Probability Theory:
The Logic of Science

by

E. T. Jaynes

Wayman Crow Professor of Physics

Washington University

St. Louis, MO 63130, U. S. A.

Dedicated to the Memory of Sir Harold Jeffreys,

who saw the truth and preserved it.

Fragmentary Edition of Sept 1993

PROBABILITY THEORY – THE LOGIC OF SCIENCE

Short Contents

PROBABILITY THEORY – THE LOGIC OF SCIENCE

Long Contents

PART A – PRINCIPLES and ELEMENTARY APPLICATIONS

PREFACE

The following material is addressed to readers who are already familiar with applied mathematics at the advanced undergraduate level or preferably higher; and with some field, such as physics, chemistry, biology, geology, medicine, economics, sociology, engineering, operations research, *etc.*, where inference is needed.† A previous acquaintance with probability and statistics is not necessary; indeed, a certain amount of innocence in this area may be desirable, because there will be less to unlearn.

We are concerned with probability theory and all of its conventional mathematics, but now viewed in a wider context than that of the standard textbooks. Every Chapter after the first has "new" (*i.e.,* not previously published) results that we think will be found interesting and useful. Many of our applications lie outside the scope of conventional probability theory as currently taught. But we think that the results will speak for themselves, and that something like the theory expounded here will become the conventional probability theory of the future.

History: The present form of this work is the result of an evolutionary growth over many years. My interest in probability theory was stimulated first by reading the work of Harold Jeffreys (1939) and realizing that his viewpoint makes all the problems of theoretical physics appear in a very different light. But then in quick succession discovery of the work of R. T. Cox (1946), C. E. Shannon (1948) and G. Pólya (1954) opened up new worlds of thought, whose exploration has occupied my mind for some forty years. In this much larger and permanent world of rational thinking in general, the current problems of theoretical physics appeared as only details of temporary interest.

The actual writing started as notes for a series of five lectures given at Stanford University in 1956, expounding the then new and exciting work of George Pólya on "Mathematics and Plausible Reasoning". He dissected our intuitive "common sense" into a set of elementary qualitative desiderata and showed that mathematicians had been using them all along to guide the early stages of discovery, which necessarily precede the finding of a rigorous proof. The results were much like those of James Bernoulli's "Art of Conjecture" (1713), developed analytically by Laplace in the late 18'th Century; but Pólya thought the resemblance to be only qualitative.

However, Pólya demonstrated this qualitative agreement in such complete, exhaustive detail as to suggest that there must be more to it. Fortunately, the consistency theorems of R. T. Cox were enough to clinch matters; when one added Pólya's qualitative conditions to them the result was a proof that, if degrees of plausibility are represented by real numbers, then there is a uniquely determined set of quantitative rules for conducting inference. That is, any other rules which conflict with them will necessarily violate an elementary desideratum of rationality or consistency.

But the final result was just the standard rules of probability theory, given already by Bernoulli and Laplace; so why all the fuss? The important new feature was that these rules were now seen as uniquely valid principles of logic in general, making no reference to "chance" or "random variables"; so their range of application is vastly greater than had been supposed in the conventional probability theory that was developed in the early twentieth Century. As a result, the imaginary distinction between "probability theory" and "statistical inference" disappears, and the field achieves not only logical unity and simplicity, but far greater technical power and flexibility in applications.

In the writer's lectures, the emphasis was therefore on the quantitative formulation of Pólya's viewpoint, so it could be used for general problems of scientific inference, almost all of which arise out of incomplete information rather than "randomness". Some personal reminiscences about Pólya and this start of the work are in Chapter 5.

† By "inference" we mean simply: deductive reasoning whenever enough information is at hand to permit it; inductive or plausible reasoning when – as is almost invariably the case in real problems – the necessary information is not available. Thus our topic is the optimal processing of incomplete information.

But once the development of applications started, the work of Harold Jeffreys, who had seen so much of it intuitively and seemed to anticipate every problem I would encounter, became again the central focus of attention. My debt to him is only partially indicated by the dedication of this book to his memory. Further comments about his work and its influence on mine are scattered about in several Chapters.

