Bayesian," although some slippage is probably unavoidable.

[2] Jaynes has a majestic work in progress on probability theory and its applications to which this primer is much indebted. Draft versions of *Probability Theory: The Logic of Science* are available on the World-Wide Web at http://bayes.wustl.edu. A very useful collection of Jaynes' papers is R. D. Rosenkrantz, ed., *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics* (Dordrecht, Holland: D. Reidel, 1983).

[3] D. S. Sivia, *Data Analysis: A Bayesian Tutorial* (Oxford: Clarendon, 1996).

We would like to develop a method of inference that enables us to answer the following questions:

> Given the data *D* and any relevant background information *I,* what is the probability that proposition *A* is true? That proposition *B* is true? That proposition *C* is true?…

Any worthwhile theory of scientific inference must provide some way of answering these questions.

The term *probability* is used in its intuitive sense: it is a measure of the plausibility of a proposition based on the given evidence. The probability of a proposition *A* given evidence *E* is a measure of our assessment of the likelihood that *A* is true based upon evidence *E.* A probability, in other words, is a measure of our state of knowledge.[4] Notice that we never discuss the probability of a proposition in and of itself. We always refer to the probability of a proposition *given* or *conditioned upon* some evidence, which should be stated as explicitly as possible, because we always determine the likelihood of a proposition with the aid of some information, even if it is just lumped implicitly into the background information *I*.

A. The calculus of plausible reasoning

In 1946, Richard T. Cox showed how this intuitive formulation of the inference problem could be mathematically formulated in a powerful and flexible manner.[5] Cox assumed the following minimal requirements:

1. The plausibility of proposition *A* given information *I* can be represented by a real number. The plausibilities are ordered so that more plausible propositions are represented by greater real numbers than are less plausible propositions.
2. Our reasoning with plausibilities must be logically consistent.

These assumptions are hardly restrictive. As mentioned above, the gathering of numerical data lies at the heart of the scientific approach; it is reasonable to expect, then, that our measure of plausibility in such problems can be given numerical form.[6] The insistence on logical consistency is the key to the whole development that follows. Cox's aim (and ours) is to formulate a theory of rational, honest inference. There is nothing more common in daily life

---

[4] In philosophical terminology, probabilities are *epistemological* objects.

[5] R. T. Cox, "Probability, Frequency, and Reasonable Expectation," *American Journal of Physics* **14** (1946), pp. 1–13.

[6] This assumption has been questioned, although numerical representations of plausibilities are natural in quantitative work. For discussions, see Jaynes, *Logic of Science*, Appendix A, and I. J. Good, *Probability and the Weighing of Evidence* (London: Charles Griffin, 1950).

than to jump to conclusions based on deep-seated emotional desires, unstated preferences, inadequate evidence, and (possibly unknown) biases. We have set ourselves a much different problem: we take the greatest pains to explicitly note every scrap of evidence that we allow ourselves to use in the inference and then we require rigorous logical consistency in carrying out the inference—no matter how disagreeable the results.

Under these assumptions (using the notation of Appendix A), Cox proved the following startling result: the plausibilities that we started with can be mapped in a one-to-one fashion into another set of numerical functions that *must* obey the following rules:

$$\text{(Product rule)} \qquad p(AB|E) = p(A|E)p(B|AE) \tag{1}$$

$$\text{(Sum rule)} \qquad p(A|E) + p(\overline{A}|E) = 1, \tag{2}$$

where AB here represents the logical product of propositions A and B, $\overline{A}$ the logical complement (negation) of A, and the vertical bar represents conditioning on the evidence E. (Appendix B contains an outline of Cox's proof.)

These two rules are just the ordinary product and sum rules of probability theory, although they appear here in a much different light. What Cox proved, in essence, is that any method of plausible reasoning that is to be logically consistent is equivalent to probability theory (which is entirely based on the sum and product rules). In other words, probability theory is the only consistent way to reason in the presence of incomplete or uncertain information, which is always the situation we find in science.

From this conclusion it follows that any method of drawing conclusions from numerical data that does not derive from the sum and product rules of probability theory is logically inconsistent. Fortunately, a large portion of current statistical practice (e.g., Student's t distribution, maximum likelihood methods) can be shown to be special cases of the sum and product rules under certain simplifying assumptions. But some common statistical methods (e.g., Blackman-Tukey methods of spectral analysis, data filtering) do not appear to follow from the sum and product rules, and could well be logically inconsistent. They should therefore be viewed with skepticism.[7]

As mentioned in Appendix A, probability theory can be viewed as an extension of logical reasoning to deal with propositions that are merely plausible rather than certain. Recall from Appendix A that, in a given logical environment *I, A* implies *B* if the two propositions satisfy

[7] In the cases where ad hoc methods have been used, probability theory, as it turns out, solves the problem more efficiently and provides superior results.

$A = AB$. Inserting this into the product rule, Eq. (1), we find that $p(B|AI) = 1$. In other words, *B* is certain given *A* in the logical environment *I*. This result shows that certainty is represented by a probability of unity, and hence impossibility, by the sum rule, is represented by a probability of zero. Exclusion can also be treated similarly. For if *A* excludes *B* in logical environment *I,* then $AB = 0$, and the product rule implies $p(B|AI) = 0$. Clearly, then, $0 < p(B|AI) < 1$ represents a continuous spectrum of cases where *B* ranges from impossibility to certainty given *A*.

As will become clear in the discussion that follows, the powerful extension of classical logical methods provided by probability theory provides just the type of structured method needed in scientific inference.

From this point on, Cox's theorems will recede somewhat into the background. All of the results derived below will follow directly from the product and sum rules of probability theory. But we should keep in mind that these rules are the only consistent framework for plausible reasoning.

B. The basic principle of logical inference

Cox's results now allow us to formulate the basic principle of inference precisely. Let *A* be a proposition we are interested in for some given phenomenon. Let E_1, E_2, … be all the relevant evidence we can muster related to *A*. This evidence could include many different sets of experimental data, theoretical hunches, background information, anything. Following Jaynes,[8] we then have:

The fundamental principle of inference
To form an estimate of the likely truth or falsity of *A* on the basis of evidence E_1, E_2, …, compute the probability $p(A|E_1E_2\ldots)$ using the sum and product rules of probability theory.

