BAYESIAN METHODS: GENERAL BACKGROUND
An Introductory Tutorial

E.T. Jaynes
St. John's College and Cavendish Laboratory
Cambridge CB2 1TP
England

We note the main points of history, as a framework on which to hang many background remarks concerning the nature and motivation of Bayesian/Maximum Entropy methods. Experience has shown that these are needed in order to understand recent work and problems. A more complete account of the history, with many more details and references, is given in Jaynes (1978).

The following discussion is essentially nontechnical; the aim is only to convey a little introductory "feel" for our outlook, purpose, and terminology, and to alert newcomers to common pitfalls of misunderstanding.

INDEX

HERODOTUS

The necessity of reasoning as best we can in situations where our information is incomplete is faced by all of us, every waking hour of our lives. We must decide what to do next, even though we cannot be certain what the consequences will be. Should I wear a raincoat today, eat that egg, cross that street, talk to that stranger, tote that bale, buy that book?

Long before studying mathematics we have all learned, necessarily, how to deal with such problems intuitively, by a kind of plausible reasoning where we lack the information needed to do the formal deductive reasoning of the logic textbooks. In the real world, some kind of extension of formal logic is needed.

And, at least at the intuitive level, we have become rather good at this extended logic, and rather systematic. Before deciding what to do, our intuition organizes the preliminary reasoning into stages: (a) try to foresee all the possibilities that might arise; (b) judge how likely each is, based on everything you can see and all your past experience; (c) in the light of this, judge what the probable consequences of various actions would be; (d) now make your decision.

From the earliest times this process of plausible reasoning preceding decisions has been recognized. Herodotus, in about 500 BC, discusses the policy decisions of the Persian kings. He notes that a decision was wise, even though it led to disastrous consequences, if the evidence at hand indicated it as the best one to make; and that a decision was foolish, even though it led to the happiest possible consequences, if it was unreasonable to expect those consequences.

So this kind of reasoning has been around for a long time, and has been well understood for a long time. Furthermore, it is so well organized in our minds in qualitative form that it seems obvious that: (a) the above stages of reasoning can be reproduced in a quantitative form by a mathematical model; (b) such an extended logic would be very useful in such areas as science, engineering, and economics, where we are also obliged constantly to reason as best we can in spite of incomplete information, but the number of possibilities and amount of data are far too great for intuition to keep track of.

BERNOULLI

A serious, and to this day still useful, attempt at a mathematical representation was made by James Bernoulli (1713), who called his work "Ars Conjectandi", or "The Art of Conjecture", a name that might well be revived today because it expresses so honestly and accurately what we are doing. But even though it is only conjecture, there are still wise and foolish ways, consistent and inconsistent ways, of doing it. Our extended logic should be, in a sense that was made precise only much later, an optimal or "educated" system of conjecture.

First one must invent some mathematical way of expressing a state of incomplete knowledge, or information. Bernoulli did this by enumerating a set of basic "equally possible" cases which we may denote by $(x_1, x_2 \ldots x_N)$, and which we may call, loosely, either events or propositions. This defines our "field of discourse" or "hypothesis space" $H_\emptyset$. If we are concerned with two tosses of a die, $N = 6^2 = 36$.

Then one introduces some proposition of interest A, defined as being true on some specified subset H(A) of M points of $H_\emptyset$, false on the others. M, the "number of ways" in which A could be true, is called the multiplicity of A, and the probability of A is defined as the proportion p(A) = M/N.

The rules of reasoning consist of finding the probabilities p(A), p(B), etc., of different propositions by counting the number of ways they can be true. For example, the probability that both A and B are true is the proportion of HØ on which both are true. More interesting, if we learn that A is true, our hypothesis space contracts to H(A) and the probability of B is changed to the proportion of H(A) on which B is true. If we then learn that B is false, our hypothesis space may contract further, changing the probability of some other proposition C, and so on.

Such elementary rules have an obvious correspondence with common sense, and they are powerful enough to be applied usefully, not only in the game of "Twenty Questions", but in some quite substantial problems of reasoning, requiring nontrivial combinatorial calculations. But as Bernoulli recognized, they do not seem applicable to all problems; for while we may feel that we know the appropriate HØ for dice tossing, in other problems we often fail to see how to define any set HØ of elementary "equally possible" cases. As Bernoulli put it, "What mortal will ever determine the number of diseases?" How then could we ever calculate the probability of a disease?

