PROBABILITY THEORY AS LOGIC[†]

E. T. JAYNES
Wayman Crow Professor of Physics
Washington University, St. Louis MO 63130, USA

Abstract: At the 1988 workshop we called attention to the "Mind Projection Fallacy" which is present in all fields that use probability. Here we give a more complete discussion showing why probabilities need not correspond to physical *causal* influences, or "propensities" affecting mass phenomena. Probability theory is far more useful if we recognize that probabilities express fundamentally *logical inferences* pertaining to individual cases. We note several examples of the difference this makes in real applications.

CONTENTS

[†] Presented at the Ninth Annual Workshop on Maximum Entropy and Bayesian Methods, Dartmouth College, New Hampshire, August 14, 1989. In the Proceedings Volume, *Maximum Entropy and Bayesian Methods*, Paul F. Fougere, Editor, Kluwer Academic Publishers, Dordrecht, Holland (1990). The present version was substantially revised, corrected, and extended 5/1/94.

INTRODUCTION

"Man is surely mad. He cannot make a worm; yet he makes Gods by the dozen." — Montaigne.

It seems that mankind has always been occupied with the problem of how to deal with ignorance. Primitive man, aware of his helplessness against the forces of Nature but totally ignorant of their causes, would try to compensate for his ignorance by inventing hypotheses about them. For educated people today, the idea of directing intelligences willfully and consciously controlling every detail of events seems vastly more complicated than the idea of a machine running; but to primitive man (and even to the uneducated today) the opposite is true. For one who has no comprehension of physical law, but is aware of his own consciousness and volition, the natural question to ask is not: "*What* is causing it?", but rather: "*Who* is causing it?"

The answer was to invent Gods with the same consciousness and volition as ourselves, but with the additional power of psychokinesis; one in control of the weather, one in control of the seas, and so on. This personification of Nature must have been going on for thousands of years before it started producing permanent written records, in ancient Egypt and Greece. It appears that the adult citizens of those times really believed very literally in all their local Gods.

This oldest of all devices for dealing with one's ignorance, is the first form of what we have called the "Mind Projection Fallacy". One asserts that the creations of his own imagination are real properties of Nature, and thus in effect projects his own thoughts out onto Nature. It is still rampant today, not only in fundamentalist religion, but in every field where probability theory is used.

Of course, we are not arguing against a scientist's practice of formulating hypotheses about what is happening in Nature. Indeed, we see it as the highest form of creativity – far transcending mere mathematical manipulative skill – to conceive the right hypothesis in any field, out of one's educated imagination. Copernicus, Newton, Faraday, Darwin, Mendel, Pasteur, Wegener, Einstein are our heroes for having done this.

The difference between an imaginative scientist on the one hand, and primitive man and religious fundamentalists on the other, is that the scientist clearly recognizes the creations of his imagination as *tentative working hypotheses* to be tested by observation; and he is prepared to test and reject a hundred different hypotheses in order to find the right one.

THE MIND PROJECTION FALLACY

The writer became fully aware of this fallacy only recently, after many years of being vaguely troubled by the kind of logic that is used in a dozen different fields. Eventually there came that sudden flash of understanding of what was causing this.

I first learned about Bose and Fermi statistics, as an undergraduate, by this argument: "You and I cannot distinguish between the particles: *therefore*, the particles behave differently than if we could." In some vague way, the logic of this bothered me. It seemed to be claiming for man the powers of psychokinesis formerly reserved for those Gods. But a few years later, as a graduate student in Berkeley, I heard J. R. Oppenheimer expound the same argument, in the obvious conviction that this was an expression of deep new wisdom, a major advance in thinking.

Oppy proceeded to give a dozen other "physical" arguments in support of quantum theory, using the same pattern of logic. I was, of course, too cowed by his authority to raise any open objection, and made strenuous efforts to understand such arguments as lines of rational thought. But I never succeeded, and so for me quantum theory has always seemed to have more the character of a religion than a science.

Then in studying probability theory, it was vaguely troubling to see reference to "gaussian random variables", or "stochastic processes", or "stationary time series", or "disorder", as if the property of being gaussian, random, stochastic, stationary, or disorderly is a real property, like the property of possessing mass or length, existing in Nature. Indeed, some seek to develop statistical tests to determine the presence of these properties in their data.

There was a short phase of studying philosophy, hoping that the answer to what was troubling me might be found there. But the reasoning of philosophers was far worse in the same vague way, and it was clear that they understood even less than I did about the meaning of quantum theory and probability theory (at least, I was familiar with the mathematics of the theories and could solve real problems by applying it). There were several more experiences like this in other fields.

A change started with the much appreciated opportunity to spend a year at St. John's College, Cambridge, in the quiet contemplation that is possible only when one is free of all other responsibilities. This started some new lines of thought which finally congealed a few years later. It was in late 1987 that the sudden flash of understanding came, and I saw that the things that had been troubling me vaguely for 40 years were all the *same* basic error and that this error occurs not only in science and philosophy; it is ubiquitous in every area where people try to think.

Why did it take so long to see this? We can reason clearly only in areas where we have an established symbolism for the concepts; but this error had no name. Even after sensing it intuitively, there was a struggle to find an appropriate name for this vaguely seen thing. Finally the term "Mind Projection Fallacy" seemed to be the only one that expressed the idea without calling up wrong connotations.

As soon as the error had a definite name and description, it was much easier to recognize. Once one has grasped the idea, one sees the Mind Projection Fallacy everywhere; what we have been taught as deep wisdom, is stripped of its pretensions and seen to be instead a foolish *non sequitur*. The error occurs in two complementary forms, which we might indicate thus:

(A) (My own imagination) $\longrightarrow$ (Real property of Nature)

(B) (My own ignorance) $\longrightarrow$ (Nature is indeterminate)

Form (B) arose out of quantum theory; instead of covering up our ignorance with fanciful assumptions about reality, one accepts that ignorance but attributes it to Nature. Thus in the Copenhagen interpretation of quantum theory, whatever is left undetermined in a pure state ψ is held to be unknown not only to us, but also to Nature herself. That is, one claims that ψ represents a physically real "propensity" to cause events in a statistical sense (a certain proportion of times on the average over many repetitions of an experiment) but denies the existence of physical causes for the individual events below the level of ψ. Its zealots accuse those who speculate about such causes of being "obsolete, mechanistic materialists", to borrow a favorite phrase.