According to Cox's theorem, this method provides the unique, logically consistent method of inferring the probable truth of *A* on the basis of the evidence E_1, E_2, …. If we desire to represent our intuitive notion of probability numerically, and if we desire to maintain logical consistency, Cox informs us that we must use this principle. There is no other choice.

C. Some useful results: Bayes' theorem and marginalization

For future reference, this section contains several results whose use may not be immediately apparent but that will prove to be indispensable in realistic data analysis problems. For pedagogical purposes, you may find it advisable to omit this section on a first reading, referring to

[8] Jaynes, *Logic of Science,* Chapter 4.

it when necessary. In any case, all the following results follow directly from Boolean algebra and the sum and product rules of probability theory.[9]

Bayes' theorem. Let *A, B,* and *I* be propositions. Then:

$$p(B|AI) = \frac{p(B|I)p(A|BI)}{p(A|I)}. \tag{3}$$

Proof. Because Boolean algebra is commutative, we have $p(AB|I) = p(BA|I)$. The theorem results from the application of the product rule to each side of this equation.

Boolean algebra and the rules of probability theory can be combined to yield another important theorem.

Theorem. (Generalized sum rule.) Define *A, B,* and *I* as above. Then:

$$p(A + B|I) = p(A|I) + p(B|I) - p(AB|I). \tag{4}$$

Proof. See Appendix C.

Corollary. If *A* and *B* are mutually exclusive (i.e., $AB = 0$), then $p(AB|I) = 0$ and:

$$p(A + B|I) = p(A|I) + p(B|I). \tag{5}$$

Now we are in a position to prove the important property called *marginalization*. Let $(A_1, A_2, \ldots, A_n)$ be a set of exhaustive, mutually exclusive propositions (Appendix A). Let *B* be an arbitrary proposition. We then have the following:

Theorem. (Marginalization.) For $(A_1, A_2, \ldots, A_n)$ exhaustive and mutually exclusive:

$$p(B|I) = \sum_{i=1}^{n} p(BA_i|I) = \sum_{i=1}^{n} p(B|A_iI)p(A_i|I). \tag{6}$$

Proof. The theorem follows by induction from the corollary and the product rule.

As will be seen below in the section on the analysis of data, marginalization allows us to deal with "nuisance parameters"—parameters that unavoidably enter into the analysis of a physical

[9] For a careful development based on axioms essentially equivalent to ours, see Good, *Probability and the Weighing of Evidence*, Chapter 3.

situation but whose values are unimportant to the problem at hand—by summing (or integrating) over them. This device, which follows directly from the basic rules of probabilistic inference, is completely unknown in classical statistical practice and is a significant advantage of the probabilistic approach.

D. Assigning probabilities by maximum entropy

Cox's theorem informs us that probability theory (i.e., the sum and product rules) provides the only logically consistent formalism for plausible inference. But it does not provide us with any principles for assigning numerical probabilities in a given problem. For any given set of propositions, there are infinitely many numerical probability schemes that will satisfy the basic sum and product rules. Each particular numerical assignment for a probability, $p(A|BI)$, represents a different degree of logical implication between *A* and *B*. In other words, each different numerical value corresponds to a different logical environment. What we need is a method or methods for assigning probability distributions in a particular logical environment. Without such principles, it is difficult to proceed very far in realistic inference situations.

There are several approaches to assigning probabilities. One of the most useful lines of attack on this problem was initiated in the work of Shannon[10] on information theory in the 1940s and was expanded by Jaynes[11] and others in the 1950s and subsequently.[12] Space limitations preclude all but the most cursory discussion of this interesting and important development. Suppose that on background information *I,* one is able to define a set $(A_1, A_2, \ldots, A_n)$ of mutually exclusive, exhaustive propositions. The probability distribution associated with this set of propositions in the logical environment *I* is $\{p_i = p(A_i|I)\}_{i=1}^{n}$. Shannon proved under very general conditions that there is a unique, positive, additive numerical measure of the spread or uncertainty of a probability distribution. This unique measure is called the *entropy* (or *information entropy)* of the distribution and is given by the formula:

$$H = -\sum p_i \log p_i \,. \tag{7}$$

The logarithm can be taken to any base; a common choice is base 2, and the entropy is then measured in bits. Other common choices are base 10, or base *e* (i.e., natural logarithms). Of course, the probabilities are always constrained by the sum rule, which in this case means that they must sum to unity.

---

[10] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal* **27** (1948), pp. 379–423, 623–656.

[11] See the first two papers in Rosenkrantz's collection.

[12] An advanced work that repays careful study is S. Kullback, *Information Theory and Statistics* (New York: Wiley, 1959).

Example 1. (Discrete distribution.) Let $p_i = \frac{1}{n}$, then $H = \log n$.

The entropy defined by Eq. (7) is essentially a logarithmic measure of the amount of choice involved in the partition $(A_1, A_2, \ldots, A_n)$. Note that if any one of the propositions in the partition is certain, the entropy is zero. This measure is additive in the following sense: if the partition can be "factored" into a product of independent subpartitions, the total entropy is obtained simply by adding the entropies of the subpartitions. In essence, this is simply a restatement of the well-known property of logarithms that the log of a product is equal to the sum of the logs of the factors. If the subpartitions are not independent, one needs to incorporate a term that reflects the conditioning of one subpartition on the other. The explicit formulas for these cases are important in developing information theory, but will not be considered further here.

Example 2. Consider the experiment of rolling two dice simultaneously. There are 36 possible outcomes that we assume are equally weighted ("fair" dice). The total entropy is given by $H = \log 36 = \log(6 \times 6) = \log 6 + \log 6 = 2 \log 6 = 5.2$ bits. This result is twice the entropy of the roll of a single die.

By taking limits it is possible to pass from a discrete distribution to a continuous distribution. The most commonly occurring continuous distribution is the Gaussian distribution, given by:

$$p(x|I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(x-\mu)^2 \;\; 2\sigma^2\right], \tag{8}$$

where μ, σ are real parameters describing the mean and standard deviation of the distribution, and x is a real variable. The sum rule in the continuous case generalizes to an integral, with the integral from $-\infty$ to ∞ equal to unity.

Example 3. For continuous distributions, the sum in the entropy formula generalizes to an integral and it is possible to show that the entropy of a Gaussian distribution is $H = \log(\sigma\sqrt{2\pi e})$ (here e is the base of the natural logarithms). Notice that the correct measure of the spread of the distribution is the *logarithm* of the standard deviation, rather than the standard deviation itself.

