Edwin T. Jaynes

THE INTUITIVE INADEQUACY OF CLASSICAL STATISTICS

INTRODUCTION

The title of this talk, suggested to me in Professor Agazzi's kind letter of invitation, caused a moment of uncertainty. What variety of statistics is meant by "classical"? J.R. Oppenheimer [1955] held that in science the word "classical" has a special meaning: "[...] it means 'wrong'. That is, the classical theory is the one which is wrong, but which was held yesterday to be right." And indeed, probabilists appear to follow the Oppenheimer usage more or less consistently, advocates of frequency definitions calling Bayesian views "classical"; while Bayesians call frequency views "classical".

Although Oppenheimer (or, "Oppy" as we all called him) was my respected teacher, his usage has always been rather mind-wrenching for me, because in other fields "classical" carries the opposite connotation of "having great and timeless merit". Classical music, sculpture, and architecture are the kind I like. So, in which way should one interpret this title? In the end I opted for Oppy on pragmatic grounds.

If by "classical" professor Agazzi meant "Bayesian" then this would be a very short talk indeed; for I do not believe that there are any intuitive inadequacies in Bayesian statistics. Of course, this is not to say that Bayesian methods require no further technical development; that development has been my main concern for

Epistemologia VII (1984), Fascicolo Speciale - Special Issue.
Probability, Statistics, and Inductive Logic, pp. 43-74.

thirty years. Nevertheless, many of the recent advances in statistics consist of realizing that, on closer examination, past objections to Bayesian methods were either misapplications or empty ideological slogans.

For example, de Finetti's famous exchangeability theorem vindicated Laplace's derivation of the Rule of Succession, which has been attacked by a long list of writers starting with Venn [1866]. Laplace's critics thought that he was assuming all kinds of metaphysical nonsense; in fact, he was assuming only exchangeability. To the best of my knowledge, no claim of an "intuitive inadequacy" in Bayesian methods has been sustained by modern reexamination of these issues.

In the following I shall take "classical", "frequentist", "orthodox", and "sampling theory" as approximately synonymous, in about as close agreement as one can come to a current usage that is not itself quite consistent.

With this interpretation, Professor Agazzi's suggested title becomes highly appropriate, although I wish it were otherwise. One would prefer to take a positive stance: pro-Bayesian rather that anti-orthodox; and that is, of course, my real purpose here. But a principle of logic tells us that theories cannot be proved right, only wrong.

Orthodox significance tests extend this to inference as well as deduction. To argue for an hypothesis H, one proceeds indirectly: first invent a "null hypotheses" H_0 that denies H, then argue againts H_0. Like it or not, that is the methodology we must adopt here also — and not only for logical, but also for psychological, reasons.

It is now about 50 years since de Finetti's first landmark contribution to this field, which started conceptual thinking about probability toward the right track. It is 42 years since Jeffreys demonstrated the superior power and generality of Bayesian methods, not in the abstractness of σ-algebras or philosophical taste, but in the arena of real applications.

Indeed, the variety of applications demonstrated by Jeffreys included all those for which "Student", Fisher, Neyman, and Pearson had developed "orthodox" methods. Since 1960, the Bayesian

literature has grown at an ever-increasing rate, and applications have extended to a dozen new fields with uniformly great success.

A mathematically convenient subclass of Bayesian methods, maximum entropy, succeeds best in just those generalized inverse problems where orthodox methods met their greatest difficulty. Today, with computer programs capable of dealing with dozens to thousands of simultaneous constraints, the Burg-Shore maximum-entropy spectral analysis of time series in geophysics and speech processing, and the Skilling-Gull-Frieden maximum-entropy image reconstructions in optics and radio astronomy, are in constant use. They are routinely extracting detailed information from data in a way that would not have been believed possible a few years ago. On the theoretical side, a mass of Bayes-optimality theorems has accumulated that leave no loophole visible to me. One might think that the point has been made.

Yet all this has had, to the best of my knowledge, zero effect on the thinking and teaching of most purveyors of orthodox statistical doctrine. Their newest textbooks continue to expound the lore of the 1930's and 1940's, complacently ignoring the Bayesian advances that would improve many of their own results, and vastly extend the range of useful applications of their methods.

Why is it that pointing out and demonstrating superior methods – even proving their optimality – has no effect on advocates of orthodox methods? Traditionally, one points to ideological barriers to acceptance of Bayesian notions. But I wonder whether ideology is really such a strong force today in statistics, when it is in such disfavor in all the surrounding fields. Perhaps a more easy explanation is simple inertia. Independently of all ideology, orthodox methods work as well as they did 45 years ago. Persons who were trained to be satisfied with that performance, and continue to study only the problems of 45 years ago, see no need for anything different.

After all, Ptolemaic epicycles still account for the facts of astronomy as well as they did 500 years ago, and the caloric theory of heat still accounts for the facts of thermodynamics as well as it did 200 years ago. Anyone who is satisfied with that performance can ignore what has happened since, and continue using them. But

to continue teaching them is a more serious matter.

Instead of this complacency, I should think that orthodox teachers would be very troubled by the following situation. Who have made the important advances in statistical practice in this Century? Others will judge differently, but my own list is: "Student", Jeffreys, Fisher, Wiener, von Neumann, Shannon, Wald, Zellner, Burg, Skilling. Here we find a chemist, a physicist, a eugenicist, two mathematicians, an economist, an astronomer, two engineers – and only one professional statistician! Whatever list one makes, I think he will find that most of the important advances have come from outside the profession, and had to make their way against the opposition of most statisticians.

Now in physics, if we were to discover that even ten percent of the important advances were coming from non-physicists and being opposed by professional physicists, we would conclude that there was something very drastically wrong in physics teaching. But we have avoided this embarrassment, because physicists stopped teaching the caloric theory some time ago even though it is still as usable as it ever was.