In the years 1957–1970 the lectures were repeated, with steadily increasing content, at many other Universities and research laboratories.‡ In this growth it became clear gradually that the outstanding difficulties of conventional "statistical inference" are easily understood and overcome. But the rules which now took their place were quite subtle conceptually, and it required some deep thinking to see how to apply them correctly. Past difficulties which had led to rejection of Laplace's work, were seen finally as only misapplications, arising usually from failure to define the problem unambiguously or to appreciate the cogency of seemingly trivial side information, and easy to correct once this is recognized. The various relations between our "extended logic" approach and the usual "random variable" one appear in almost every Chapter, in many different forms.

Eventually, the material grew to far more than could be presented in a short series of lectures, and the work evolved out of the pedagogical phase; with the clearing up of old difficulties accomplished, we found ourselves in possession of a powerful tool for dealing with new problems. Since about 1970 the accretion has continued at the same pace, but fed instead by the research activity of the writer and his colleagues. We hope that the final result has retained enough of its hybrid origins to be usable either as a textbook or as a reference work.

In view of the above, we repeat the sentence that Charles Darwin wrote in the Introduction to his *Origin of Species*: "I hope that I may be excused for entering on these personal details, as I give them to show that I have not been hasty in coming to a decision." But it might be thought that work done thirty years ago would be obsolete today. Fortunately, the work of Jeffreys, Pólya and Cox was of a fundamental, timeless character whose truth does not change and whose importance grows with time. Their perception about the nature of inference, which was merely curious thirty years ago, is very important in a half–dozen different areas of science today; and it will be crucially important in all areas 100 years hence.

Foundations: From thirty years of experience with its applications in hundreds of real problems, our views on the foundations of probability theory have evolved into something quite complex, which cannot be described in any such simplistic terms as "pro–this" or "anti–that". For example our system of probability could hardly, in style, philosophy, and purpose, be more different from that of Kolmogorov. What we consider to be fully half of probability theory as it is needed in current applications – the principles for assigning probabilities by logical analysis of incomplete information – is not present at all in the Kolmogorov system.

Yet when all is said and done we find ourselves, to our own surprise, in agreement with Kolmogorov and in disagreement with his critics, on nearly all technical issues. As noted in Appendix A, each of his axioms turns out to be, for all practical purposes, derivable from the Pólya–Cox desiderata of rationality and consistency. In short, we regard our system of probability as not contradicting Kolmogorov's; but rather seeking a deeper logical foundation that permits its extension in the directions that are needed for modern applications. In this endeavor, many problems have been solved, and those still unsolved appear where we should naturally expect them: in breaking into new ground.

As another example, it appears at first glance to everyone that we are in very close agreement with the de Finetti system of probability. Indeed, the writer believed this for some time. Yet when all is said and done we find, to our own surprise, that little more than a loose philosophical

‡ Some of the material in the early Chapters was issued in 1958 by the Socony–Mobil Oil Company as Number 4 in their series "Colloquium Lectures in Pure and Applied Science".

agreement remains; on many technical issues we disagree strongly with de Finetti. It appears to us that his way of treating infinite sets has opened up a Pandora's box of useless and unnecessary paradoxes; nonconglomerability and finite additivity are examples discussed in Chapter 15.

Infinite set paradoxing has become a morbid infection that is today spreading in a way that threatens the very life of probability theory, and requires immediate surgical removal. In our system, after this surgery, such paradoxes are avoided automatically; they cannot arise from correct application of our basic rules, because those rules admit only finite sets and infinite sets that arise as well–defined and well–behaved limits of finite sets. The paradoxing was caused by (1) jumping directly into an infinite set without specifying any limiting process to define its properties; and then (2) asking questions whose answers depend on how the limit was approached.

For example, the question: "What is the probability that an integer is even?" can have any answer we please in (0, 1), depending on what limiting process is to define the "set of all integers" (just as a conditionally convergent series can be made to converge to any number we please, depending on the order in which we arrange the terms).