The maximum entropy method of assigning probabilities may be sketched as follows. Suppose that the logical environment I provides a set of constraints on the probability distribution. Then the maximum entropy distribution is the distribution obtained by maximizing the entropy [Eq. (7)] *subject* to the constraints. (The technical tools needed for finding the constrained maximum are Lagrange multipliers and the calculus of variations.)

Example 4. Suppose that *I* gives us no constraints on the *n* possible outcomes (except, of course, that the probabilities sum to unity). Then it can be shown that the distribution that has maximum entropy is simply the uniform distribution of Example 1, i.e., $p_i = 1/n$. What this result means is simply that if *I* provides no reason to favor one outcome over the other, the distribution that best expresses our lack of knowledge is the distribution with each possible outcome equally weighted, an intuitively pleasing result.

Example 5. Suppose, on information *I,* we expect that the distribution will have a finite mean and standard deviation, call them μ, σ. Otherwise, we assume nothing. It can be shown that the distribution that maximizes the entropy subject to these constraints is simply the Gaussian distribution of Eq. (8).

As the preceding examples indicate, maximum entropy distributions are the most conservative distributions in the following sense: a maximum entropy distribution is the most "spread-out" distribution possible subject to very explicit constraints. It assigns the maximum possible weight to each outcome subject to the constraints. The use of maximum entropy distributions is therefore a useful preventive measure against introducing any implicit assumptions in an inference. This property is the reason that Gaussian distributions are so successful in a wide range of practical applications. Example 4 indicates that if you postulate that the distribution has a finite variance, then there is no distribution that can be spread out more (or, equivalently, assumes less about the situation) than the Gaussian. It's as safe a distribution as can be used in any situation where you have good reason to believe that the variance of the distribution is finite but where you have no other particular information—which is often the case in describing experimental uncertainties, for example.

This has been a very brief introduction to a large and important topic. Much more detailed and intensive discussion of entropy and the maximum entropy method can be found in the literature.[13]

III. Scientific Inference: The Analysis of Data

The formalism for inference developed above can be used directly for the analysis and interpretation of data.[14] Suppose that we have a physical phenomenon about which we have background information *I,* data *D,* and an exhaustive set of mutually exclusive hypotheses $\{H_i\}_{i=1}^{n}$. Bayes' theorem tells us to calculate the probability of H_i given *D* and *I* as

[13] See, e.g., R. D. Levine and M. Tribus, eds., *The Maximum Entropy Formalism* (Cambridge: The MIT Press, 1979), and B. Buck and V. A. Macaulay, eds., *Maximum Entropy in Action* (Oxford: Clarendon, 1991).

[14] Or in any situation in which we have to reason on the basis of incomplete or uncertain information.

$$p\left(H_i|DI\right) = \frac{p\left(H_i|I\right)p\left(D|H_iI\right)}{p\left(D|I\right)}. \tag{9}$$

By marginalization, we can express the denominator as a sum:

$$p\left(D|I\right) = \sum_{i=1}^{n} p\left(DH_i|I\right) = \sum_{i=1}^{n} p\left(D|H_iI\right)p\left(H_i|I\right), \tag{10}$$

showing that it is simply a normalization constant designed to ensure that

$$\sum_i p\left(H_i|DI\right) = 1. \tag{11}$$

We can therefore write Bayes' theorem as follows:

$$p\left(H_i|DI\right) \propto p\left(H_i|I\right)p\left(D|H_iI\right), \tag{12}$$

where the proportionality constant is fixed by Eq. (11). This relation is often expressed as:

$$\text{posterior probability} \propto \text{prior probability} \times \text{likelihood}, \tag{13}$$

in which the posterior probability (or simply the posterior) reflects our knowledge of the plausibility of the hypothesis in the new logical environment provided by the data and the background information; the prior probability (or prior) reflects our knowledge of the plausibility of the hypothesis in light of the background information; and the likelihood measures how the data update our initial plausibility assessments. The basic principle of scientific inference informs us that the posterior contains all the information we need to make the most rational judgment possible about the hypothesis in light of the data and the background information.

We can compute the relative plausibilities of hypotheses H_i and H_j by applying Bayes' theorem to each and then taking the ratio:

$$\frac{p\left(H_i|DI\right)}{p\left(H_j|DI\right)} = \frac{p\left(H_i|I\right)}{p\left(H_j|I\right)} \times \frac{p\left(D|H_iI\right)}{p\left(D|H_jI\right)}, \tag{14}$$

which gives the odds ratio of the two posteriors in terms of the ratio of the priors and the likelihood ratio. All of these results are intuitively reasonable.

In performing an inference, computing the likelihood usually causes no particular problems, but the prior is often a source of worry. There is no reason for such anxiety. The prior is simply a probability distribution like any other: all the standard methods for assigning a prior distribution are available, in particular the method of maximum entropy. It is important to explicitly take into account any constraints implied by the background information *I,* but this should be done in the assignment of any probability distribution, not just priors. As an example, highly relevant background information might simply be that we know a given parameter we are trying to estimate (e.g., a mass, charge, voltage, etc.) is positive and/or is less than some (possibly large) upper limit. Suppose that otherwise we have no reason to prefer one value of the parameter to another. Subject to these constraints, the maximum entropy prior distribution would be zero for negative values of the parameter, constant between zero and the upper limit, and zero again for values greater than the upper limit. Notice that a distribution of this type is normalizable (in the literature, such distributions are called *proper*). If we desire, we can take the upper limit to infinity at the *end* of the calculation; after doing so, conclusions that do not depend strongly on the exact value of the upper limit will come clearly to the fore. It is important that we do not take such a limit before we have completed the calculation, which could result in misleading and completely avoidable errors.

A. Parameter estimation

One of the most common practical problems in data analysis is that of estimating one or more parameters from a given (noisy) data set. Examples include the estimation of a charge, mass, angular momentum, temperature, frequency, isotopic ratio, and so on. The reader will undoubtedly be able to provide examples from other areas. In many cases of interest, the physical situation under which the data were gathered implies that several parameters are necessarily involved, only some of which may be of interest to the analyst. In such cases, the excess parameters are expressively called *nuisance* parameters. As will be seen shortly, these parameters are easily dealt with in probability theory by marginalization. A brief discussion of how the rules of probability theory can be applied to continuous parameters is given in Appendix D.