Let us deliver a short Sermon on this. Faced with this problem, there are two different attitudes one can take. The conventional one, for many years, has been to give up instantly and abandon the entire theory. Over and over again modern writers on statistics have noted that no general rule for determining the correct HØ is given, ergo the theory is inapplicable and into the waste-basket it goes.

But that seems to us a self-defeating attitude that loses nearly all the value of probability theory by missing the point of the game. After all, our goal is not omniscience, but only to reason as best we can with whatever incomplete information we have. To demand more than this is to demand the impossible; neither Bernoulli's procedure nor any other that might be put in its place can get something for nothing.

The reason for setting up HØ is not to describe the Ultimate Realities of the Universe; that is unknown and remains so. By definition, the function of HØ is to represent what we know; it cannot be unknown. So a second attitude recommends itself; define your HØ as best you can -- all the diseases you know -- and get on with the calculations.

Usually this suggestion evokes howls of protest from those with conventional training; such calculations have no valid basis at all, and can lead to grotesquely wrong predictions. To trust the results could lead to calamity.

But such protests also miss the point of the game; they are like the reasoning of a chess player who thinks ahead only one move and refuses to play at all unless the next move has a guaranteed win. If we think ahead two moves, we can see the true function and value of probability theory in inference.

When we first define HØ, because of our incomplete information we cannot be sure that it really expresses all the possibilities in the real world. Nor can we be sure that there is no unknown symmetry- breaking influence at work making it harder to realize some possibilities than others. If we knew of such an influence, then we would not consider the x_i equally likely.

To put it in somewhat anthropomorphic terms, we cannot be sure that our hypothesis space HØ is the same as Nature's hypothesis space HN. The conventional attitude holds that our calculation is invalid unless we know the "true" HN; but that is something that we shall never know. So, ignoring all protests, we choose to go ahead with that shaky calculation from HØ, which is the best we can actually do. What are the possible results?

Suppose our predictions turn out to be right; i.e. out of a certain set of propositions A_1, A_2, ... A_m the one A_k that we thought highly likely to be true (because it is true on a much larger subset of HØ than any other) is indeed confirmed by observation. That does not prove that our HØ represented correctly all those, and only those, possibilities that exist in Nature, or that no symmetry-breaking influences exist. But it does show that our HØ is not sufficiently different from Nature's HN to affect this prediction. Result: the theory has served a useful predictive purpose, and we have more confidence in our HØ. If this success continues with many different sets of propositions, we shall end up with very great confidence in HØ. Whether it is "true" or not, it has predictive value.

But suppose our prediction does indeed turn out to be grotesquely wrong; Nature persists in generating an entirely different A_j than the one we favoured. Then we know that Nature's HN is importantly different from our HØ, and the nature of the error gives us a clue as to how they differ. As a result, we are in a position to define a better hypothesis space H1, repeat the calculations to see what predictions it makes, compare them with observation, define a still better H2, ... and so on.

Far from being a calamity, this is the essence of the scientific method. HØ is only our unavoidable starting point.

As soon as we look at the nature of inference at this many-moves-ahead level of perception, our attitude towards probability theory and the proper way to use it in science becomes almost diametrically opposite to that expounded in most current textbooks. We need have no fear of making shaky calculations on inadequate knowledge; for if our predictions are indeed wrong, then we shall have an opportunity to improve that knowledge, an opportunity that would have been lost had we been too timid to make the calculations.

Instead of fearing wrong predictions, we look eagerly for them; it is only when predictions based on our present knowledge fail that probability theory leads us to fundamental new knowledge.

Bernoulli implemented the point just made, and in a more sophisticated way than we supposed in that little Sermon. Perceiving as noted that except in certain gambling devices like dice we almost never know Nature's set HN of possibilities, he conceived a way of probing HN, in the case that one can make repeated independent observations of some event A; for example, administering a medicine to many sick patients and noting how many are cured.

We all feel intuitively that, under these conditions, events of higher probability M/N should occur more often. Stated more carefully, events that have higher probability on HØ should be predicted to occur more often; events with higher probability on HN should be observed to occur more often. But we would like to see this intuition supported by a theorem.