In our view, an infinite set cannot be said to possess any "existence" and mathematical properties at all – at least, in probability theory – until we have specified the limiting process that is to generate it from a finite set. In other words, we sail under the banner of Gauss, Kronecker, and Poincaré rather than Cantor, Hilbert, and Bourbaki. We hope that readers who are shocked by this will study the indictment of Bourbakism by the mathematician Morris Kline (1980), and then bear with us long enough to see the advantages of our approach. Examples appear in almost every Chapter.

Comparisons: For many years there has been controversy over "frequentist" versus "Bayesian" methods of inference, in which the writer has been an outspoken partisan on the Bayesian side. The record of this up to 1981 is given in an earlier book (Jaynes, 1983). In these old works there was a strong tendency, on both sides, to argue on the level of philosophy or ideology. We can now hold ourselves somewhat aloof from this because, thanks to recent work, there is no longer any need to appeal to such arguments. We are now in possession of proven theorems and masses of worked–out numerical examples demonstrating the facts of actual performance. As a result, the superiority of Bayesian methods is now a thoroughly demonstrated fact in a hundred different areas. We point this out in some detail whenever it makes a substantial difference in the final results. Thus we continue to argue vigorously for the Bayesian methods; but we ask the reader to note that our arguments now proceed by citing facts rather than proclaiming a philosophical or ideological position.

However, neither the Bayesian nor the frequentist approach is universally applicable, so in the present more general work we take a broader view of things. Our theme is simply: *Probability Theory as Extended Logic.* The "new" perception amounts to the recognition that the mathematical rules of probability theory are not merely rules for calculating frequencies of "random variables"; they are also the unique consistent rules for conducting inference (*i.e.* plausible reasoning) of any kind, and we shall apply them in full generality to that end.

It is true that all "Bayesian" calculations are included automatically as particular cases of our rules; but so are all "frequentist" calculations. Nevertheless, our basic rules are broader than either of these, and in many applications our calculations do not fit into either category.

To explain the situation as we see it presently: The traditional "frequentist" methods which use only sampling distributions are usable and useful in many particularly simple, idealized problems; but they represent the most proscribed special cases of probability theory, because they presuppose conditions (independent repetitions of a "random experiment" but no relevant prior information) that are hardly ever met in real problems. This approach is quite inadequate for the current needs of science.

In addition, frequentist methods provide no technical means to eliminate nuisance parameters or to take prior information into account, no way even to use all the information in the data when sufficient or ancillary statistics do not exist. Lacking the necessary theoretical principles, they force one to "choose a statistic" on intuitive grounds rather than from probability theory, and then to invent arbitrary *ad hoc* devices (such as unbiased estimators, confidence intervals, power functions, tail–area significance tests) not contained in the rules of probability theory. But, as Cox's theorems guarantee, such devices always generate inconsistencies or absurd results when applied far enough; we shall see dozens of examples.

Bayesian methods represent a great generalization, and are adequate to deal with what we might call "well–developed" problems of inference. As Harold Jeffreys demonstrated, they have a superb analytical apparatus, able to deal effortlessly with the technical problems on which frequentist methods fail. Therefore they enable us to solve problems of far greater complexity than can be discussed at all in frequentist terms. One of our main purposes is to show how all this capability was contained already in the simple product and sum rules of probability theory interpreted as extended logic, with no need for – indeed, no room for – any *ad hoc* devices.

But before Bayesian methods can be used, a problem must be developed beyond John Tukey's "exploratory phase" to the point where it has enough structure to determine all the needed apparatus (a model, sample space, hypothesis space, prior probabilities, sampling distribution). Almost all scientific problems pass through an initial exploratory phase in which we have need for inference, but the frequentist assumptions are invalid and the Bayesian apparatus is not yet available. Indeed, some of them never evolve out of the exploratory phase. Problems at this level call for more primitive means of assigning probabilities directly out of our incomplete information.

For this purpose, the Principle of Maximum Entropy has at present the clearest theoretical justification and is the most highly developed computationally, with an analytical apparatus as powerful and versatile as the Bayesian one. To apply it we must define a sample space, but do not need any model or sampling distribution. In effect, entropy maximization creates a model for us out of our data, which proves to be optimal by so many different criteria* that it is hard to imagine circumstances where one would not want to use it in a problem where we have a sample space but no model.