Suppose first that all the parameters in a given problem are interesting. Denote them collectively by $\theta = (\theta_1, \theta_2, \ldots, \theta_r)$. The basic principle of inference tells us that we need to compute the posterior probability of θ conditioned on the data D and background information I. The posterior is given by Bayes' theorem as

$$p(\theta|DI) = Kp(\theta|I)p(D|\theta, I), \tag{15}$$

where K is a constant determined by the normalization

$$\int_{-\infty}^{\infty} d\theta\, p(\theta|DI) = 1. \tag{16}$$

That the sum of the probabilities is unity means that we are implicitly assuming as part of I that the model with which we relate the parameters to the data is correct for some value of the parameters. In that sense, we are interpreting the data *in light of* a given model. In the model comparison discussion below, we demonstrate how probability theory addresses the question of deciding whether one model is more justified by the data than another. For now we keep the assumption that the model is correct, and proceed to investigate what the data tell us about the parameters of that model.

Example 6. (Experimental errors.) Suppose we want to determine the true value of a parameter θ from noisy data D. Suppose on background information I, we can write the relationship between D and θ as:

$$D_i = \theta + e_i, \tag{17}$$

where the $\{e_i\}$ are measurement errors. Using maximum entropy arguments as outlined above, we can most safely model our knowledge of the errors by parametrizing them as independent, Gaussian errors; any other distribution would have a lower entropy (assuming, of course, a finite error variance). For the purposes of this example, assume that background information I implies that the errors have a common and known standard deviation σ. Then the likelihood is given as:

$$p(D|\theta,\sigma,I) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(D_i-\theta)^2\right\}, \tag{18}$$

where the conditioning on σ (indicating that it is assumed to be known) is explicitly indicated. The posterior is then given by:

$$p(\theta|D,\sigma,I) = Ap(\theta|I)\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(D_i-\theta)^2\right\}, \tag{19}$$

where A is the normalization constant (i.e., independent of θ). By expanding the sum in the exponent, it is straightforward to show that the posterior can be written in the following form:

$$p(\theta|D,\sigma,I) = A' p(\theta|I) \exp\left\{-\frac{1}{2(\sigma^2/n)} \sum_{i=1}^{n} \left(\theta - \overline{D}\right)^2\right\}, \tag{20}$$

where, as usual, $\overline{D} = \sum D_i/n$ is the mean of the data, and all the terms independent of θ have been absorbed into the normalization constant. This form of the posterior easily brings out the standard result that the posterior is a Gaussian centered around the mean of the data, with a spread σ^2/n. For reasonable quantities of data, the Gaussian is going to be so sharply peaked around the data mean that the prior is relatively unimportant in the inference, especially if it is a constant (or slowly varying) over the range of interest. Notice that probability theory automatically extracts all the relevant information from the data, in this case by averaging, and discards the irrelevant parts of the data (by absorption into the normalization constant). This information extraction is an automatic benefit of the application of probabilistic inference. The data mean is—for this example—a *sufficient* statistic. Bayes' theorem automatically generates sufficient statistics, if they exist. Although the results are identical with those given by the standard statistical treatments, they appear here in a different light as simple and direct results of the product and sum rules and of the various simplifying assumptions that were made, which are not always clearly brought out in standard treatments. Moreover, this simple example can be generalized in ways that the standard methods are ill-equipped to handle.

Example 7. (Maximum likelihood estimation.) Let us take the same problem as in Example 5, but now assume, on background *I,* that our model of the data is given by:

$$D_i = f(\theta) + e_i, \tag{21}$$

where *f* is a known function (perhaps provided by a theorist) that relates the parameter of interest to the data. In this case the posterior is given by:

$$p(\theta|D,\sigma,I) = A p(\theta|I) \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(D_i - f(\theta)\right)^2\right\}. \tag{22}$$

If we assume that the posterior is slowly varying or constant in the region of interest, then the best estimate of θ, call it $\hat{\theta}$, is provided by locating the maximum value of the likelihood. The uncertainty of this estimate is easily given by investigation of the spread of the posterior in the vicinity of the maximum. The sum in the exponent of the likelihood is the usual least-squares sum. Once again, this example shows how the standard statistical algorithms—if they are logically consistent—emerge naturally from probability theory under certain simplifying assumptions.

One aspect of data analysis that probability theory deals with naturally—but which standard methods have no mechanism for handling—is the problem of nuisance parameters. One physical example of such a situation is the transmission of a signal through a medium (e.g., electromagnetic radiation through the atmosphere). The problem usually involves several parameters characterizing the interaction of the signal with the medium. If all we are interested in is inferring the properties of the signal (e.g., source strength), then the medium parameters are purely a nuisance, but they have a clear effect on the data that are collected. How are we to handle them?

Denote the parameters of interest collectively by θ, and the nuisance parameters collectively by φ. Then marginalization allows us to express the posterior for the interesting parameters as:

$$p(\theta|DI) = \int d\varphi\, p(\theta,\varphi|DI), \tag{23}$$

where the sums are generalized to integrals for continuous variables in the usual way. Bayes' theorem enables us to reverse the conditioning in the integrand:

$$p(\theta|DI) = K\int d\varphi\, p(\theta,\varphi|I)p(D|\theta,\varphi,I). \tag{24}$$

Again, the basic structure of prior × likelihood emerges under the integral sign. If, based on I, we have no reason to believe the interesting parameters are dependent on the nuisance parameters, we can factor the prior using the product rule:

$$p(\theta,\varphi|I) = p(\theta|I)p(\varphi|\theta,I) = p(\theta|I)p(\varphi|I). \tag{25}$$

Inserting this into the integral yields:

$$p(\theta|DI) = Kp(\theta|I)\int d\varphi\, p(\varphi|I)p(D|\theta,\varphi,I). \tag{26}$$

Example 8. Suppose that the conditions of Example 6 hold, except that we now have a nuisance parameter φ that enters through the data model:

$$D_i = f(\theta,\varphi) + e_i \ . \tag{27}$$

The posterior for the interesting parameter is given by:

$$p(\theta|D,\sigma,I) = Ap(\theta|I)\int d\varphi\, p(\varphi|I)\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(D_i - f(\theta,\varphi)\right)^2\right\}. \tag{28}$$

If, on background information *I,* the prior for φ is sharply peaked about some value φ_0 (i.e., if we have very good information about the value of the nuisance parameter), then in the limit the prior becomes a delta function and Eq. (28) collapses to the form of Eq. (22) with φ replaced by φ_0, which accords with our intuitive expectations. In realistic applications, the marginalization integrals often have to be computed numerically, although it is surprising how far analytical knowledge of Gaussian integrals can go in solving real problems.