Yet no experiment could possibly demonstrate that no cause exists; the most that one could ever conclude is that no cause was found. But if we ask, "Well, how hard did you *try* to find the cause?" we will be told: "I didn't try at all, because the theory assures us there is none." Then in what sense can one say that any experiments confirm such a theory? How can anyone feel confident that no causes exist at the submicroscopic level when experiments give no evidence for this and experimenters do not search for them? Clearly, such a claim is pure type (B) Mind Projection Fallacy.

It is evident that this pattern of thought is also present throughout orthodox statistics, whenever someone states, or implies, that his probabilities are real causative agents *en masse* for events that are not determined, individually, by anything. And we see that there can be no such thing as a statistical test for "absence of cause" or "randomness" or "disorder" for the same reason that there is no test for ugliness or foolishness; those qualities exist only in the eye of the observer. Now let us see one aspect of this in a specific example.

EXAMPLE: THE POISSON DISTRIBUTION

At our Cambridge meeting in 1988, the fallacy of supposing that conditional probabilities must express a real physical causation in Nature (but one which operates only statistically), was illustrated by the example of drawing two balls from an Urn, comparing forward inference which may express such a causation with backward inference which cannot. Now let us give less trivial (and more useful) calculations which illustrate that in order to conduct sound reasoning we are not only permitted, but *required*, to use conditional probabilities as logical inferences, in situations where physical causation could not be operative.

The elementary Poisson distribution sampling problem provides a very nice example. The Poisson distribution is usually derived as a limiting "low counting rate" approximation to the binomial distribution, but it is instructive to derive it by using probability theory as logic, directly from the statement of independence of different time intervals, using only the primitive product and sum rules. Thus define the prior information:

$I \equiv$ "There is a positive real number λ such that, given λ, the probability that an event A, or count, will occur in the time interval $(t, t+dt)$ is $p(A|\lambda I) = \lambda dt$. Furthermore, knowledge of λ makes any information Q about the occurrence or nonoccurrence of the event in any other time interval irrelevant to this probability: $p(A|\lambda QI) = p(A|\lambda I)$."

In orthodox statistics one would not want to say it this way, but instead would claim that λ is the sole causative agent present; the occurrence of the event in any other time interval exerts no *physical influence* on what happens in the interval dt. Our statement is very different.

Denote by $h(t)$ the probability there is no count in the time interval $(0, t)$. Now the proposition:

$$R \equiv \text{"No count in } (0, t+dt)\text{"} \tag{1}$$

is the conjunction of the two propositions:

$$R = [\text{"No count in } (0,t)\text{"}] \cdot [\text{"No count in } (t, t+dt)\text{"}] \tag{2}$$

and so, by the independence of different time intervals, the product rule gives:

$$h(t+dt) = h(t) \cdot [1 - \lambda dt] \tag{3}$$

or $\partial h/\partial t + \lambda h(t) = 0$. The solution, with the evident initial condition $h(0) = 1$, is

$$h(t) = e^{-\lambda t} \quad . \tag{4}$$

Now consider the probability, given λ and I, of the proposition

$B \equiv$ "In the interval $(0, t)$ there are exactly n counts, which happen at the times $(t_1, \cdots, t_n)$ within tolerances $(dt_1, \cdots, dt_n)$, where $(0 < t_1, \cdots < t_n < t)$."

This is the conjunction of $(2n+1)$ propositions:

$$B = [\text{no count in } (0, t_1)] \cdot (\text{count in } dt_1) \cdot [\text{no count in } (t_1, t_2)] \cdot (\text{count in } dt_2) \cdots$$
$$[\text{no count in } (t_{n-1}, t_n)] \cdot (\text{count in } dt_n) \cdot [\text{no count in } (t_n, t)].$$

so by the product rule and the independence of different time intervals, the probability of this is the product of all their separate probabilities:

$$p(B|\lambda I) = [e^{-\lambda t_1}] \cdot (\lambda dt_1) \cdot [e^{-\lambda(t_2 - t_1)}] \cdots [e^{-\lambda(t_n - t_{n-1})}] \cdot (\lambda dt_n) \cdot [e^{-\lambda(t - t_n)}]$$

or, writing the proposition B now more explicitly as $B =$ '$dt_1 \cdots dt_n$',

$$p(dt_1 \cdots dt_n|\lambda t I) = e^{-\lambda t}\, \lambda^n\, dt_1 \cdots dt_n\ , \qquad (0 < t_1, \cdots < t_n < t) \tag{5}$$

Then what is the probability, given λ, that in the interval $(0, t)$ there are exactly n counts, whatever the times? Since different choices of the count times represent mutually exclusive propositions, the continuous form of the sum rule applies:

$$p(n|\lambda t I) = \int_0^t dt_n \cdots \int_0^{t_3} dt_2 \int_0^{t_2} dt_1\ e^{-\lambda t}\ \lambda^n$$

or,

$$p(n|\lambda t I) = e^{-\lambda t}\ \frac{(\lambda t)^n}{n!} \tag{6}$$

the usual Poisson distribution. Without the time ordering in our definition of B, different choices of count times would not all be mutually exclusive events, so the sum rule would not apply in the above way.

As noted, conventional theory obtains this same formula from the premise that events in disjoint time intervals exert no physical influences on each other; the only causative agent operating is λ. Some authors have turned this around, and supposed that if we verify (6) in the frequency sense, that proves that the events were indeed causally independent!