The teaching of orthodox statistics seems to be based on a different psychology. Those ideological barriers – and just plain inertial ones – are still there, and as long as the old methods are usable at all, the fact that something better is available is not enough to overcome them. Rather, we must exhibit important real problems of current interest, which any statistician's clients would expect him to be able to handle, as a matter of professional competence – but for which orthodox methods lead to unacceptable results, or to no results at all. This is, unfortunately, the psychological reason why we must take a negative rather than a positive stance.

All statistical methods, of course, consist of using probability theory in different ways, and to find such problems we need to understand how the reasoning formats of orthodox and Bayesian methods fit into probability theory.

RELATION TO PROBABILITY THEORY

As mathematical system, probability theory consists simply of the basic product and sum rules

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

$$p(A \mid B) + p(\sim A \mid B) = 1 \tag{2}$$

and their consequences. All schools of thought accept these as correct — at least on finite sets and their well-behaved limits, which are all we need consider here. In probability theory, as in any other area of applied mathematics, some paradoxes await those who jump carelessly into an infinite set without considering the limiting process needed to define its properties — but that is their problem, not ours.

We have, then, a well-defined and noncontroversial mathematical machine. But before this machine will run and perform useful services for us, it must be plugged into the real world. It is over how to make the connection that the centuries-old controversies (or, more accurately, misunderstandings) swirl.

To state the difference most succinctly, orthodoxy holds that Equations (1) and (2) are only rules for calculating frequencies; Bayesians claim that they are the general rules for conducting inference of any kind. If I may betray my youth in the fringes of electrical engineering, one camp uses only the minimum connection (data and noise) that will enable the machine to run at all: the other believes that it is capable of delivering far more useful work if fed full three-phase power (prior information as well).

The personalistic approach of Savage [1954] represents an intermediate transitional form, in which one recognized the legitimacy of connecting that third wire, but did not propose any specific way of accomplishing this. It is now of more historical than technical interest because today's applications require that third wire to be fully operational, carrying just as much current (information) as the other two. However, some personal appreciations of Jimmie Savage are given at the end of this work.

In any approach, the reasoning format one can use is determined by the techniques used to make these connections between the mathematics and the real world. In principle, orthodox theory recognizes such a connection only when it consists of empirically observable frequencies. But one can work in statistics for a long time without ever encountering a real problem in which the data actually consist of frequencies. Therefore, to maintain this viewpoint, if frequencies are not already inherent in the nature of the problem and the data, they must be implanted by artificial means.

The technique for doing this is well known. Given some data D_1, we imbed it in a "sample space" $\{D_1, D_2, \ldots, D_N\}$ containing other data that one postulates might have been observed, but were not. Then one introduces a "sampling distribution" consisting of the probabilities

$$(3) \qquad p(D_i \mid H), \qquad 1 \leqslant i \leqslant N$$

that the data set D_i would be observed if some hypothesis H were true. The frequency connection is then made by asserting that, for example, $p(D_2 \mid H)$ is the frequency with which the unobserved data set D_2 would have been obtained in the long run if the experiment were repeated indefinitely with H constantly true. Usually, such sampling distributions are the only probabilities one is allowed to use for inference, because a probability is not considered respectable until a frequency interpretation is bestowed upon it; and this special blessing is reserved for sampling distributions.

But where does the orthodox statistician obtain all this knowledge? What determines the "true" sample space and the "true" sampling distribution? How can one know what the actual frequencies would be? Surely, when one asserts the long-run results of an arbitrarily long sequence of experiments that have not been performed, he is drawing upon a vivid imagination; and not on any fund of actual knowledge of the phenomenon. How is it possible that for decades claims of great "scientific objectivity" for this approach have not been effectively challenged? It cannot have been only physicists and Bayesians who perceive the lack of substance in such pretensions.

The answer must be that for decades workers have been cowed by the oppressive weight of authority in this field. Indeed, Jimmie Savage [1962a] used just this term in recalling his own early experiences. The path of least resistance – also the one safest for one's worldly career – is to put up a public front, giving lip service to things which we believe to be false, remaining silent on what we see as the truth, out of fear of the "clamor of the Boeotians". We know that Newton, Gauss, and von Neumann delayed publication of some of their most original ideas for this reason.

It is not only in science that this false public front is expedient. At the turn of the century, Jules Massenet enjoyed enormous public success with his religious and operatic music; but he said privately to Vincent d'Indy, "I don't believe in all that creeping Jesus stuff, but the public likes it, and we must always agree with the public."

Illusions of objectivity are preserved, not so much by authority imposed from above, but by the ring of authority in an official language that encourages them. "It is a gaussian random process" sounds very much like a statement of physical fact; i.e., something that is true or false independently of anybody's state of knowledge. The notions of sample space, population from which we draw, and sampling frequencies, are almost always represented as if they were physical facts. Like religion, this gives a certain feeling of security that the statistical "public" likes.

Yet almost everyone has lucid moments in which he recognizes that these representations cannot be really true. Fisher [1956] observes:

> [...] the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination through the hypotheses he has decided to test [...]

Lindley [1971] notes:

> A statistician faced with some data often imbeds it in a family of possible data that is just as much a product of his fantasy as is a prior distribution. A good example occurred recently in a paper of Edwards [1970]. The data here are the distribution of human blood-groups in the world at the present day. What repetitions of this experiment are envisaged to provide a sample space?

But these lucid moments are rare, and the illusions of authority artificially created by our language continue to dominate the way we formulate, and think about, problems of inference. In the case of a time series, that language almost forces us to believe that there exists a "true but unknown" frequency distribution, mean, covariance, power spectrum, etc. which we are to estimate by various means. In occasional lucid moments we must recognize that these things are only figments of our imagination. What does it mean to "estimate" a figment? Our real goal, almost always, is to obtain a predictive distribution.