Example 9. (Left as an exercise for the reader.) The conditions of the problem are the same as Example 6, except that now σ is unknown. Treat σ as a nuisance parameter and integrate it out; take the prior for σ to be uniform on $(0,\infty)$; and show that the resulting posterior distribution is the well-known Student *t* distribution for $n - 2$ degrees of freedom, where n is the number of data points.

Jaynes[15] has emphasized that in integrating out nuisance parameters, we are not losing any information. On the contrary, a simple application of marginalization and the product rule shows that

$$p(\theta|DI) = \int d\varphi\, p(\theta,\varphi|DI) = \int d\varphi\, p(\theta|\varphi,DI)p(\varphi|DI). \tag{29}$$

Notice the form of this equation. It is an average of the posterior for θ weighted by the posterior distribution for the nuisance parameter φ. That is to say, marginalization automatically estimates the nuisance parameter from the data and incorporates that information into the estimate for the interesting parameter—all of which is done extremely efficiently.

B. Model comparison

In the parameter estimation problem discussed above, a given model was specified on the basis of background information before the parameter inference was carried out. In many instances, however, one would like to decide which of several competing models is most justified by the data. Probability theory has the tools to help us with this problem as well.

The set of different models must be specified in advance to make the problem of choosing between models based on the data mathematically and conceptually well defined. It is impossible—even in principle—to determine whether a single model is the "true" model without specifying any competing alternatives. Certainly, one should not reject a specific model without also giving some alternative model or models that are to be preferred based on some data. Of course, it is entirely possible in any given case that the "true" model of a phenomenon is not included in the specified set of models. If so, probability theory then tells you which of the given models is most preferred by the data and the background information. If none of the

[15] Jaynes, *Logic of Science,* Chapter 7.

models is strongly preferred, then that is a powerful indication that new theories are needed—an inherently useful fact. Further discussion of these topics, and a comparison with traditional methods of hypothesis testing, are to be found in the excellent review article by Loredo.[16]

It is interesting to understand why the alternative models must be given explicitly.[17] Suppose to the contrary that we have only one model available, call it *M*. The plausibility of *M* given the data and any background information is given by Bayes' theorem as:

$$p(M|DI) = \frac{p(M|I)p(D|MI)}{p(D|I)}. \tag{30}$$

We can assign the prior for *M* based on *I* by, say, maximum entropy, and the likelihood of *D* given *M* and *I* can be computed from the model equations, so the numerator poses no difficulties in calculating the plausibility of *M*. But we also need to compute the denominator. In parameter estimation, this was computed by normalizing the posterior. In this context, the normalization sum becomes

$$p(D|I) = p(DM|I) + p(D\overline{M}|I) = p(D|MI)p(M|I) + p(D|\overline{M}I)p(\overline{M}|I). \tag{31}$$

The first term in the sum is no problem and can be computed like the numerator of Eq. (29). But notice the last term! It instructs us to compute the likelihood of the data given that the model *M* is false. This is impossible even in principle. You can compute a likelihood given a model, but you cannot compute a likelihood given only that a model is false. Probability theory is telling us that in order to decide the degree to which a model is justified by the data, we have to specify one or more alternative models. It makes no sense to ask if a data set determines whether a theory is "right" or "wrong" in the abstract; the data can only determine whether one theory is more justified than another. Every comparison must have at least two terms.

Denote the set of models under consideration by $(M_1, M_2, \ldots, M_r)$. Each model includes specific equations and parameters for modeling the data *D*. The parameter sets may vary from model to model. Probability theory informs us that to decide whether model *j* is preferred by the data over model *k,* we should compute the ratio of posterior probabilities (the odds ratio) according to Bayes' theorem:

---

[16] T. J. Loredo, "From Laplace to Supernova 1987A: Bayesian Inference in Astrophysics," in P. F. Fougère, ed., *Maximum Entropy and Bayesian Methods* (Dordrecht: Kluwer, 1990), pp. 81–142. Also available on the WWW at http:// bayes.wustl.edu.

[17] The following argument is due to Sivia, p. 88.

$$\frac{p\left(M_j|DI\right)}{p\left(M_k|DI\right)} = \frac{p\left(M_j|I\right)}{p\left(M_k|I\right)} \times \frac{p\left(D|M_jI\right)}{p\left(D|M_kI\right)}. \tag{32}$$

This equation shows, as usual, that the ratio of posteriors is given by the ratio of priors multiplied by a factor that indicates how the data contribute to our knowledge. The latter factor is sometimes referred to as the Bayes factor.[18]

Let the parameter set of the *j*th model be denoted by θ_j. Then marginalization and the product rule allow us to compute the Bayes factor explicitly as:

$$\frac{p\left(D|M_j\right)}{p\left(D|M_k\right)} = \frac{\int d\theta_j\, p\left(\theta_j|M_jI\right)p\left(D|\theta_jM_jI\right)}{\int d\theta_k\, p\left(\theta_k|M_kI\right)p\left(D|\theta_kM_kI\right)}\ . \tag{33}$$

Each integrand in this equation has the form of a posterior distribution familiar from parameter estimation. The model comparison procedure automatically estimates all the parameters (both nuisance and interesting) involved in each model from the data and integrates over all the possible parameter values.

Example 10. Suppose we have data *D* and two competing models to explain the data. The first model contains two independent parameters (α,β) and the second only one parameter θ. Suppose, given *I,* there is no prior reason to prefer one model to the other. With these assumptions, the odds ratio becomes:

$$\frac{p\left(M_1|DI\right)}{p\left(M_2|DI\right)} = \frac{\int d\alpha d\beta\, p\left(\alpha,\beta|M_1I\right)p\left(D|\alpha,\beta,M_1I\right)}{\int d\theta\, p\left(\theta|M_2I\right)p\left(D|\theta,M_2I\right)}. \tag{34}$$

These integrals can be computed by the methods developed above. It is important to note that because we are dealing with different numbers of different parameters, we have to carefully keep track of the various normalization constants in Eq. (33). They do not in general cancel out in computing the ratio.