This is an astonishing conclusion, when we note that one could design a hundred different mechanisms (or write a hundred different computer programs), which in various ways that are completely deterministic, generate the seemingly "random" data. That is, the time of the next event is completely determined by the times of the previous events by some complicated rule. Yet all of them could constrain the long–run frequencies to agree with (6) without showing any signs of correlations.

If an experimenter did not know what that complicated rule was, there is almost no chance that he could discover it merely by accumulation of more data. Then the Mind Projection Fallacy might lead him to claim that no rule exists; and we seem to be back to quantum theory. This is why "randomness" is a slippery, undefined, and unverifiable notion.

Now consider the next problem: let $0 < t_1 < t_2$ and let n_1 and n_2 be the numbers of counts in the time intervals $(0, t_1)$ and $(0, t_2)$. What is the forward conditional probability $p(n_2|n_1, \lambda, t_1, t_2, I)$? By the aforementioned logical independence, the probability that there are $(n_2 - n_1)$ counts in (t_1, t_2) still has the form (1) independent of n_1, so

$$p(n_2|n_1, \lambda, t_1, t_2, I) = e^{-\lambda(t_2 - t_1)}\ \frac{[\lambda(t_2 - t_1)]^{n_2 - n_1}}{(n_2 - n_1)!}\ , \quad t_1 < t_2, \quad n_1 \le n_2 \tag{7}$$

Then what is the joint probability for n_1 and n_2? By the product rule,

$$p(n_1 n_2|\lambda t_1 t_2 I) = p(n_1|\lambda t_1 I)\, p(n_2|n_1 \lambda t_1 t_2 I)$$

$$= \left[e^{-\lambda t_1} \frac{(\lambda t_1)^{n_1}}{n_1!}\right] \cdot \left[e^{-\lambda(t_2 - t_1)} \frac{[\lambda(t_2 - t_1)]^{n_2 - n_1}}{(n_2 - n_1)!}\right] \tag{8}$$

Now this can be rearranged into

$$\left[e^{-\lambda t_2} \frac{(\lambda t_2)^{n_2}}{n_2!}\right] \cdot \left[\binom{n_2}{n_1} \left(\frac{t_1}{t_2}\right)^{n_1} \left(1 - \frac{t_1}{t_2}\right)^{n_2 - n_1}\right] \tag{9}$$

and we recognize the first factor as the unconditional probability $p(n_2|\lambda t_2 I)$, so by the other way of writing the product rule,

$$p(n_1 n_1|\lambda t_1 t_2 I) = p(n_2|\lambda t_2 I)\, p(n_1|n_2 \lambda t_1 t_2 I) \tag{10}$$

the backward conditional distribution must be given by the binomial:

$$p(n_1|n_2\lambda t_1 t_2 I) = \binom{n_2}{n_1}\left(\frac{t_1}{t_2}\right)^{n_1}\left(1-\frac{t_1}{t_2}\right)^{n_2-n_1}, \qquad (0 \le n_1 \le n_2) \tag{11}$$

But this is totally different from $p(n_2|n_1\lambda t_1 t_2)$; it does not even contain λ!

When we reason forward from given n_1 to inferred n_2, knowledge of λ makes n_1 irrelevant for predicting the number of counts after t_1. In conventional "random variable" probability theory one might think that λ is always the sole relevant quantity because it is the sole physical causative agent; and therefore $p(n_1|n_2\lambda t_1 t_2) \equiv p(n_1|\lambda t_1)$. But our analysis shows that when we reason backward from n_2 to n_1, knowledge of λ does not make n_2 irrelevant; on the contrary, knowledge of n_2 makes λ irrelevant!

We could hardly make the point more strongly that *physical* dependence and *logical* dependence are very different things. Some may find this result so disconcerting that their first reaction is to doubt the correctness of (11). If you find yourself with such feelings, please consider: If you already knew the actual number n_2 of events in the long interval, how would you then use knowledge of λ to improve your estimate of n_1 beyond what is given by (11)?

The point is that knowledge of λ does not determine n_2; it gives us only probabilities for different values of n_2. But if we know the *actual value* of n_2 over an interval that includes $(0, t_1)$, common sense surely tells us that this takes precedence over anything that we could infer from λ. That is, possession of the datum n_2 makes the original sampling probabilities (those conditional only on λ) irrelevant to the question we are asking.

In the above we considered λ known in advance (*i.e.* specified in the statement of the problem). More realistically, we will not know λ exactly, and therefore information about either n_1 or n_2 will enable us to improve our knowledge of λ and take this into account to improve our estimates of other things. How will this change our results?

Consider the case that λ is unknown, but suppose for simplicity that it is known not to be varying with time. Then we are to replace (6) – (10) by extra integrations over λ. Thus in place of (6) we have

$$p(n|I) = \int p(n\lambda|I)d\lambda = \int p(n|\lambda I)\, p(\lambda|I)\, d\lambda \tag{6a}$$

where $p(n|\lambda I)$ is given by (6), and $p(\lambda|I)$ is the prior probability density function (*pdf*) for λ. In a similar way, (7) is replaced by a probability mixture of the original distributions:

$$p(n_2|n_1, t_1, t_2, I) = \int p(n_2|n_1, \lambda, t_1, t_2, I)\, p(\lambda|I) d\lambda\,, \qquad t_1 < t_2,\ \ n_1 \le n_2 \tag{7a}$$

and the joint probability for n_1 and n_2 becomes

$$p(n_1 n_2|t_1 t_2 I) = \int p(n_1 n_2|\lambda t_1 t_2 I)\, p(\lambda|I)\, d\lambda \tag{8a}$$

where the integrand is given by (8); again it can be rearranged as in (9), yielding