Bayesian and maximum entropy methods differ in another respect. Both procedures yield the optimal inferences from the information that went into them, but we may choose a model for Bayesian analysis; this amounts to expressing some prior knowledge – or some working hypothesis – about the phenomenon being observed. Usually such hypotheses extend beyond what is directly observable in the data, and in that sense we might say that Bayesian methods are – or at least may be – speculative. If the extra hypotheses are true, then we expect that they will improve the Bayesian results; if they are false, the inferences will likely be worse.

On the other hand, maximum entropy is a nonspeculative procedure, in the sense that it invokes no hypotheses beyond the sample space and the evidence that is in the available data. Thus it predicts only observable facts (functions of future or past observations) rather than values of parameters which may exist only in our imagination. It is just for that reason that maximum

* These concern efficient information handling; for example, (1) The model created is the simplest one that captures all the information in the constraints (Chapter 11); (2) It is the unique model for which the constraints would have been sufficient statistics (Chapter 8); (3) If viewed as constructing a sampling distribution for subsequent Bayesian inference from new data D, the only property of the measurement errors in D that are used in that subsequent inference are the ones about which that sampling distribution contained some definite prior information (Chapter 7). Thus the formalism automatically takes into account all the information we have, but avoids assuming information that we do not have. This contrasts sharply with orthodox methods, where one does not think in terms of information at all, and in general violates both of these desiderata.

entropy is the appropriate (safest) tool when we have very little knowledge beyond the raw data; it protects us against drawing conclusions not warranted by the data. But when the information is extremely vague it may be difficult to define any appropriate sample space, and one may wonder whether still more primitive principles than Maximum Entropy can be found. There is room for much new creative thought here.

For the present, there are many important and highly nontrivial applications where Maximum Entropy is the only tool we need. The planned second volume of this work is to consider them in detail; usually, they require more technical knowledge of the subject–matter area than do the more general applications studied in this volume. All of presently known statistical mechanics, for example, is included in this, as are the highly successful maximum entropy spectrum analysis and image reconstruction algorithms in current use. However, we think that in the future the latter two applications will evolve on into the Bayesian phase, as we become more aware of the appropriate models and hypothesis spaces.

Mental Activity: As one would expect already from Pólya's examples, probability theory as extended logic reproduces many aspects of human mental activity, sometimes in surprising and even disturbing detail. In Chapter 5 we find our equations exhibiting the phenomenon of a person who tells the truth and is not believed, even though the disbelievers are reasoning consistently. The theory explains why and under what circumstances this will happen.

The equations also reproduce a more complicated phenomenon, divergence of opinions. One might expect that open discussion of public issues would tend to bring about a general concensus. On the contrary, we observe repeatedly that when some controversial issue has been discussed vigorously for a few years, society becomes polarized into two opposite extreme camps; it is almost impossible to find anyone who retains a moderate view. Probability theory as logic shows how two persons, given the same information, may have their opinions driven in opposite directions by it, and what must be done to avoid this.

In such respects, probability theory is undoubtedly telling us something about the way our own minds operate when we form intuitive judgments, of which we may not have been consciously aware. Some may feel uncomfortable at these revelations; others may see in them useful tools for psychological, sociological, or legal research.

What is 'safe'? We are not concerned here only with abstract issues of mathematics and logic. One of the main practical messages of this work is the great effect of prior information on the conclusions that one should draw from a given data set. Currently much discussed issues such as environmental hazards or the toxicity of a food additive, cannot be judged rationally if one looks only at the current data and ignores our prior information about the phenomenon. As we demonstrate, this can lead us to greatly overestimate or underestimate the danger.