IV. Conclusions

We have now developed probability theory far enough to tackle a vast array of data analysis problems. The theory we have explored is the unique, consistent method of reasoning logically in the presence of uncertainty. When viewed through the lens of probability theory, each and

[18] For a discussion of the utility of the logarithm of the Bayes factor in hypothesis testing, see Jaynes, *Logic of Science*, Chapter 4.

every data analysis problem is simply a problem of logical inference. We seek to draw valid conclusions in the logical environment provided by the data and all other relevant information. This point of view leads to clearly defined and often simple calculational schemes. All the important tools—the use of probabilities to represent states of knowledge, the product and sum rules as the basis of consistent plausible reasoning, maximum entropy distributions, marginalizing nuisance parameters—are conceptually complete and at your command. Each specific problem will require its own special calculations, often numerical, but all share a unified conceptual approach. Probability theory has armed us with a powerful array of methods to organize our thoughts and our calculations. How you choose to apply these methods is up to you.

Appendix A. The Algebra of Propositions

When we are confronted with a physical situation to study, the first thing we do—consciously or not—is to determine a set of propositions that fully describe the situation at hand. Three examples of increasing complexity are given below. An indefinite number of further examples may be easily produced.

Example A1 (coin toss)

A = "The coin will land heads up."

B = "The coin will land tails up." (= NOT A)

Example A2 (reprocessing plants)

A = "The plant is reprocessing weapon-grade plutonium."

B = "The plant is reprocessing reactor-grade plutonium."

C = "The plant is not reprocessing any fuel at all."

Example A3 (physical measurements)

A_1 = "The temperature T lies between T_1 and T_2."

A_2 = "The temperature T lies between T_2 and T_3."

⋮

A_{n-1} = "The temperature T lies between T_{n-1} and T_n."

Propositions such as these are known as true-false (TF) propositions. They are essential to scientific discourse and are so ubiquitous that they are rarely noticed. As a matter of notation, we will always denote TF propositions by italicized, capital Latin letters, and we will simply call them propositions or assertions.

Assertions obey a natural algebra based on the operations AND (logical product or conjunction), OR (logical sum or disjunction), and NOT (logical complement or negation). There is nothing mysterious about these operations; if you have ever spent any time searching a library database for an author AND a given subject, then you are familiar with them.[19] This algebra is called a Boolean algebra. We will adopt the following notation for logical products, sums, and complements:

$$\begin{aligned} &A \text{ AND } B = AB \\ &A \text{ OR } B = A + B \\ &\text{NOT } A = \overline{A} \end{aligned} \tag{A1}$$

[19] One can formally define these operations in terms of "truth tables." See any book on mathematical logic or Boolean algebra, e.g., Flora Dinkines, *Introduction to Mathematical Logic* (New York: Appleton-Century, 1964).

For completeness, note that disjunctions are assumed to be inclusive, that is to say, *A* OR *B* means "either *A* or *B* or both."

Symbolic relations for the algebra of propositions were developed in the middle of the 19th century by the English mathematician George Boole and others as part of an effort to formalize the rules of logical inference.[20] The important facts about the Boolean algebra of propositions are summarized in the following equations, which may be taken as axioms of the algebra:

(Associative)	$(AB)C = A(BC) = ABC$ $(A + B) + C = A + (B + C) = A + B + C$	(A2)
(Distributive)	$A(B + C) = AB + AC$	(A3)
(Commutative)	$AB = BA$ $A + B = B + A$	(A4)
(Idempotent)	$A^2 = AA = A$ $A + A = A$	(A5)
(Duality)	$\overline{AB} = \overline{A} + \overline{B}$ $\overline{(A + B)} = \overline{A}\,\overline{B}$ $\overline{\overline{A}} = A$	(A6)
(Identities)	$A \cdot 1 = A, \quad A + 1 = 1$ $A \cdot 0 = 0, \quad A + 0 = A$	(A7)
(Exhaustive)	$A + \overline{A} = 1$	(A8)

The symbols *A, B,* …, 1, 0, and the operations of conjunction, disjunction, negation, and equality are implicitly defined by these axioms. Notice in particular the role of the two identities 1 and 0. We may think of these as the sure proposition and the impossible proposition, respectively. Following Cox, we will call these elements of the algebra the logical truism and the logical absurdity.[21]

Intuitively, we would expect that 1 be the negation of 0 and vice versa. We can prove this formally from the axioms by substituting 0 for *A* in (A8) and using (A7). Using this result, and negating (A8), we then have $A\overline{A} = 0$, or the conjunction of any proposition with its complement is the absurdity.

[20] An interesting essay that this presentation draws heavily upon is R. T. Cox, "Of Inference and Inquiry, An Essay in Inductive Logic," in Levine and Tribus, eds., *The Maximum Entropy Formalism*.

[21] Ibid.

The following theorem provides a useful condition for proving that two assertions are equivalent.

Theorem. Let C be a given assertion. If A, B satisfy the following conditions:

$$AC = BC, \quad A + C = B + C \tag{A9}$$

then $A = B$. In words, if two assertions have the same conjunction and disjunction with a given assertion, then the two assertions are equal.

Proof. By disjoining A on both sides of the equation $AC = BC$, and simplifying the result, we find that $A = A + BC$. By repeating this process with B instead of A, we find that $B = B + AC$. We can express these two equations as $A = A + BC = (A + B)(A + C)$ and as $B = B + AC = (B + A)(B + C)$. But, by hypothesis, $A + C = B + C$, and so $A = B$.

This theorem may be used to show that the complement of a given assertion is unique, for if we assumed that two different assertions could function as a complement of A, then both would have to satisfy (A8) and its negation. But this would imply that both would have the same conjunction and disjunction with A, and the above theorem states that they must be the same assertion.

Notice that the operations of conjunction and negation alone form a *complete* set of operations, for the disjunction of two assertions may be defined, using the duality relations, in terms of negation and conjunction alone. This result applies to any logical expression, no matter how complicated.