$$p(n_1 n_2|t_1 t_2 I) \int \left[e^{-\lambda t_2}\frac{(\lambda t_2)^{n_2}}{n_2!}\right]\ \left[\binom{n_2}{n_1}\left(\frac{t_1}{t_2}\right)^{n_1}\left(1-\frac{t_1}{t_2}\right)^{n_2-n_1}\right] p(\lambda|I)\, d\lambda$$

but now we recognize in this the factor

$$p(n_2|t_1t_2I) = \int p(n_2|\lambda t_1 t_2 I)\, p(\lambda|I)\, d\lambda \tag{9a}$$

and so from the product rule in the form $p(n_1 n_2|t_1t_2I) = p(n_2|t_1t_2I)\; p(n_1|n_2t_1t_2I)$ we conclude that

$$p(n_1|n_2t_1t_2I) = \binom{n_2}{n_1}\left(\frac{t_1}{t_2}\right)^{n_1}\left(1-\frac{t_1}{t_2}\right)^{n_2-n_1}, \qquad (0 \le n_1 \le n_2) \tag{11a}$$

in agreement with (11); this result is the same whether λ is known or unknown. Since this derivation allowed full opportunity for updated knowledge of λ to be taken into account, (11a) is a Bayesian predictive distribution.

In reasoning from n_1 to n_2, the difference between (7) and (7a) represents the penalty we pay for not knowing λ exactly; but in reasoning from n_2 to n_1 there is no penalty. Indeed, if the probability (11) is independent of λ when λ is known, it is hard to see how it could matter if λ was unknown; and probability theory so indicates. Possession of the datum n_2 makes the original sampling probabilities – whether λ is known or unknown – irrelevant to the question we are asking.

DISCUSSION

The phenomena we have just demonstrated are true much more generally. Conventional sampling probabilities like $p(n|\lambda t)$ are relevant only for "pre-data" considerations; making predictions before we have seen the data. But as soon as we start accumulating data, our state of knowledge is different. This new information necessarily modifies our probabilities for the remaining data in a way that is incomprehensible to one who tries to interpret probabilities as expressing physical causation or long–run relative frequencies; but as (11) and (11a) illustrate, this updating appears automatically when we use probability theory as logic.

For example, what is the probability that in 10 binomial trials we shall find 8 or more successes? The binomial sampling distribution might assign to this event a probability of 0.46. But if the first 6 trials yield only 3 successes, then we know with certainty that we shall *not* get 8 or more successes in those 10 trials; the sampling probability 0.46 becomes irrelevant to the question we are asking.

How would a conventional probabilist respond to this example? He can hardly deny our conclusion, but he will get out of it by saying that conventional probability theory does not refer to the individual case as we were trying to do; it makes statements only about long–run relative frequencies, and we agree.

But then we observe that probability theory as logic *does* apply to the individual case, and it is just that individual case that concerns us in virtually all real problems of scientific inference (*i.e.*, reasoning as best we can when our information is incomplete). The binomial distribution (11) will yield a more reliable estimate of n_1 than will the Poisson distribution (6) *in each individual case* because it contains cogent information, pertaining to that individual case, that is not in (6).

Orthodox probabilists, who use only sampling probability distributions and do not associate them with the individual case at all, are obliged to judge any estimation method by its performance "in the long run"; *i.e.* by the sampling distribution of the estimates when the procedure is repeated many times. That is of no particular concern to a Bayesian, for the same reason that a person with a ten–digit hand calculator has no need for a slide rule. The real job before us is to make the best estimates possible from the information we have *in each individual case*; and since Bayesians already have the solution to that problem, we have no need to discuss a lesser problem.

To see that long–run performance is indeed a lesser problem, note that even if we had found a procedure whose long–run performance is proved to be as good as can be obtained (for example, which achieves the minimum possible mean–square error), that would not imply that this procedure is best – or even tolerably good – in any particular individual case. One can trade off better performance for one class of samples against worse performance for another in a way that has no effect on long–run performance, but has a very large effect on performance in the individual case.

We do, of course, want to know how accurate we can expect our estimates to be; but the proper criterion of this is not the sampling distribution, but the width of the Bayesian posterior probability distribution for the parameter. This gives an indication of the accuracy of our estimate *in each individual case*, not merely a smeared–out average over all cases. In this sense also, the sampling distribution is the answer to a lesser problem, and the sampling distribution criterion of performance is not the one an informed scientist really wants.

The relevant question for a scientist is not: "How accurately would the estimate agree with the true value of the parameter in the long run over all possible data sets?" but rather: "How accurately does the one data set that I actually have determine the true value of the parameter?" This is the question that Bayes' theorem answers. When no sufficient statistic exists (or if one uses an estimator which is not a sufficient statistic, even though one does exist) the answers to these questions can be very different, and the sampling distribution criterion can be misleading.

This was perceived already by R. A. Fisher in the 1930's. He noted that different data sets, even though they may lead to the same estimate, may still justify very different claims of accuracy because they have different spreads, and he sought to correct this by his conditioning on "ancillary statistics". But ancillary statistics do not always exist, and when they do, as noted by Bretthorst (1988), this procedure is mathematically equivalent to applying Bayes' theorem.

Indeed Bayes' theorem generates a log–likelihood that is spread over a range corresponding to that of the data, whether or not ancillary statistics exist. For example, let θ be a location parameter: $p(x_i|\theta) = f(x_i - \theta)$. Then $\log L(\theta) = \Sigma_i \log f(x_i - \theta)$. The point is that the width of the likelihood function is determined, not merely by the width of the sampling function $f(x - \theta)$, but even more by the spread in values of the *actual data*. Thus it automatically compensates when the data have spread greater or less than expected, so the claimed accuracy is always that indicated by the data. This is one of its built–in safety features; it protects us against being misled about the accuracy of our estimates.

If we choose estimates by sampling distribution criteria, the conclusions we draw will depend, not just on the data that we actually have, but on what other data sets one thinks might have been observed, but were not. Thus an estimation procedure that works well for the data set that we have observed can be rejected on the grounds that it would have worked poorly for some other data set that we have not observed! We return to this point in "Optional Stopping" below, and see the can of worms it opens up.