Bernoulli proved the first mathematical connection between probability and frequency, today known as the weak law of large numbers. If we make n independent observations and find A true m times, the observed frequency $f(A) = m/n$ is to be compared with the probability $p(A) = M/N$. He showed that in the limit of large n, it becomes practically certain that f(A) is close to p(A). Laplace showed later that as n tends to infinity the probability remains more than 1/2 that f(A) is in the shrinking interval $p(A) \pm q$, where $q^2 = p(1-p)/n$.

There are some important technical qualifications to this, centering on what we mean by "independent"; but for present purposes we note only that often an observed frequency f(A) is in some sense a reasonable estimate of the ratio M/N in Nature's hypothesis space HN. Thus we have, in many cases, a simple way to test and improve our HØ in a semiquantitative way. This was an important start; but Bernoulli died before carrying the argument further.

BAYES

Thomas Bayes was a British clergyman and amateur mathematician (a very good one - it appears that he was the first to understand the nature of asymptotic expansions), who died in 1761. Among his papers was found a curious unpublished manuscript. We do not know what he intended to do with it, or how much editing it then received at the hands of others; but it was published in 1763 and gave rise to the name "Bayesian Statistics". For a photographic reproduction of the work as published, with some penetrating historical comments, see Molina (1963). It gives, by lengthy arguments that are almost incomprehensible today, a completely different kind of solution to Bernoulli's unfinished problem.

Where Bernoulli had calculated the probability, given N, n, and M, that we would observe A true m times (what is called today the "sampling distribution"), Bayes turned it around and gave in effect a formula for the probability, given N, n, and m, that M has various values. The method was long called "inverse probability". But Bayes' work had little if any direct influence on the later development of probability theory.

LAPLACE

In almost his first published work (1774), Laplace rediscovered Bayes' principle in greater clarity and generality, and then for the next 40 years proceeded to apply it to problems of astronomy, geodesy, meteorology, population statistics, and even jurisprudence. The basic theorem appears today as almost trivially simple; yet it is by far the most important principle underlying scientific inference.

Denoting various propositions by A, B, C, etc., let AB stand for the proposition "both A and B are true", $\bar{A}$ = "A is false", and let the symbol p(A:B) stand for "the probability that A is true, given that B is true". Then the basic product and sum rules of probability theory, dating back in essence to before Bernoulli, are

$$p(AB|C) = p(A|BC)p(B|C) \tag{1}$$

$$p(A|B) + p(\bar{A}|B) = 1 \tag{2}$$

But AB and BA are the same proposition, so consistency requires that we may interchange A and B in the right-hand side of (1). If $p(B|C) > 0$, we thus have what is always called "Bayes' Theorem" today, although Bayes never wrote it:

$$p(A|BC) = p(A|C)\ p(B|AC)/p(B|C) \tag{3}$$

But this is nothing more than the statement that the product rule is consistent; why is such a seeming triviality important?

In (3) we have a mathematical representation of the process of learning; exactly what we need for our extended logic. $p(A|C)$ is our "prior probability" of A, when we know only C. $p(A|BC)$ is its "posterior probability", updated as a result of acquiring new information B. Typically, A represents some hypothesis, or theory, whose truth we wish to ascertain, B represents new data from some observation, and the "prior information" C represents the totality of what we knew about A before getting the data B.

For example -- a famous example that Laplace actually did solve -- proposition A might be the statement that the unknown mass M_S of Saturn lies in a specified interval, B the data from observatories about the mutual perturbations of Jupiter and Saturn, C the common sense observation that M_S cannot be so small that Saturn would lose its rings; or so large that Saturn would disrupt the solar system. Laplace reported that, from the data available up to the end of the 18th

Century, Bayes' theorem estimates M_S to be (1/3512) of the solar mass, and gives a probability of .99991, or odds of 11,000:1, that M_S lies within 1 per cent of that value. Another 150 years' accumulation of data has raised the estimate 0.63 per cent.

The more we study it, the more we appreciate how nicely Bayes' theorem corresponds to -- and improves on -- our common sense. In the first place, it is clear that the prior probability $p(A|C)$ is necessarily present in all inference; to ask "What do you know about A after seeing the data B?" cannot have any definite answer -- because it is not a well-posed question -- if we fail to take into account, "What did you know about A before seeing B?".