Although unnecessary for the development of the theory, two interpretations of the axiomatic structure may be helpful. Take "1" to symbolize a given finite set of objects (e.g., the books in a library). Let "0" be the null set. Let the assertions A, B, etc., be represented by the proper subsets of 1. If we take the conjunction of A and B to be the set theoretic intersection of A and B, disjunction to be the union of sets, and negation to be the complement of A with respect to 1 (i.e., all the objects in 1 that are not in A), then we have a representation of the algebra (A2)—(A8) in terms of finite sets.

A second interpretation is in terms of digital logic circuits. The assertions are the respective inputs to and outputs from the logic gates. Let "1" represent "high" voltages (e.g., +5 V) and let "0" represent "low" voltages (e.g., 0 V). Let conjunction be represented by AND gates, disjunction by OR gates, and negation by NOT gates. This representation is of fundamental importance in the design of computers.

In this calculus, logical relations are represented by algebraic equations. Perhaps the most important aspect of logic is *implication.* In terms of Boolean algebra, we can represent implication as follows. Let us call the background information *I* that leads us to believe that one assertion implies another a *logical environment.* We shall say a particular logical environment asserts that *A* implies *B* if *A* and *B* satisfy the following logical equation:

$$A = AB. \tag{A10}$$

This equation states simply that—for this particular logical environment—wherever we have assertion *A* we can automatically include assertion *B*—it comes along for free. This encapsulates precisely what we intuitively mean when we say something implies something else. By disjoining both sides of this equation with *B* and using Eqs. (A3) and (A7), we can derive the equivalent condition:

$$B = A + B. \tag{A11}$$

Using the definition (A10) of implication, we can also prove the *transitive property of implication:* In a given logical environment, let *A* imply *B* and *B* imply *C*. Then *A* implies *C* in that same environment. The proof is left as an exercise. This property is, of course, a critical aspect of logical deduction.

The opposite of implication is *exclusion.* Intuitively, when we say that one thing excludes something else, we mean that the joint occurrence of the two things is impossible. This property can be formulated precisely as follows. In a given logical environment *I, A* and *B* are mutually exclusive if they satisfy

$$AB = 0. \tag{A12}$$

As we would expect, it is possible to show that if *A* excludes *B,* then *A* implies the complement of *B*. In the main text, we show how the notion of probability unites the concepts of implication and exclusion as opposite extremes of a continuum of plausibilities. One general way of viewing probability theory, therefore, is as an *extension* of logical analysis to handle propositions that are neither certainly true nor certainly false, but merely probable.

Finally, we explicitly illustrate the concept of an *exhaustive* set of propositions. The set $(A_1, A_2, \ldots, A_n)$ of assertions is exhaustive in a given logical environment if the disjunction of all the members of the set is equivalent to the logical truism. Symbolically, an exhaustive set is defined by:

$$A_1 + A_2 + \ldots + A_n = \sum A_j = 1. \tag{A13}$$

We may interpret the condition of exhaustion as saying that at least one of the propositions in the set must be true.

If in addition an exhaustive set of assertions is also pairwise mutually exclusive, the set is called a *partition*. The additional requirement of mutual exclusivity means that the conjunction of any pair of assertions in the partition is equivalent to the logical absurdity. In total, we may think of a partition as a set of assertions of which one, and only one, must be true. Notice that the axiom (A8) and its negation state that every assertion and its complement form a partition.

One simple relation that is often useful in dealing with partitions and hypothesis testing is the following. If $\left(A_1, A_2, \ldots, A_n\right)$ is a partition, then the complement of any proposition in the partition is given by:

$$\overline{A_i} = A_1 + A_2 + \ldots + A_n \quad (\text{exclude } A_i) \tag{A14}$$

The proof of this uses the theorem proved in the beginning of the appendix and follows along the lines of the proof that showed complements are unique.

Appendix B. Outline of Cox's Theorem

The results in the text followed directly from the sum and product rules of probability theory. Cox supplemented this axiomatic basis of probability theory by showing that these two rules followed from some very general criteria of consistency. The purpose of this appendix is to outline Cox's arguments. For a detailed discussion, see Jaynes,[22] on which this presentation is based. Cox's original paper is also very readable and concise.[23]

As mentioned in the text, Cox assumed that plausibilities could be represented by real numbers. To fix notation, denote the plausibility of a proposition A given evidence E by $A|E$. Similarly, the plausibility of the logical product of A and B given E is $AB|E$, etc. For the purposes of the discussion, we assume that some method of computing plausibilities has been given. As Cox noted in his original paper, there is a degree of indeterminacy here, for any one-to-one function of the plausibilities yields another, equivalent set of plausibilities. But we will show that, regardless of the initial choice of plausibilities, logical consistency requires that there exist a mapping of the plausibilities into other numerical quantities that obey the sum and product rules of probability theory.

How might the plausibility of AB given E be related to the individual plausibilities? To decide whether the product AB is true, given E, we might first determine the plausibility of A given E, and then the conditional plausibility of B given A and E. Since the logical product is commutative, we could also switch A and B in the preceding statement. Defining the variables x, y, as follows:

$$\begin{aligned} x &= A|E \\ y &= B|AE \end{aligned} \tag{B1}$$

the previous statements can be represented symbolically as:

$$AB|E = F\left(x, y\right), \tag{B2}$$

where F is an unknown function. What we will show is that requiring consistency in our reasoning provides a severe constraint on the form of F.

Now let's move on to the joint plausibility of three propositions A, B, and C, given E. The associativity of Boolean algebra tells us that $ABC = \left(AB\right)C = A\left(BC\right)$. According to Eq. (B2), we can work out the plausibility $ABC|E$ in two equivalent ways. First,

---

[22] Jaynes, *Logic of Science,* Chapter 2.