This supposed frequency connection encounters some problems even for estimating a quantity, such as the velocity of light, which everyone believes is something real; and not a figment. Suppose we are trying to estimate an "objectively real" location parameter θ from a single measurement; i.e., we use the sampling distribution $p(x \mid \theta) = f(x - \theta)$ for the observed datum x. We can think of $(x - \theta)$ as the "error" or "noise" in the measurement. But this sampling distribution might describe two quite different things: (A) the frequencies of errors in many repetitions of the measurement; (B) the probabilities of error in the specific measurement actually made. Orthodoxy fails to distinguish between these meanings; yet it is clearly (B) and not (A) that is directly relevant to our inference about θ.

Indeed, if we had independent evidence (B) telling us our actual error, the frequency of errors (A) in other measurements would be completely irrelevant. As a moment's thought will show, for validity of the orthodox reasoning which draws inferences about the present case from interpretation (A) it is necessary not only that we have ample data from other measurements to determine those frequencies, but also that we have *no* information about the actual error in the present measurement, beyond those frequencies. I know of no experiment in which these conditions were met.

This reminds us of Fisher's admonition that the validity of some orthodox reasoning depends on the absence of recognizable subsets. In his last book (Fisher [1956]) he recognizes also that fiducial inference is valid only when we have no prior information. This work gives considerable support to a conjecture sometimes heard

– that if Fisher were alive today he would now be a Bayesian.

If one examines the actual procedures – as distinguished from the precepts – used in setting up orthodox statistical problems, it is seen that sampling distributions are not, in practice, determined by any consideration of frequencies. They are calculated from theoretical models (Bernoulli trials, Poisson process, ARMA model, etc.) or simply adopted by convention (iid normal errors, Weibull failure law, etc.). It would not be easy to cite an orthodox work which presented empirical evidence that the sampling distribution used was indeed a real frequency.

The Bayesian reasoning format seeks to relate the mathematics to the real world in a totally different way. We do not proceed indirectly through supposed connections with frequencies of imagined data sets that might have been observed but were not. Our mathematics is attached directly to the real world by the fact that our probability distributions represent our actual state of knowledge of that world; i.e., they are conditional on the one real data set that was observed, and on the prior information that we actually do have. That is our connection with "reality" and I submit that it is vastly more objective and scientific than one that has to conjure up frequencies in an imaginary universe, while ignoring cogent prior information in this universe.

There is a famous expression of the opposite view. Norman Campbell asserted that anyone who tried to claim that probability in a physics experiment meant anything different from frequency, "[...] would convince us of nothing but his ignorance of physics." On the contrary, throughout the history of this subject the persons who have argued most strongly against frequency interpretations have been physicists. As a professional physicist, I can assure you that there are reasons for this in the fact that physicists are trained to direct their attention to the cause-effect relations controlling real physical phenomena – an attention that is conspicuously missing in frequentist descriptions of those phenomena.

Indeed, it appears to me that maintaince of a frequency view *requires* one to ignore virtually all the professional knowledge that physicists have – because that constitutes "prior information" that would invalidate a naive frequency interpretation. This corre-

sponds closely to what we noted five paragraphs before.

Of course, if the only information we have about a phenomenon is its observed frequency, then the probability we shall assign to it will be equal to that frequency. But this, far from standing in conflict with Bayesian principles, is an elementary mathematical consequence of those principles. I have the impression that this has been demonstrated adequately in every Bayesian work since Laplace's memoir of 1774 on the *Probability of Causes.* Laplace, Maxwell, Gibbs, Poincaré, Jeffreys, and Cox have not succeeded in persuading anyone of their ignorance of physics.

However, after all these criticisms we must admit that orthodox practice is often far more defensible than orthodox precepts, because common sense is powerful enough to make one lay the latter aside when they become too obstructive to sanity. A good example is the tail-area criterion for orthodox significance tests, where the precept is surely wrong; yet in some cases I should defend the actual practice against criticisms made by other Bayesian.

TAIL-AREA REASONING

It is true that, in an orthodox significance test or confidence interval not based on sufficient statistics, using the tail area of a sampling distribution as a criterion can lead to absurdly false conclusions. But this absurdity is clear also to orthodox statisticians; and the tests in common use do use sufficient statistics. As a result, the orthodox, fiducial, and Bayesian results cannot be very different unless the Bayesian has essential prior information not usable by the others. But let us have a little fun with tail areas anyway.

Jeffreys [1939] poked fun at the orthodox tail-area reasoning in an amusing tongue-in-cheek way, noting that the null hypothesis is rejected "[...] because it has not predicted observable results that have not occurred. This seems a remarkable procedure." I should like to add my own version of this joke: In the orthodox test, the sole basis for decision is probabilities conditional on the null hypothesis H_0. Suppose, then, that we reject H_0. Surely, we must

also reject probabilities conditional on H_0; but then what is the *a posteriori* justification for the decision? Orthodox logic saws off its own limb.

This version applies to more than tail-area thinking. Any rule for rejecting hypotheses based only on sampling distributions conditional on the very hypotheses being tested, gets itself into this limb-sawing logical difficulty. The Bayesian reasoning format avoids it, since we decide on the grounds of probabilities of the hypotheses being tested, conditional on the data and the prior information.

Still, this is really only a joke; for any Bayesian who examines the mathematics unblinded by his own ideology can discover that if the orthodox test is based on sufficient statistics then there are mathematical connections. Indeed, consider the orthodox t-text, F-test, or tests for the parameter of the binomial or Poisson distribution. If we test a simple null hypothesis against a one-sided alternative, the orthodox conclusions are not merely similar to, but identical with, the Bayesian conclusions from a noninformative prior (Jaynes [1976]). If such connections did not exist, common sense examination of the results would have rejected orthodox tail-area tests long ago.