Even this platitude has not always been perceived by those who do not use Bayes' theorem and go under the banner: "Let the data speak for themselves!" They cannot, and never have. If we want to decide between various possible theories but refuse to supplement the data with prior information about them, probability theory will lead us inexorably to favour the "Sure Thing" theory ST, according to which every minute detail of the data was inevitable; nothing else could possibly have happened. For the data always have a much higher probability on ST than on any other theory; ST is always the maximum likelihood solution over the class of all theories. Only our extremely low prior probability for ST can justify rejecting it.

Secondly, we can apply Bayes' theorem repeatedly as new pieces of information B_1, B_2, ... are received from the observatories, the posterior probability from each application becoming the prior probability for the next. It is easy to verify that (3) has the chain consistency that common sense would demand; at any stage the probability that Bayes' theorem assigns to A depends only on the total evidence $B_{tot} = B_1 \ldots B_k$ then at hand, not on the order in which the different updatings happened. We could reach the same conclusion by a single application of Bayes' theorem using B_{tot}.

But Bayes' theorem tells us far more than intuition can. Intuition is rather good at judging what pieces of information are relevant to a question, but very unreliable in judging the relative cogency of different pieces of information. Bayes' theorem tells us quantitatively just how cogent every piece of information is.

Bayes' theorem is such a powerful tool in this extended logic that, after 35 years of using it almost daily, I still feel a sense of excitement whenever I start on a new, nontrivial problem; because I know that before the calculation has reached a dozen lines it will give me some important new insight into the problem, that nobody's intuition has seen before. But then that surprising result always seems intuitively obvious after a little meditation; if our raw intuition was powerful enough we would not need extended logic to help us.

Two examples of the fun I have had doing this, with full technical details, are in my papers "Bayesian Spectrum and Chirp Analysis" given at the August 1983 Laramie Workshop, and "Highly Informative Priors" in the Proceedings Volume for the September 1983 International Meeting on Bayesian Statistics, Valencia, Spain (Jaynes, 1985). In both cases, completely unexpected new insight from Bayes' theorem led to quite different new methods of data analysis and more accurate results, in two problems (spectrum analysis and seasonal adjustment) that had been treated for decades by non-Bayesian methods. The Bayesian analysis took into account some previously neglected prior information.

Laplace, equally aware of the power of Bayes' theorem, used it to help him decide which astronomical problems to work on. That is, in which problems is the discrepancy between prediction and observation large enough to give a high probability that there is something new to be found? Because he did not waste time on unpromising research, he was able in one lifetime to make more of the important discoveries in celestial mechanics than anyone else.

Laplace also published (1812) a remarkable two-volume treatise on probability theory in which the analytical techniques for Bayesian calculations were developed to a level that is seldom surpassed today. The first volume contains, in his methods for solving finite difference equations, almost all of the mathematics that we find today in the theory of digital filters. An English translation of this work, by Professor and Mrs. A.F.M. Smith of Nottingham University, is in preparation.

Yet all of Laplace's impressive accomplishments were not enough to establish Bayesian analysis in the permanent place that it deserved in science. For more than a Century after Laplace, we were deprived of this needed tool by what must be the most disastrous error of judgement ever made in science.

In the end, all of Laplace's beautiful analytical work and important results went for naught because he did not explain some difficult conceptual points clearly enough. Those who came after him got hung up on inability to comprehend his rationale and rejected everything he did, even as his masses of successful results were staring them in the face.

JEFFREYS

Early in this Century, Sir Harold Jeffreys rediscovered Laplace's rationale and, in the 1930's, explained it much more clearly than Laplace did. But still it was not comprehended; and for thirty more years Jeffreys' work was under attack from the very persons who had the most to gain by understanding it (some of whom were living and eating with him daily here in St. John's College, and had the best possible opportunity to learn from him). But since about 1960 comprehension of what Laplace and Jeffreys were trying to say has been growing, at first slowly and today quite rapidly.