[23] Cox, "Probability, Frequency, and Reasonable Expectation."

$$ABC|E = (AB)C|E = F\left[AB|E, C|ABE\right] = F\left\{F\left[A|E, B|AE\right], C|ABE\right\}, \tag{B3}$$

where Eq. (B2) has been used twice in succession. But we can also work out $ABC|E$ as

$$ABC|E = A(BC)|E = F\left[A|E, BC|AE\right] = F\left\{A|E, F\left[B|AE, C|ABE\right]\right\}. \tag{B4}$$

As the notation suggests, we require logical consistency in that these two ways should yield equivalent results. With *x* and *y* defined as in Eq. (B1), and defining $z = C|ABE$, consistency then requires that *F* satisfy the functional equation:

$$F\left[F(x, y), z\right] = F\left[x, F(y, z)\right]. \tag{B5}$$

It is easy to verify that the particular solution $F(x, y) = xy$ satisfies Eq. (B5). In general, the constraint that this complicated-looking functional equation puts on the form of *F* can be deduced without an unreasonable amount of effort if we assume that *F* is differentiable. It can be shown by differentiation that there exists a continuous, monotonic function *w* that satisfies:

$$w\left[F(x, y)\right] = w(x)w(y). \tag{B6}$$

In other words, we can always define a new set of plausibilities that satisfies:

$$w(AB|E) = w(A|E)w(B|AE), \tag{B7}$$

which is the product rule.

Logical consistency also provides a relationship between the plausibility of a proposition *A,* given *E,* and the plausibility of its negation $\overline{A}$. The following heuristic discussion brings out the essential points. We are looking for a function *f* that relates these plausibilities:

$$w\left(\overline{A}|E\right) = f\left[w(A|E)\right]. \tag{B8}$$

Clearly, consistency requires that the same relation should hold if *A* and $\overline{A}$ are switched, so that:

$$w(A|E) = f\left[w\left(\overline{A}|E\right)\right]. \tag{B9}$$

Thus we are looking for a self-reciprocal function:

$$f\big(f(x)\big) = x, \tag{B10}$$

where $x = w\big(A|E\big)$. By extending this type of reasoning to more complicated Boolean expressions, Cox obtained another functional equation involving *f* whose solution he determined to be:

$$f(x) = \big(1 - x^m\big)^{1/m}, \tag{B11}$$

where *m* is an arbitrary constant. This equation yields:

$$w\big(A|E\big)^m + w\big(\overline{A}|E\big)^m = 1. \tag{B12}$$

Now notice that the product rule can be exponentiated with the result:

$$w\big(AB|E\big)^m = w\big(A|E\big)^m w\big(B|AE\big)^m. \tag{B13}$$

To complete the demonstration, define the probabilities by $p\big(A|E\big) = w\big(A|E\big)^m$. Then the final results for the sum and product rules are exactly those given in the text [Eqs. (1, 2)]. So no loss of generality results from putting $m = 1$ in Eqs. (B11)–(B13).

Because for any Boolean algebra the operations of conjunction and negation form a complete set (Appendix A), the sum and product rules are sufficient to compute the probability of any Boolean expression, no matter how involved.

What do these conclusions demonstrate? They show that if we have any reasonable scheme for assigning numerical plausibilities to propositions, logical consistency demands that there exist a one-to-one transformation of these plausibilities to another set of numerical functions that obey the rules of probability theory. In other words, any consistent scheme of assigning plausibilities must be *equivalent* to a scheme that obeys the rules of probability theory.

Very basic considerations of consistency and plausibility—which no acceptable theory of logical inference could violate—imply that any acceptable theory which purports to represent states of knowledge by numerical plausibilities in a consistent fashion is at bottom just probability theory.

Appendix C. Proof of the Generalized Sum Rule

This appendix is devoted to a proof of the generalized sum rule (Eq. 4). By duality, we can write $A + B = \overline{\overline{A}\,\overline{B}}$; combining this with the sum and product rules allows us to state that:

$$p\left(A + B|I\right) = 1 - p\left(\overline{A}\,\overline{B}|I\right) = 1 - p\left(\overline{A}\middle|\overline{B}I\right)p\left(\overline{B}|I\right). \tag{C1}$$

The sum rule then gives $p\left(\overline{A}\middle|\overline{B}I\right) = 1 - p\left(A\middle|\overline{B}I\right)$, which, when inserted into Eq. (C1), gives:

$$p\left(A + B|I\right) = 1 - \left[1 - p\left(A\middle|\overline{B}I\right)\right]p\left(\overline{B}|I\right) = p\left(B|I\right) + p\left(A\overline{B}|I\right), \tag{C2}$$

where the product rule has been used to recombine the last two terms. Now use the sum and product rules again to expand $p\left(A\overline{B}|I\right) = p\left(\overline{B}|AI\right)p\left(A|I\right) = p\left(A|I\right) - p\left(AB|I\right)$, and we finally get the generalized sum rule:

$$p\left(A + B|I\right) = p\left(A|I\right) + p\left(B|I\right) - p\left(AB|I\right). \tag{C3}$$

Appendix D. Continuous Parameters

Let θ be a continuous real parameter. Defining two propositions by

$$\begin{aligned} A_1 &: -\infty < \theta \le \theta_1 \\ \overline{A}_1 &: \theta_1 < \theta < \infty, \end{aligned} \tag{D1}$$

we can write for the probability of A_1:

$$p\left(A_1|I\right) \equiv F\left(\theta_1|I\right), \tag{D2}$$

where F is a function called the *cumulative distribution function* (cdf). (This F is not to be confused with the F of Appendix B.) If F is differentiable, then the fundamental theorem of calculus implies that we have

$$F\left(\theta_1|I\right) = \int_{-\infty}^{\theta_1} d\theta\, f\left(\theta|I\right), \tag{D3}$$

where

$$f\left(\theta|I\right) \equiv \frac{dF}{d\theta} \tag{D4}$$

is called the *probability distribution function* (pdf). The sum rule implies the normalization

$$p\left(A_1|I\right) + p\left(\overline{A}_1|I\right) = \int_{-\infty}^{\infty} d\theta\, f\left(\theta|I\right) = 1, \tag{D5}$$

indicating that the sums generalize to integrals in the usual way.

By abuse of notation, we write $f\left(\theta|I\right) = p\left(\theta|I\right)$, which permits us to write

$$p\left(\theta_1 \le \theta \le \theta_1 + d\theta_1|I\right) = p\left(\theta_1|I\right)d\theta_1 \tag{D6}$$

for the probability of the proposition H: $\theta_1 \le \theta \le \theta_1 + d\theta_1$. These equations can be easily generalized to several continuous parameters, with single integrals replaced by multiple integrals. By writing the probabilities in this way, it is straightforward to show that Bayes' theorem takes the same form for pdfs (because the differentials in numerator and denominator cancel) as for probabilities of propositions. The product rule also carries over to pdfs as expected.

Los Alamos
NATIONAL LABORATORY
Los Alamos, New Mexico 87545