It has always seemed to me that one of the most ridiculous spectacles in statistics is the orthodox texbook writer who warns his reader not to use those awful Bayesian methods on these problems; and then offers as a "more objective" alternative an orthodox method which leads, in the entire class of problems considered, to precisely the same result. Clearly, in the view of such writers the Bayesian's sin cannot lie in any property of his actual calculations; but only in his failure to pronounce the proper incantations over them.

But for this same reason, it would be ridiculous for me to reject an orthodox tail-area test; and then advocate a Bayesian test that leads to precisely the same result. Indeed, a Bayesian will – and I think properly and necessarily – use tail-area criteria for some of his own decisions. But if there are no sufficient or ancillary statistics, or if there are nuisance parameters, the Bayesian will avoid the absurdities into which orthodox tail-area reasoning can lead,

because he will use the tail-area of a posterior distribution that contains all the relevant information; and not the tail-area of a sampling distribution that contains only part of the information.

Several examples illustrating these remarks are given in Jaynes [1976]. Since the result is not a criticism of tail-area reasoning *per se*, we must look further.

PRIOR INFORMATION

Savage ([1954], p. 4) finds the difficulty with orthodox methods of inference in the fact that on this viewpoint one cannot use probability to express the degree of plausibility of anything but a "random variable". This is indeed a serious difficulty, as I have also pointed out several times. But this too has not been a fatal defect, because if our ideology forbids us to use probability for expressing the "measure of trust" to be put in an hypothesis, the human mind is still able to invent any number of *ad hockeries* – Chi-squared, likelihood, significance level, confidence level, power functions – to replace it. These substitutes, if not optimal, were at least usable, and from a pragmatic standpoint frequentist practice did not actually suffer very much from them.

Of course, from an aesthetic standpoint this collection of *ad hockeries* now appears to us as awkward, often far from optimal, and – worst of all – unnecessary. The principles of probability theory, Equations (1) and (2) above, already contain these things. Jeffreys [1939] has shown this, more clearly and in more detail, fifteen years before Savage did. But scientific practice is not determined by aesthetic considerations; as noted above, to have any effect a difficulty must not only extend to the pragmatic level, it must be so serious that is stops us from functioning altogether.

The basic suggestion that I want to make here is that the really fatal objection to orthodox methods is one that Savage never recognized and which applies to his own position as well (although this is largely my own fault; see the concluding remarks). Neither frequentist nor personalistic approaches give us any definite procedure for taking prior information into account. Of course, as

long as we work only on problems where we have almost no prior information, this does not seem to be a serious difficulty. Unfortunately, both Fisher and Savage seem to have concentrated all their attention on such problems.

Fisher, in his maxim "Let the data speak for themselves", appears to consider it almost a principle of morality that we *must not* allow ourselves to be influenced by prior information; at any rate, that is the interpretation of his position that others have made. As the writer remembers very well, this morality was infused into students in the 1940's; we were taught that it is not only illogical, but a reprehensible breach of "scientific objectivity" to use prior information. Savage's *Principle of Precise Measurement* is a faint echo of this, holding that it is all right to use diffuse priors because they won't make much difference anyway.

John W. Tukey [1978] has noted that this attitude toward prior information puts orthodox statistics in a curious position. It is held to be decent to use judgment in deciding what parameters should be present in a model; but then indecent to use judgment to help us in estimating their values. Yet, as Tukey observes, before 1973 judgment offered a far better basis than all the world's time series for estimating the factors related to oil prices in econometric models.

To clinch our arguments here, we need only amplify Tukey's observation, no doubt carrying his line of thought further than he himself would wish to. But what he pointed out leads us to a class of real problems of inference in which orthodox methods fail so completely that they cannot be redeemed by any *ad hockery*. In fact, such problems have been well known for thousands of years.

THE GENERALIZED INVERSE PROBLEM

To the best of the writer's knowledge, neither Fisher nor Savage ever considered a generalized inverse problem, although they have been the crux of medical diagnosis since Hippocretes. Indeed, in the statistics classroom an orthodox professor may teach his students to ignore prior information; yet if his personal physician

ignored his medical history in diagnosis, he would hold the man guilty of malpractice.

But a physician innocent of malpractice is reasoning according to a totally different format than that taught in the orthodox statistical classroom. When apprised of your symptoms, he does not start by imbedding them in an imaginary sample space. Instead of thinking about the class of all symptoms you might have had but don't, he thinks about the class of all disorders consistent with the symptoms you do have. The first one he will test for is the one which, in that class, appears *a priori* the most likely, from your medical history.

[In a discussion of this in June 1981, John Tukey objected that a psychiatrist might very well wonder: "If the facts were A, what is the probability that this patient would tell me B?" Perhaps he knows something about this that I don't; so let us agree that the physician in our story is *not* a psychiatrist.]

The reasoning format here is in a sense that opposite of that supposed by orthodox statistics. Let us state it in very general, symbolic terms. There are a number of conceivable "diseases" or states of nature $[x_1, x_2,\ldots,]$ and we obtain data y that we write as

$$y = Ax \tag{4}$$

determined by the unknown true state x, where A is a deterministic operator, assumed known. However, A is singular (i.e., the same y could result from more than one x), and so we cannot invert this relations to determine x form y. We must be content with making some guess, or "estimate"

$$\hat{x} = By \tag{5}$$

where B is a "resolvent" operator to be chosen.

We have here no "noise", and therefore no sampling distribution except in the rudimentary sense that $p(y_i \mid x_j) = 1$ or 0, depending on wheter y_i is or is not the data set resulting from state x_j. The observed data y_{obs} tells us only that the true x_j must lie in the class C for which $p(y_{\text{obs}} \mid x_j) = 1$.