A common error is to assume a linear response model without threshold when judging the effects of radioactivity or the toxicity of some substance. Presumably there is no threshold effect for cumulative poisons like heavy metal ions (mercury, lead), which are eliminated only very slowly if at all. But for virtually every organic substance (such as saccharin or cyclamates), the existence of a finite metabolic rate means that there must exist a finite threshold dose rate, below which the substance is decomposed or eliminated so rapidly that it has no ill effects. If this were not true, the human race could never have survived to the present time, in view of all the things we have been eating.

Indeed, every mouthful of food you and I have ever taken contained many billions of kinds of complex molecules whose structure and physiological effects have never been determined – and many millions of which would be toxic or fatal in large doses. We cannot doubt that we are daily ingesting thousands of substances that are far more dangerous than saccharin – but in amounts that are safe, because they are far below the various thresholds. But at the present time there is hardly any substance except common drugs, for which we actually know the threshold.

Therefore, the goal of inference in this field should be to estimate not only the slope of the response curve, but *far more importantly*, to decide whether there is evidence for a threshold; and if so, to estimate its magnitude (the "maximum safe dose"). For example, to tell us that a sugar substitute is dangerous in doses a thousand times greater than would ever be encountered in practice, is hardly an argument against using the substitute; indeed, the fact that it is necessary to go to kilodoses in order to detect any ill effects at all, is rather conclusive evidence, not of the danger, but of the *safety*, of a tested substance, and probability theory confirms this whenever it is allowed to do so (that is, whenever we use a model that is flexible enough to admit the possibility of a threshold). A similar overdose of sugar would be far more dangerous, leading not to barely detectable harmful effects, but to sure, immediate death by diabetic coma; yet nobody has proposed to ban the use of sugar in food.

Kilodose effects are irrelevant because we do not take kilodoses; in the case of a sugar substitute the important question is: *What are the threshold doses for toxicity of a sugar substitute and for sugar, compared to the normal doses*? If that of a sugar substitute is higher, then the rational conclusion would be that the substitute is actually safer than sugar, as a food ingredient. To analyze one's data in terms of a model which does not allow even the possibility of a threshold effect, is to prejudge the issue in a way that can lead to false conclusions from any amount of data.

We emphasize this in the Preface because false conclusions of just this kind are now not only causing major economic waste, but also creating unnecessary dangers to public health. Society has only finite resources to deal with such problems, so any effort expended on imaginary dangers means that real dangers are going unattended. Use of models which correctly represent the prior information that scientists have about the mechanism at work (such as metabolic rate, chemical reactivity) can prevent such folly in the future.

Style of Presentation: In part A, expounding principles and elementary applications, most Chapters start with several pages of verbal discussion of the nature of the problem. Here we try to explain the constructive ways of looking at it, and the logical pitfalls responsible for past errors. Only then do we turn to the mathematics, solving a few of the problems of the genre. In part B, expounding more advanced applications, we can concentrate more on the mathematical details.

The writer has learned from much experience that this primary emphasis on the logic of the problem, rather than the mathematics, is necessary in the early stages. For modern students, the mathematics is the easy part; once a problem has been reduced to a definite mathematical exercise, most students can solve it effortlessly and extend it endlessly, without further help from any book or teacher. It is in the conceptual matters (how to make the initial connection between the real–world problem and the abstract mathematics) that they are perplexed and unsure how to proceed.

Recent history demonstrates that anyone foolhardy enough to describe his own work as "rigorous" is headed for a fall. Therefore, we shall claim only that we do not knowingly give erroneous arguments. We are conscious also of writing for a large and varied audience, for most of whom clarity of meaning is more important than "rigor" in the narrow mathematical sense.

There are two more, even stronger reasons for placing our primary emphasis on logic and clarity. Firstly, no argument is stronger than the premises that go into it, and as Harold Jeffreys noted, those who lay the greatest stress on mathematical rigor are just the ones who, lacking a sure sense of the real world, tie their arguments to unrealistic premises and thus destroy their relevance. Jeffreys likened this to trying to strengthen a building by anchoring steel beams into plaster. An argument which makes it clear intuitively *why* a result is correct, is actually more trustworthy and more likely of a permanent place in science, than is one that makes a great overt show of mathematical rigor unaccompanied by understanding.