But if anyone insists on seeing it, the Bayesian can of course calculate the sampling distribution for his estimates. Needless to say, Bayesian procedures look very good by sampling theory criteria, even though they were not derived with such criteria in mind. In fact, analysis of many cases has shown that the Bayesian point and interval estimates based on noninformative priors (say, the mean $\pm$ standard deviation of the posterior distribution) are almost always the best available by sampling theory criteria. With informative priors, Bayesian methods advance into an area where sampling theory cannot follow at all.

For example, the frequency distribution of errors in estimating n_1 that result from using the Bayesian predictive distribution (11a) will be better (more accurate in the long run) than the distribution that one gets from using the direct sampling distribution (6). To a Bayesian, this is obvious without any calculation; for if a method is better in each individual case, how could it fail to be better also in the long run?

Orthodox probabilists would not accept that argument, but they would be convinced by comparing the two sampling distributions of the estimates, either analytically or experimentally. A simple hand–waving argument leads us to predict that the mean–square error with (11a) will be less than that with (6) by a factor $[1-(t_1/t_2)]$. Although we could demonstrate this by an explicit sampling theory calculation, it would be more interesting to conduct "Monte Carlo" computer experiments to check our claim.

Of course, probability theory as logic need not be applied only to the individual case. It can be applied equally well to prediction of long–run relative frequencies, if that happens to be the thing of interest. Indeed, it can sometimes make better predictions, because by using the notion of probability of an hypothesis it has the means for taking into account relevant information that "random variable" theory cannot use.

The philosophical difference between conventional probability theory and probability theory as logic is that the former allows only sampling distributions, interprets them as physically real frequencies of "random variables", and rejects the notion of probability of an hypothesis as being meaningless. We take just the opposite position: that the probability of an hypothesis is the fundamental, necessary ingredient in all inference, and the notion of "randomness" is a red herring, at best irrelevant.

But although there is a very great philosophical difference, where is the functional difference? As illustrated above, by "probability theory as logic" we mean nothing more than applying the standard product and sum rules of probability theory to whatever propositions are of interest in our problem. The first reaction of some will be: "What difference can this make? You are still using the same old equations!" To see why it makes a difference, consider some case histories from Statistical Mechanics and Artificial Intelligence.

STATISTICAL MECHANICS

Mark Kac (1956) considered it a major unsolved problem to clarify how probability considerations can be introduced into physics. He confessed that he could not understand how one can justify the use of probability theory as Boltzmann used it, in writing down a simultaneous probability distribution $f(x, v; t)\, d^3x\, d^3v$ over position and velocity of a particle, because:

> "In every probabilistic model in physics and in all other sciences there must be some lack of specification over which you can average. - - That's the whole problem as to how probability can be introduced in kinetic theories of mechanics. - - - I am unable to find a probabilistic model which will lead to the full Boltzmann equation. I will show you how one can very easily be led to the equation in velocity space, however. - - - Once we have spatial homogeneity, then we have a lack of specification of position. And consequently we have wide freedom to average over all possible positions. If you don't have spatial homogeneity, then the problem becomes over–determined. There's absolutely no room, or at least I can't find any room, to introduce a stochastic element. I don't know what's random anymore, and so I cannot find a stochastic model which will lead to the full Boltzmann equation."

Mark Kac was a fine mathematician, but he had this mental hangup which prevented him from comprehending the notion of a probability referring to an individual case, rather than an "ensemble" of cases. So he was reduced to inventing clever mathematical models, instead of realistic physical situations, for his probability analyses. In a very idealized model he found that he could get a Boltzmann–like equation only if the n–particle probability distribution factored into a product of single–particle distributions. This led him to say of the Boltzmann equation:

> "- - - it is philosophically rather peculiar. Because if you believe in it you must ask yourself why nature prepares for you at time zero such a strange factorized distribution. Because otherwise you can't get Boltzmann's equation."

We see here the Mind Projection Fallacy, in the form of a belief that his n–particle probability distributions were real things existing in Nature. The answer which I gave Mark Kac at the time was: "Nature does not prepare distributions, factorized or otherwise; she prepares *states*." But his thinking was so far removed from this viewpoint that he thought I was joking.

Today we could explain the point a little better: "The probability distributions in phase space used by Maxwell, Boltzmann, and Gibbs are not realities existing in Nature; they are descriptions of incomplete human information about Nature. They yield the best predictions possible from the information contained in them. Probability distributions are not "right" or "wrong" in a factual sense; rather, some distributions make better predictions than others because they contain more relevant information. With a factorized distribution, getting knowledge of the position of one particle would tell us nothing about the position of any other. But at soon as the particles interact, knowing the position of one *does* tell us something about where the others are. Therefore the probability distributions which lead to the best physical predictions for interacting particles are not factorized."

But we think that Kac's mathematical conclusion is quite correct; the Boltzmann equation does, in effect, suppose factorization. But then it follows that it cannot take full account of the effect of interactions. Indeed, as soon as the particles interact, a factorized n–particle probability distribution cannot predict correctly either the equation of state or the hydrodynamic equations of motion. One can do much better by using the nonfactorized distribution of Gibbs, which contains more relevant information than does the Boltzmann distribution, in just the same sense that (11) contains more relevant information than does (6). Mark Kac had this important fact in his grasp but could not see it because of his conceptual hangup over the notion of probability; and so he never appreciated the superiority of Gibbs' methods, and continued trying to justify Boltzmann's methods.

We could give other recent case histories of workers (Feller, Uhlenbeck, Montroll) who were highly competent mathematically, but were conceptually such captives of the Mind Projection Fallacy that it prevented them from seeing important results that were already present in their equations. For others, this fallacy prevents them from finding any useful results at all. For example, Pool (1989) quotes a current worker in statistical mechanics stating as one of its long–standing problems:

> "Where does the randomness necessary for statistical behavior come from if the universe is at heart an orderly, deterministic place?"