This is a fairly realistic description of the medical diagnosis problem, since it is typical that a given cause produces definite symptoms, but a given symptom may have many different causes. It is even more realistic for many problems of inference in present technology, where the essence of the problem does not lie in random "noise" perturbing our data, but rather in the fact that our information, although essentially noiseless, is incomplete.

For example, in spectral analysis we have accurate measurements of a function $f(t)$, but only on a finite set of times $\{t_1,...,t_n\}$. Given this noiseless but incomplete information, make the best estimate of its power spectrum

$$p(\omega) = \left| \int f(t) e^{i\omega t}\, dt \right|^2 . \tag{6}$$

Or in image reconstruction, let $\{x_1,...,x_n\}$ be the luminances of the elements of a true scene, while our data consist of noiseless but incomplete values for the elements of our image

$$y_i = \Sigma_j A_{ij} x_j \quad , \quad 1 \leqslant i \leqslant m < n \tag{7}$$

and A_{ij} is the digitized point-spread function of our imperfect telescope.

In each of these cases, the data can tell us only that the true spectrum or the true scene must lie in a certain class C of possibilities, but orthodox statistical principles, finding a constant likelihood for all of them, are helpless to make a definite decision within that class.

Returning to our general format, $y = Ax$, $\hat{x} = By$, it would appear that any rational method for choosing the resolvent operator B must have, at the very minimum, the property that the estimate $\hat{x}$ lies in the class C of *possible* causes: for all x,

$$y = Ax = A\hat{x} = ABy = ABAx$$

The resolvent operator must therefore be a generalized inverse:

$$ABA = A \tag{8}$$

By a "pure generalized inverse" problem we mean one in which there is no noise, and the likelihood is strictly rectangular, so Eq. (8) contains all that orthodox statistics can tell us.

However, some orthodox statisticians have not even complied with condition (8). For example, in the limit of zero noise, "window" methods of spectral estimation (Brillinger [1980]) do not satisfy condition (8), and so these estimates can conflict with what we know from deductive reasoning. This takes two forms: (1) the spectrum, by definition non-negative, can nevertheless be estimated to be negative; and (2) the estimated spectrum disagrees with the autocovariance data at every data point where the window function $W \neq 1$. Yet, as we deplored in our opening remarks, such methods are still being put into new textbooks and taught.

By contrast, the Bayesian method deals with generalized inverse problems without any difficulty, leading automatically to estimates that satisfy condition (8). The posterior probability $p(x \mid yI)$, where I stands for prior information, is zero outside class C, and within that class it is proportional to the prior probability $p(x \mid I)$; just the medical diagnostician's common-sense reasoning.

In real problems we often have a great deal of highly cogent prior information, so that for a Bayesian the choice between different possibilities in class C may be extremely sharp and definite. For example, suppose class C contains only two possibilities, x_1 and x_2. But x_1 is compatible with $W_1 = 10^{10}$ different quantum states, while x_2 can be realized in $W_2 = 10^{20}$ independent ways. With a multiplicity ratio $W_2 / W_1 = 10^{10}$, our decision problem is not really very difficult. With sufficient knowledge of the laws of physics, these multiplicity factors can be calculated, often by quite nontrivial combinatorial methods.

If our prior information consists entirely of multiplicity factors (as it does in most of the current problems), then the Bayesian's optimal estimate $\hat{x} = By$ will be that one which, in class C, has the greatest multiplicity $W(x)$. At this point, perhaps our story will start to sound familiar again; for the quantity $H(x) = \log W(x)$ is just what we call the "entropy" of x.

The original pure generalized inverse problem, therefore, was just the statistical mechanics of Boltzmann and Gibbs. Their algorithm:

the macroscopic state which is overwhelmingly more likely than any other is the one which has maximum entropy subject to the constraints of the data and the laws of physics. As we hope to show elsewhere, all of presently known statistical mechanics – equilibrium and nonequilibrium – is contained in this algorithm, suitably generalized to allow for space-time variations. A preliminary survey is given in Jaynes [1980].

What is new and exciting in statistics today is, however, the quite recent realization that this maximum entropy principle applies in a beautiful way to the aforementioned spectral analysis and image reconstruction problems, and to "black hole thermodynamics" in general relativity. A spectrum function, a scene, or a black hole possesses a calculable entropy; and the one which has maximum entropy subject to the available data is overwhelmingly the most likely to occur in Nature. It was surprising to all of us – although mathematically elementary – to realize that in such problems multiplicity ratios of 10^{10} or more are not unusual. Further details may be found in Burg [1975], Childers [1978], Gull and Daniell [1978], Bekenstein [1981]; and in a mass of other papers currently in process of writing or publication, for which references are not yet available.

Today, generalized inverse problems have become so important in engineering (for example, control systems) that mathematicians have devoted a great deal of attention to them. However most, being still encumbered by frequentist views of probability, have not recognized these as being problems of inference at all (there is no "randomness" in sight); and so they call them "ill-posed problems." We have, for example, the treatise of Tikhonov and Arsenin [1977] which explores a variety of *ad hoc* algorithms for their inversion, but never at any point recognizes that the only rational basis for choosing one algorithm over another lies in our prior information. From a Bayesian standpoint, these problems are often very well posed, with unique and useful solutions. The services of a few Bayesians are much needed in this field.

Generalized inverse problems are also, or very soon will be, important in economics. Here, many struggle with the seemingly ill-posed problem of seasonal adjustment of economic time series. The

U.S. Census Bureau's XII program for this has been in use since the middle 1960's; but in 1979 three different Government-sponsored committees of Statisticians were searching for improved methods.

The Bell Laboratories SABL program was announced recently (Cleveland et al. [1980]) as a major advance over XII. However, from the writer's discussions with Cleveland in December 1980, it appears that SABL is still an entirely orthodox procedure, with the defects that one can readily anticipate; in problems where the essence lies not in the "randomness" of the data, but in its incompleteness, one had better do some very careful searching of his fund of prior information, and use it. In this field also, the services of a few Bayesians are clearly needed; although W.S. Cleveland is a former student of Jimmie Savage, the personalistic approach has never taken seriously the explicit conversion of prior information into prior probabilities.