This strange history is only one of the reasons why, today, we Bayesians need to take the greatest pains to explain our rationale, as I am trying to do here. It is not that it is technically complicated; it is the way we have all been thinking intuitively from childhood. It is just so different from what we were all taught in formal courses on "orthodox" probability theory, which paralyze the mind into an inability to see the distinction between probability and frequency. Students who come to us free of that impediment have no difficulty in understanding our rationale, and are incredulous that anyone could fail to comprehend it.

My Sermons are an attempt to spread the message to those who labour under this handicap, in a way that takes advantage of my own experience at a difficult communication problem. Summarizing Bernoulli's work was the excuse for delivering the first Sermon establishing, so to speak, our Constitutional Right to use HØ even if it may not be the same as HN.

Now Laplace and Jeffreys inspire our second Sermon, on how to choose HØ given our prior knowledge; a matter on which they made the essential start. To guide us in this choice there is a rather fundamental "Desideratum of Consistency": in two problems where we have the same state of knowledge, we should assign the same probabilities.

As an application of this desideratum, if the hypothesis space HØ has been chosen so that we have no information about the x_i beyond their enumeration, then as an elementary matter of symmetry the only consistent thing we can do is to assign equal probability to all of them; if we did anything else, then by a mere permutation of the labels we could exhibit a second problem in which our state of knowledge is the same, but in which we are assigning different probabilities.

This rationale is the first example of the general group invariance principle for assigning prior probabilities to represent "ignorance". Although Laplace used it repeatedly and demonstrated its successful consequences, he failed to explain that it is not arbitrary, but required by logical consistency to represent a state of knowledge. Today, 170 years later, this is still a logical pitfall that causes conceptual hangups and inhibits applications of probability theory.

Let us emphasize that we are using the word "probability" in its original -- therefore by the usual scholarly standards correct -- meaning, as referring to incomplete human information. It has, fundamentally, nothing to do with such notions as "random variables" or "frequencies in random experiments"; even the notion of "repetition" is not necessarily in our hypothesis space.

In cases where frequencies happen to be relevant to our problem, whatever connections they may have with probabilities appear automatically, as mathematically derived consequences of our extended logic (Bernoulli's limit theorem being the first example). But, as shown in a discussion of fluctuations in time series (Jaynes, 1978), those connections are often of a very different nature than is supposed in conventional pedagogy; the predicted mean-square fluctuation is not the same as the variance of the first-order probability distribution.

So to assign equal probabilities to two events is not in any way an assertion that they must occur equally often in any "random experiment"; as Jeffreys emphasized, it is only a formal way of saying "I don't know". Events are not necessarily capable of repetition; the event that the mass of Saturn is less than (1/3512) had, in the light of Laplace's information, the same probability as the event that it is greater than (1/3512), but there is no "random experiment" in which we expect those events to occur equally often.

Of course, if our hypothesis space is large enough to accommodate the repetitions, we can calculate the probability that two events occur equally often.

To belabour the point, because experience shows that it is necessary: In our scholastically correct terminology, a probability p is an abstract concept, a quantity that we assign theoretically, for the purpose of representing a state of knowledge, or that we calculate from previously assigned probabilities using the rules (1) - (3) of probability theory. A frequency f is, in situations where it makes sense to speak of repetitions, a factual property of the real world, that we measure or estimate. So instead of committing the error of saying that the probability is the frequency, we ought to calculate the probability p(f)df that the frequency lies in various intervals df -- just as Bernoulli did.

In some cases our information, although incomplete, still leads to a very sharply peaked probability distribution p(f); and then we can indeed make very confident predictions of frequencies. in these cases, if we are not making use of any information other than frequencies, our conclusions will agree with those of "random variable" probability theory as usually taught today. Our results do not conflict with frequentist results whenever the latter are justified. From a pragmatic standpoint (i.e., ignoring philosophical stances and looking only at the actual results), "random variable" probability theory is contained in the Laplace-Jeffreys theory as a special case.

But the approach being expounded here applies also to many important real problems -- such as the "pure generalized inverse" problems of concern to us at this Workshop -- in which there is not only no "random experiment" involved, but we have highly cogent information that must be taken into account in our probabilities, but does not consist of frequencies.

A theory of probability that fails to distinguish between the notions of probability and frequency is helpless to deal with such problems. This is the reason for the present rapid growth of Bayesian/ Maximum Entropy methods -- which can deal with them, and with demonstrated success. And