Statements like this are spread throughout the literature of quantum theory and statistical mechanics. People who believe that probabilities are physically real, are thereby led to doubt the reality of mechanical causes; eventually they come to doubt the reality of physical objects like atoms.

Once we have learned how to use probability theory as logic, we are free of this mental hangup and able at least to perceive, if not always solve, the real problems of science. Most of those long–standing problems of statistical mechanics are seen as non–problems. We do not seek to explain "statistical behavior" because there is no such thing; what we see in Nature is *physical* behavior, which does not conflict in any way with deterministic physical law. Quite the contrary, probability theory as logic easily explains what we see, as a *consequence* of deterministic physical law.

We are not puzzled by "irreversibility" because (one of those important results which has been in our equations for over a Century, but is still invisible to some), given the present macrostate, the overwhelming majority of *all possible* microstates lead, via just those evil, deterministic mechanical equations of motion, to the *same* macroscopic behavior; just the reproducible behavior that we observe in the laboratory. So what else is there to explain? There would be a major mystery to be explained if this behavior were *not* observed.

The Maximum Entropy formulation makes these things obvious from the start, because it sees Statistical Mechanics not as a physical theory of "random behavior", but as a process of inference: predicting macroscopic behavior as best we can from our incomplete information about microstates. In this endeavor, as in any other problem of inference, we never ask, "Which quantities are random?" The relevant question is: "Which quantities are known, and which are unknown?" Indeed, it appears to us that whenever we get down to a specific calculation, all of us are obliged to use the term "random" merely as a synonym for "unknown"

ARTIFICIAL INTELLIGENCE

This field provides examples differing in detail, but not in the basic situation. Recently, its stagnation has been noted by many, leading to the appearance of articles of the genre: "*Whatever happened to AI?*" We can tell you what has happened by noting two recent references.

Peter Cheeseman (1988) in an eight–page article, tried to point out the need for Bayesian inference in AI, only to be buried under an avalanche of criticism (a total of 62 pages by 26 authors), which prompted a 14–page reply by Cheeseman. To elicit such a response must mean that Cheeseman's needle struck a rather sensitive nerve. Most of the critics simply refused to take note of his message (which was that Bayesian methods solve some currently important AI problems) and attacked his work on other grounds. Obviously, we cannot go into all the counter–arguments here, but we can indicate their general flavor.

The first critic objected to Bayesian methods on the grounds that they do not tell us how to create hypotheses (although neither do any other methods). This is like criticizing a computer because it will not push its own buttons; of course, it is up to us to tell Bayesian theory *which* problem we want it to solve. Would anybody want it otherwise?

The second critic complained that Bayesian inference "seems to be ruled out as a candidate for representing commonsense reasoning" on the grounds that people often reason differently. Indeed they do, particularly in AI. As we have been pointing out for many years, people who lack Bayesian training commit all kinds of errors of reasoning, which Bayesian methods would have corrected.

The third critic wrote so confusingly that I have no idea what he was trying to say. The fourth was concerned with Cheeseman's failure to recite the long history of the subject. The fifth and sixth rejected Cox's theorems on the grounds that he assumed differentiability, evidently unaware that Aczél (1966) removed that assumption long ago. Another critic resorted to name–calling, accusing Cheeseman of being a "necessarian"; and even worse, a *physicist*!!

[Indeed, one with some training in physics is in a good position to perceive the logical situation here, because we are familiar with it in other contexts. The present context is that we learn about real physical phenomena via probability theory, although probability is not itself a physical phenomenon. We also learn about the size and shape of objects via light, although light does not itself possess the properties of size and shape.]

And so it went on and on, critics calling up every conceivable subterfuge in order to avoid having to recognize what Bayesian methods *actually do* in real problems. It was like the rantings of a desperately ill patient who refuses to take the one medicine that could save him, and accuses the doctors of trying to poison him.

Our second example is explicit enough so that we can indicate one thing that Bayesian methods can do for AI. Dan Shafer (1989) tries to explain what is called in AI a "certainty factor" or "confidence factor". He states that (on a scale of 0 to 100) this "expresses the degree of confidence the user or the system has in a response or a conclusion" and warns us that this is something very different from a probability. In his words [italics mine]:

> "Probability—which predicts or describes the likelihood that *in a given group of items* any single item will have a particular attribute—does not enter into the issue."

Again, the mental hangup which cannot comprehend the notion of probability applied to an individual case.

Next he considers two propositions: A $\equiv$ "The patient's temperature is $>$ 101." and B $\equiv$ "The patient has been resting for an hour." Then we have the technical problem: suppose we have the confidence factors $c(A), c(B)$ for propositions A and B; what is the confidence factor for their conjunction AB = "The patient has a fever"? He notes "three popular methods" for calculating this: (1) the minimum; (2) the product; (3) the average. But then he notes that we now have a computer program named GURU, which is much superior because it offers the user his choice of not just three, but *seven* different rules for calculating this confidence factor! This certainly reveals the poverty of Bayesian analysis, which can offer only one solution to this problem.

A "confidence factor" is a very explicit attempt to represent degrees of plausibility by real numbers; and so the user of it is automatically at the mercy of Cox's theorems. Cox's first functional equation, expressing the associativity of Boolean algebra, shows that for the conjunction of propositions, any AI algorithm that is not mathematically equivalent to the product rule of probability theory, will contain demonstrable inconsistencies when we try to apply it to more than two propositions.

But these inconsistencies never appear in Shafer's discussion because it never reaches even the first level of comprehension, where one sees that the problem requires the notion of a conditional confidence factor $c(A|B)$. That is, knowing that a patient is not resting but exercising vigorously ought to increase one's "degree of confidence" that he has an elevated temperature, *etc.* The confidence factor for the conjunction AB cannot be assessed rationally if one fails to take this correlation into account. A potential inconsistency for three propositions hardly matters if we have not yet achieved consistency for two.