It is hardly surprising that it is geophysics, the field in which Sir Harold Jeffreys works, that has developed the most highly sophisticated methods of time series analysis. But we have no more time and must refer the reader to the extensive review of Smylie, Clarke, and Ulrych [1973].

In conclusion, it appears to the writer that in generalized inverse problems we have the ultimate "intuitive inadequacy" that will be the fatal Achilles heel of orthodox statistical principles. Orthodox concentrates attention on a "randomness" that is not present in the problem – but fails to recognize the prior information without which there is no criterion for solution. Bayesian methods, which had been shown by Jeffreys [1939] to be more powerful than orthodox methods in the very problems for which orthodox methods were developed, work just as well in generalized inverse problems, and in so doing achieve the long-needed unity of statistical inference and statistical mechanics.

ADDENDUM - IN MEMORY OF JIMMIE SAVAGE

Knowing from his own lips of the happy and productive time

Jimmie Savage spent in Italy working with Bruno de Finetti, it is impossible to close a talk about Bayesian statistics – on Italian soil, and with Professor de Finetti so much in our thoughts – without some personal remarks.

It is now ten years since Jimmie Savage left us, but the feeling of tragedy has hardly lessened. Dozens of reminiscences about him have appeared; my reason for wanting to add still another is that the man had so many sides that those who knew him see his life and influence in many different ways.

The most common reaction was shock at his sudden and unexpected passing at the relatively early age of 54. Jimmie was five years older than I, and so I have now lived five years longer than he did, and am in a position to know that 54 years is not nearly enough to accomplish all that one has in him. Still, it is quite long enough to accomplish something; the history of science and art teaches us that virtually all creative people reach the peak of their intellectual powers in their 20's and 30's. So, let us at least take note of the fact that Jimmie Savage had a longer life than did Spinoza, Shakespeare, Mozart, Bellini, Chopin, Donizetti, Bizet, Abel, Jacobi, Riemann, Maxwell, or Fermi.

The side of the tragedy that touched me most directly was our failure to reach agreement on some very fundamental issues of statistical practice. As many readers will know, each of us is on record as criticizing the other's position with respect to the determination of prior probabilities (Jaynes [1968], [1976]; Savage [1977]). Now that it is too late, I fell that I finally understand what our difficulty was, and if we could only talk again our differences would be resolved in an hour. But the sadness is not just personal; I fear that this failure may have condemned statistics to more years of useless debate over issues that do not really exist any more.

Psychologically, the difference between Savage and me in the 1960's was only a continuation of that between Fisher and Jeffreys thirthy years earlier. Both Fisher and Savage had, obviously, deeply penetrating insight that saw at once what was relevant and what was irrelevant in a problem of inference. Why then, did they not see the simple, obvious things that the physicists – Maxwell,

Boltzmann, Gibbs, Jeffreys, and I – all saw at once?

Superficially, it appeared to everybody that the differences were over the "philosophical meaning of probability"; and as long as both sides argued on that level, the stalemate was bound to continue. But as I now realize, both Fisher and Savage were advocating positions and principles that were adequate and appropriate for the problems they had in mind; while the physicists were thinking in the context of an entirely different class of problems.

In these arguments, the two sides were talking past each other, each saying what the other found meaningless in the context of his own problems. When Fisher insisted that probabilities ought to be frequencies, he was thinking of sampling distributions; when Jeffreys denied it he was thinking mostly of prior distributions. Fisher did not envisage the possibility that one would have prior information that cannot be properly expressed by choice of a model, but which was nevertheless so cogent that it must be incorporated explicitly into the process of inference; Jeffreys, with his background in physics, was thinking of just such problems as the general kind calling for inference. As long as neither emphasized these qualifications, the difference appeared to be philosophical.

My difference with Savage was very much like this. At first glance, it seemed to be a difference over what was later called "subjective Bayesian" and "objective Bayesian" views of probability. Neither of us realized that we were thinking of different problems; and so in spite of my repeated denials Savage persisted in accusing me of holding "necessary" views of probability, while I accused him of openly condoning inconsistency by failing to see the need for normative rules determining prior probabilities from our prior information.

It was from Savage's book [1954] that I first learned of de Finetti and the modern theory of exchangeable (or as they were then called, symmetric) sequences; this went immediately into my lectures at Stanford and many other places, in the period 1955-1960. Strangely enough, Jeffreys [1939], [1948] did not seem to know of de Finetti's work, which would have fit in so beautifully with his own. Even more surprising, it was clear that Savage had not understood Jeffreys' position, which was for me the definitive

statement of probability theory.

Almost upon my arrival in St. Louis in 1960, I met William and Esther Sleator and learned that thay had known Savage since boyhood, were still in touch with him, and that all his friends called him "Jimmie". This inspired me to approach Savage through them; I wrote an analysis comparing Jeffreys' theory with Savage's definition of "necessary" and tried to show that Laplace and Jeffreys were not necessarians (in fact, I do not know of anyone, in the entire history of probability theory, who has ever held such a view; Keynes is the only one who comes close to it), and their work had far more merit than Savage had recognized. This was transmitted to Savage by the Sleators.

Jimmie's reaction, characteristic of him, was not to respond until he had done a much deeper analysis of my own work. He located a copy of my Socony-Mobil lectures (Jaynes [1958]), wrote marginal comments on almost every page, and brought it with him when we finally met in Dallas, Texas on June 25, 1963. The Socony-Mobil Research Laboratories had arranged the meeting by inviting us both to visit them on that day.