Secondly, we have to recognize that there are no really trustworthy standards of rigor in a mathematics that has embraced the theory of infinite sets. Morris Kline (1980, p. 351) came close to the Jeffreys simile: "Should one design a bridge using theory involving infinite sets or the axiom of choice? Might not the bridge collapse?" The only real rigor we have today is in the operations of elementary arithmetic on finite sets of finite integers, and our own bridge will be safest from collapse if we keep this in mind.

Of course, it is essential that we follow this "finite sets" policy whenever it matters for our results; but we do not propose to become fanatical about it. In particular, the arts of computation and approximation are on a different level than that of basic principle; and so once a result is derived from strict application of the rules, we allow ourselves to use any convenient analytical methods for evaluation or approximation (such as replacing a sum by an integral) without feeling obliged to show how to generate an uncountable set as the limit of a finite one.

But we impose on ourselves a far stricter adherence to the mathematical rules of probability theory than was ever exhibited in the "orthodox" statistical literature, in which authors repeatedly invoke the aforementioned intuitive *ad hoc* devices to do, arbitrarily and imperfectly, what the rules of probability theory as logic would have done for them uniquely and optimally. It is just this strict adherence that enables us to avoid the artificial paradoxes and contradictions of orthodox statistics, as described in Chapter 17.

Equally important, this policy often simplifies the computations in two ways: (A) The problem of determining the sampling distribution of a "statistic" is eliminated; the evidence of the data is displayed fully in the likelihood function, which can be written down immediately. (B) One can eliminate nuisance parameters at the beginning of a calculation, thus reducing the dimensionality of a search algorithm. This can mean orders of magnitude reduction in computation over what would be needed with a least squares or maximum likelihood algorithm. The Bayesian computer programs of Bretthorst (1988) demonstrate these advantages impressively, leading in some cases to major improvements in the ability to extract information from data, over previously used methods.

A scientist who has learned how to use probability theory directly as extended logic, has a great advantage in power and versatility over one who has learned only a collection of unrelated *ad–hoc* devices. As the complexity of our problems increases, so does this relative advantage. Therefore we think that in the future, workers in all the quantitative sciences will be obliged, as a matter of practical necessity, to use probability theory in the manner expounded here. This trend is already well under way in several fields, ranging from econometrics to astronomy to magnetic resonance spectroscopy.

Finally, some readers should be warned not to look for hidden subtleties of meaning which are not present. We shall, of course, explain and use all the standard technical jargon of probability and statistics – because that is our topic. But although our concern with the nature of logical inference leads us to discuss many of the same issues, our language differs greatly from the stilted jargon of logicians and philosophers. There are no linguistic tricks and there is no "meta–language" gobbledygook; only plain English. We think that this will convey our message clearly enough to anyone who seriously wants to understand it. In any event, we feel sure that no further clarity would be achieved by taking the first few steps down that infinite regress that starts with: "What do you mean by 'exists'?"

Acknowledgments: In addition to the inspiration received from the writings of Jeffreys, Cox, Pólya, and Shannon, I have profited by interaction with some 300 former students, who have diligently caught my errors and forced me to think more carefully about many issues. Also, over the years my thinking has been influenced by discussions with many colleagues; to list a few (in the reverse alphabetical order preferred by some): Arnold Zellner, George Uhlenbeck, John Tukey, William Sudderth, Stephen Stigler, John Skilling, Jimmie Savage, Carlos Rodriguez, Lincoln Moses, Elliott Montroll, Paul Meier, Dennis Lindley, David Lane, Mark Kac, Harold Jeffreys, Bruce Hill, Stephen Gull, Jack Good, Seymour Geisser, Willy Feller, Anthony Edwards, Morrie de Groot, Phil Dawid, Jerome Cornfield, John Parker Burg, David Blackwell, and George Barnard. While I have not agreed with all of the great variety of things they told me, it has all been taken into account in

one way or another in the following pages. Even when we ended in disagreement on some issue, I believe that our frank private discussions have enabled me to avoid misrepresenting their positions, while clarifying my own thinking; I thank them for their patience.

E. T. Jaynes
August 1993