If we can judge from Shafer's exposition, the AI theory of confidence factors is stumbling about in a condition more primitive than the probability theory of Cardano, 400 years ago. Yet the problems they are trying to solve are just the ones that *were* solved long ago in the Bayesian literature, as Cheeseman tried to point out.

One disadvantage of having a little intelligence is that one can invent myths out of his own imagination, and come to believe them. Wild animals, lacking imagination, almost never do disastrously stupid things out of false perceptions of the world about them. But humans create artificial disasters for themselves when their ideology makes them unable to perceive where their own self–interest lies. We predict that AI will continue to stumble about, producing results that a Bayesian considers always trivial, usually quantitatively wrong, and often qualitatively wrong, until it recogizes that to dissociate itself from Bayesian probability theory was a disastrous error.

Almost everything we have noted here applies as well to the field of fuzzy sets; but it would be repetitious.

OPTIONAL STOPPING

A different kind of comparison appears in the issue of optional stopping, which has been a point of controversy between Bayesians and Orthodoxians for 30 years. This is another case where the Mind Projection Fallacy is doing serious damage, leading researchers to erroneous conclusions and wasted effort in the area of medical testing. The issue is: can an overzealous experimenter produce evidence supporting a foregone false conclusion, by deciding when to stop taking data?

Here orthodox theory as expounded by Armitage (1975) holds that the conclusions we should draw from an experiment depend not only on the experimental procedure and the resulting data, but also on the private thoughts that went through the experimenter's mind when he took the data! How can this be?

To see how, consider again binomial sampling, observing r successes in n trials. Two medical researchers use the same treatment independently, in different hospitals. Neither would stoop to falsifying the data, but one had decided beforehand that because of finite resources he would stop after treating $n = 100$ patients, however many cures were observed by then. The other had staked his reputation on the efficacy of the treatment, and decided that he would not stop until he had data indicating a rate of cures definitely greater than 60%, however many patients that might require. But in fact, both stopped with exactly the same data: $n = 100$, $r = 70$. Should we then draw different conclusions from their experiments?

One who thinks that the important question is: "Which quantities are random?" is then in this situation. For the first researcher, n was a fixed constant, r a random variable with a certain sampling distribution. For the second researcher, r/n was a fixed constant (approximately), and n was the random variable, with a very different sampling distribution. Orthodox practice will then analyze the two experiments in different ways, and will in general draw different conclusions about the efficacy of the treatment from them.

This would really cause trouble in a high–energy physics laboratory, where a dozen researchers may collaborate on carrying out a big experiment. Suppose that by mutual consent they agree to stop at a certain point; but they had a dozen different private reasons for doing so. According to the principles expounded by Armitage, one ought then to analyze the data in a dozen different ways and publish a dozen different conclusions, from what is in fact only one data set from one experiment!

Bayesian inference cannot get us into this absurd situation, because it perceives automatically what common sense demands; that what is relevant for this inference is not the relative probabilities of imaginary data sets which were not observed, but the relative likelihoods of different parameter values, based on the one real data set which *was* observed; and this is the same for all the experimenters.

Actually, as Jimmie Savage (1962) explained long ago, we need not worry about being misled by zealots because, contrary to what many assume, it is for all practical purposes impossible to sample deliberately to a foregone conclusion that is appreciably false. From the above data, most statisticians would estimate the true cure rate to be about $f \pm \sqrt{f(1-f)/n}$, or 70% $\pm$ 5% at one standard deviation, where $f = r/n$ is the observed frequency of cures. If the true incidence of cures in the whole population is only 50%, then the probability that a zealot can ever produce data like this (*i.e.* data which lead to an estimated interval that strongly excludes the true value) is extremely small in an honestly conducted experiment, even if he samples for billions of years (see the 'iterated logarithm' calculation in Feller's textbook).

Thus to produce data strongly supporting a false conclusion it is not enough merely to be zealous; one must be actively dishonest. Today it is not the zealots, but the medical testers who follow Armitage, who mislead themselves and others.

Orthodox writers love to charge Bayesians with "subjectivity" because we use probability as a way of describing our information. But we have seen a few examples of their idea of objectivity and some of its consequences, including inability to take prior information into account, inability to get rid of nuisance parameters, using inefficient criteria of performance, and inability to see important facts. Strangest of all, orthodox teaching can lead one to draw conclusions that depend on whether an experimenter subjectively imagined data sets which were not observed! A person who does that is in no position to charge anybody with "subjectivity".

Invariably, those who attack Bayesian methods only reveal their ignorance of what Bayesian methods are. The blame for this lies with those educators who continue to teach only orthodox methods and deprecate Bayesian methods without even fully defining them, much less examining their performance. The fact is that Bayesian methods of inference easily solve technical problems on which orthodox methods break down, and they protect us automatically against the absurd errors in orthodox methods. Thus they achieve scientific "objectivity" in the true sense of that word.

RECAPITULATION

In our simplest everyday inferences, in or out of science, it has always been clear that two events may be physically independent without being logically independent; or put differently, they may be logically dependent without being physically dependent. From the sound of raindrops striking my window pane, I infer the likely existence of clouds overhead, $p(\text{clouds}|\text{sound}) \simeq 1$, although the sound of raindrops is not a physical causative agent producing clouds. From the unearthing of bones in Wyoming we infer the existence of dinosaurs long ago: $p(\text{dinosaurs}|\text{bones}) \simeq 1$, although the digging of the bones is not the physical cause of the dinosaurs.

Yet conventional probability theory cannot account for such simple inferences, which we all make constantly and which are obviously justified. As noted, it rationalizes this failure by claiming that probability theory expresses partial physical causation and does not apply to the individual case.

But if we are to be denied the use of probability theory not only for problems of reasoning about the individual case; but also for problems where the cogent information does not happen to be about a physical cause or a frequency, we shall be obliged to invent arbitrary *ad hockeries* for dealing with virtually all real problems of inference; as indeed the orthodox school of thought has done.