In the morning, in an audience of about 60 others, Jimmie listened to a 90-minute lecture by me on prior probabilities; whereupon I sat down and listened to a 90-minute lecture by him on the same topic. All afternoon we were closeted together privately, having a technical discussion unlike any I have ever had before or since. What made it unique was that we had prepared ourselves so well; each was familiar with the other's work, and came with a mass of written comments and questions. At the same time – let us admit it – each wanted to do a little probing to find out just how much the other really knew.

Finally, we were dragged out to attend a kind of banquet in our honor; but we insisted on sitting together and continuing our discussion – practically ignoring our kind hosts – all through dinner and past midnight. But our hosts managed surreptitiously to keep our wine-glasses filled, and no doubt noted with satisfaction that with each sip we became more fiendly but less coherent.

Of course, we covered a dozen different topics on which we were in full agreement; so it is a gross distortion of the scene that I now

mention only the two on which we disagreed. Our mathematical tastes were very different; in commenting on my work, up to the determination of prior probabilities he agreed with practically everything I had said, but wanted me to make the derivations more rigorous by restating them in measure-theory terms. I replied that I did not consider measure theory arguments more rigorous, only more general; but it was a generality that was not needed in problems of the real world and both our works would be intelligible and useful to a wider audience if we avoided them.

Likewise, in my comments on Savage's work, up to the determination of prior probabilities I agreed with everything he said, but felt that at the foundation level the consistency desiderata of R.T. Cox [1961] were more elementary – and therefore more compelling logically – than the notion of coherence. He replied that I ought to read de Finetti more carefully, and told me of their work together.

Indeed, each of us had written a review of Cox's book, mine (Jaynes, [1963]) higly laudatory and his (Savage [1962b]) quite critical. Appearing almost simultaneously was his little book of mutually admiring conversations with British orthodox statisticians (Savage [1962a]); and the contrast was too much for some. Jimmie told me that, as soon as my review appeared a colleague sent him a copy of it with a note asking, "Why don't you write reviews like this one? Why do you always criticize your friends and cozy up to your enemies?" I had felt exactly the same dismay, and realized only much later that he was, as always, taking what he saw as the wise course best calculated to influence both for the better.

But while I can now appreciate the "far-seeing grand strategy" aspect of his policy, it did – and still does – seem to smack of the "new morality" that persons like Jeffreys and Cox, whose work was almost entirely correct and of the greatest importance, could expect from Savage no appreciation but only criticism for the small imperfections that remained; while others whose work was almost entirely wrong and misleading received no criticism, only praise for finding one grain of truth. I could never behave in that way.

The above matters were only differences in taste, to be expected

between two people who insist on doing their own thinking; and did not really trouble either of us. Our serious disagreement was over the determination of prior probabilities. Put most briefly, he wanted prior probabilities to express prior opinions; I wanted them to express prior information. But this is just the point on which we were talking past each other. He thought of parameters as no more than real numbers, and did not associate them in his mind with any such thing as multiplicity factors. I thought of parameters as projections of a deeper "microscopic" reality onto the macroscopic world; and their multiplicities were seen as absolutely crucial to inference about them.

It is now clear that this misunderstanding was almost entirely my own fault, because I thought I had to talk down to him, removing the physicist's technical details and jargon, and stating things in a way that I thought would sound familiar to a statistician. This was, of course, just the worst thing I could possibly have done; there was no need to spare him the technical details, and the result was that I failed to put across to him even the existence of multiplicity factors in the problems I was concerned with. So he thought that I was espousing a "necessary" view of probability; and I thought that he was ignoring vital information. We never managed to clear this up.

Surely, if Savage had ever studied a pure generalized inverse problem such as physicists have always had in statistical mechanics (for example, given the energy of a system, predict its pressure), he would have perceived instantly that the lore of "diffuse priors" for a parameter does not work here. The only basis that exists for choosing one estimate over another is explicit, quantitative prior information about the laws of physics; and any method of inference that fails to offer mathematical rules for translating that prior information into an explicit prior probability assignment, is simply helpless to deal with such problems. He would have discovered the principle of maximum entropy for himself, very quickly — just as Boltzmann had done in 1877.

Savage [1976], in almost his last work, made some revealing comments about the mathematical difference between him and Fisher. He deplored the fact that Fisher seemed "ignorant of those

parts of mathematics in which I was best trained". But it seemed to me that the difference in their mathematical training was just the place where Fisher had the greatest advantage over Savage.

Once he finished his anti-Bayesian polemics and got down to the mathematics, Fisher did not waste one line on irrelevancies; he proceeded directly to the problem at hand, and did the analytical calculation that needed to be done, efficiently and correctly. Savage was so busy over minutiae of infinite sets and measure theory that, as noted also by Lindley [1981], he seems never to have got around to developing a solution or working method useful in problems of the real world.

Perhaps this was the greatest tragedy of all; if Jimmie Savage had been able to shake off his abstract mathematical training and thereby fully appreciate the work of competent, down-to-earth analysts like Jeffreys and Fisher, his own contributions to statistics might have been more solid and lasting.

As it is, Savage exerted a great and beneficial influence on the general thinking of his contemporaries — far more than Jeffreys was able to accomplish and, with his unique tact and willingness to work hard to understand another person's viewpoint, more than any other person could have accomplished. But for the permanent, functional substance of Bayesian methods we are still dependent almost entirely on the work of Jeffreys and de Finetti, done before Savage entered the field.

Arthur Holly Compton Laboratory of Physics
Washington University

REFERENCES

J.D. Bekenstein [1981], "Gravitation, the Quantum, and Statistical Physics", in Y. Neeman (ed.), *To Fulfill a Vision,* (Jerusalem Einstein Centennial), Addison-Wesley, Reading, (Mass.), 1981.

D.R. Brillinger [1980], *Time Series,* Holden-Day, San Francisco, 1980.

J.P. Burg [1975], *Maximum Entropy Spectral Analysis,* Ph.D. Thesis, Stanford University, 1975.