Therefore, if it should turn out that probability theory used as logic *is*, after all, the unique, consistent tool for dealing with such problems, a viewpoint which denies this applicability on ideological grounds would represent a disastrous error of judgment, which deprives probability theory of virtually all its real value – and even worse, deprives science of the proper means to deal with its problems.

As an analogy, small children start counting on their fingers and toes. Then suppose that the teaching of arithmetic had fallen under control of an Establishment ideology which proclaims that the rules of elementary arithmetic apply *only* to fingers and toes; and then invents a different *ad hoc* rule for counting apples, still another for counting dollars, and so on. Imagine the effect this would have on our civilization.

Our position is that this is exactly what *has* happened in probability theory; when we start thinking about probability we, like small children, tend to think in terms of concrete things like frequencies. But we have become victims of an Establishment ideology which proclaims that the concept of probability applies *only* to frequencies! The effects of this are visible all about us. In theoretical physics we see the stagnation of quantum theory, an astonishing number of physicists having left it for other fields such as biophysics, and the wheel–spinning concentration on non–problems in statistical mechanics. In experimental science we see the absurdity of orthodox methods of data analysis. In Artificial Intelligence we see the consequences of militant refusal to adopt the only methods that can deal with their unsolved problems.

Fortunately, children quickly reach a level of maturity where they can perceive the number 13 as an abstract notion in its own right, not standing necessarily for ten fingers and three toes. It is high time that science reached the level of maturity where we can grasp the idea of probability of an hypothesis as an abstract notion in its own right, necessary for organizing our reasoning in a consistent way. Of course, just as the rules of arithmetic remain valid when applied to fingers and toes, the rules of probability theory remain valid when applied to calculation of frequencies.

Note that probability theory as logic is more general than just Bayesian inference; it automatically includes calculation of sampling distributions, as in our derivation of (6), and Maximum Entropy calculations, in the situations where they are appropriate. Bayesian analysis requires that we have a model in addition to the data. If we have only an hypothesis space but no model, then MAXENT is the best we can do without more information.

It may appear that in our recent concentration on Bayesian methods we have abandoned MAXENT. Not at all; it is an accomplished fact and we are using it constantly for its original purpose: to assign our priors. It is just for that reason that, for some of us, the focus of attention has now shifted to the Bayesian sequel. The exciting "new" fact (although it would not have surprised Harold Jeffreys in the least fifty years ago) is the flexibility of Bayesian analysis. As Bretthorst (1988) demonstrates by many specific examples, it can accommodate itself easily to all kinds of complicating circumstances that would bring orthodox methods to a standstill.

Scientists, engineers, economists and statisticians who are ignorant of Bayesian methods are handicapped in the performance of their work. In physics and astronomy the greatest experts in instrumentation may conduct multi–million–dollar data gathering operations – and then present their final conclusions in the form of confidence intervals which ignore not only some highly cogent prior information, but usually part of the information in the data. It is as if the best chefs in Paris had spared no effort or expense to prepare the finest food a restaurant can offer – and then spilled half of it down the drain and served the rest on paper plates.

APPENDIX: A BASIC BLUNDER

Finally, we must comment on a curious article entitled "The Basic Bayesian Blunder", by the philosopher H. E. Kyburg (1987), which amused us at this meeting. He formulates a problem of little interest except that the mathematical issue reduces to this: It is our basic blundering Bayesian belief that a weighted average of real numbers may take on all those, and only those, values in the range spanned by them. Thus any number in (.887, .917) can be written as a weighted average of (.887, .907, .917).

Since the average 0.900 specified by Kyburg lies in that interval, Bayesians do indeed, just as he charges, believe that we can assign prior probabilities (α, β, γ) leading to that average. He devotes several pages to arguing, by reasoning that we are completely unable to follow, that there is no solution. For answer it should suffice to exhibit an infinity of solutions. The system of equations to be solved is

$$\alpha + \beta + \gamma = 1$$

$$.887\alpha + .907\beta + .917\gamma = .900$$

and one verifies at once that the exact general solution of this system is

$$(\alpha, \beta, \gamma) \;=\; \left(\frac{65 - r}{140}\;,\;\frac{43 + 3r}{140}\;,\;\frac{32 - 2r}{140}\right)$$

where r is arbitrary. To meet the further requirement of nonnegativity $(\alpha, \beta, \gamma) \geq 0$, we see by inspection that r is confined by $(-43 \leq 3r \leq 48)$. There is a continuum of prior probability assignments which meet all the specified conditions, and which therefore enable a Bayesian to incorporate additional information.

REFERENCES

P. Armitage (1975), *Sequential Medical Trials*, Thomas, Springfield, Illinois. Second edition: Blackwell, Oxford.

G. Larry Bretthorst (1988), *Bayesian Spectrum Analysis and Parameter Estimation*, Springer Lecture Notes in Statistics, Vol. 48.

Peter Cheeseman (1988), "An inquiry into computer understanding", Comput. Intell. **4**, 58-66. See also the following 76 pages of discussion.

H. E. Kyburg (1987), "The Basic Bayesian Blunder", in *Foundations of Statistical Inference*, Vol II, I. B. MacNeill & G. J. Umphrey, Editors, Reidel Publishing Company, Holland.

Mark Kac (1956), *Some Stochastic Problems in Physics and Mathematics*; Colloquium Lectures in Pure and Applied Science #2, Socony-Mobil Oil Company, Dallas, Texas.

R. Pool (1989), "Chaos Theory: How Big an Advance?", Science, **245**, 26–28.

L. J. Savage (1962) *The Foundations of Statistical Inference*, G. A. Barnard & D. R. Cox, Editors, Methuen & Co., Ltd., London

Dan Shafer (1989), "Ask the Expert", PC AI Magazine, May/June; p. 70.