D.G. Childers (ed.) [1978], *Modern Spectrum Analysis* (a reprint collection); IEEE Press and John Wiley and Sons, New York, 1978.

W.S. Cleveland, S.J. Devlin, and I.J. Terpenning [1980], "A Comparison of SABL and XII-Seasonal Calendar Adjustment"; Presented at the 140'th annual meeting, American Statistical Association, Houston, Texas, August 11-14, 1980.

R.T. Cox [1961], *The Algebra of Probable Inference,* Johns Kopkins University Press, Baltimore, 1961.

A.W.F. Edwards [1970], "Estimation of the Branch Points of a Branching Diffusion Process", *J. Roy. Stat. Soc.* (1970), B 32.

R.A. Fisher [1956], *Statistical Methods and Scientific Inference,* Hafner Publishing, New York, 1956.

S.F. Gull and G.J. Daniell [1978], "Image Reconstruction from Incomplete and Noisy Data", *Nature* (1978), 272, pp. 686-690; see alzo IEEE Proc (E) 5 (1980), 170.

E.T. Jaynes [1958] *Probability Theory in Science and Engineering,* Socony-Mobil Oil Company, Dallas (Texas), 1958.

E.T. Jaynes [1963], Review of Cox (1961), *Am. Jour. Phys.* 31 (1963), 66.

E.T. Jaynes [1968], "Prior Probabilities", IEEE Trans. Syst. Sci. and Cybern. SSC-4, Sept. 1968, pp. 227-241, Reprinted in *Concepts and Applications of Modern Decision Models,* V.M. Rao Tummala and R.C. Henshaw (eds.), Michigan State University Business Studies Series, 1976.

E.T. Jaynes [1976], "Confidence Intervals vs. Bayesian Intervals", in W.L. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science,* D. Reidel, Dordrecht (Holland), 1976.

E.T. Jaynes [1980], "The Minimum Entropy Production Principle", *Annual Reviews of Physical Chemistry* 31 (1980), pp. 579-601.

H. Jeffreys [1939], [1948], *Theory of Probability,* Oxford University Press.

D.V. Lindley [1971], "Estimation of Many Parameters", in V.P. Godambe and D.A. Sprott (eds.), *Foundations of Statistical Inference,* Holt, Rinehart and Winston, Toronto, 1971.

D.V. Lindley [1981], "L.J. Savage – His Work in Probability and Statistics", *The Writings of Leonard Jimmie Savage – A Memorial Selection,* Published by The American Statistical Ass'n and The Institute of Mathematical Statistics, 1981, pp. 37-60.

J.R. Oppenheimer [1955]. "Analogy in Science", presented at the 63rd Annual Convention of the American Psychological Ass'n, San Francisco, September 4, 1955.

L.J. Savage [1954], *The Foundations of Statistics,* John Wiley and Sons, New York, 1954.

L.J. Savage [1962a], discussion in M.S. Bartlett (ed.), *The Foundations of Statistical Inference,* Methuen, London, 1962.

L.J. Savage [1962b], Review of Cox (1961); *J. Am. Stat. Assoc.* 57 (1962), pp. 921-922.

L.J. Savage [1976], "On Rereading R.A. Fisher", J.W. Pratt (ed.), *Annals of Statistics* 4, 1976, pp. 441-500.

L.J. Savage [1977], "The Shifting Foundations of Statistics", in R.G. Colodny (ed.), *Logic, Laws, and Life,* Univ. of Pittsburgh Press, 1977, pp. 3-18.

D.E. Smylie, G.K. Clarke, and T.J. Ulrych [1973], "Analysis of Irregularities in the Earth's Rotation", in B.A. Bolt, B. Alder, and S. Feinbach (eds.), *Methods in Computational Physics,* Vol. 13, Academic Press, New York, 1973.

A.N. Tikhonov and V.Y. Arsenin [1977], *Solutions of Ill-Posed Problems,* John Wiley and Sons, New York, 1977.

J.W. Tukey [1978], "Discussion of Granger on Seasonality", in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series,* U.S. Government Printing Office, 1978, pp. 50-53.

DISCUSSION

Question Cohen to Jaynes

I'd like to raise a small point about medical diagnosis and the Bayesian approach to it. Now there is of course, as I am sure Professor Jaynes knows, a great deal of controversy in the medical literature about this question. One of the objections that have been raised against Bayesianism is in regard to the diagnosis of rare diseases. If you have a disease which is very rare, and if you train your physicians to calculate the posterior probability of a particular diagnosis by Bayesian methods, there is a risk that a correct diagnosis of your rare disease will not be made; its very low prior probability may infect the eventual estimate, as it were. The alternative technique is a system of what is called flow-chart diagnosis whereby you seek to eliminate some possibilities at each branching node (compare L. Jonathan Cohen, "Bayesianism versus Baconianism in the Evaluation of Medical Diagnoses", *British Journal for Philosophy of Science* 31 (1980), pp. 45-62). Now this is I think a little more like what one might call the classical Baconian method. There is a sort of long-term risk here that if people were to adopt Bayesian methods systematically and therefore miss quite often or relatively often the diagnosis of rare diseases would be affected and in a subjective sense these diseases would get rarer and rarer and more and more people would fail to be diagnosed who actually have them. There is a sort of double risk here.

Now of course there may be ways of adding cautions to the Bayesian method which avoid this danger. But the antiBayesians in the medical profession, as I read the literature, claim that this is not the case and that there is a standing risk in the Bayesian approach to medical diagnosis (and in teaching this to clinicians) that in actual medical practice people will miss the diagnosis of a rare disease.

Jaynes to Cohen I

First, I should stress that my topic was the principles of inference, with

medical diagnosis mentioned only to illustrate the generalized inverse situation. There was no thought of giving a serious discussion of, much less a new contribution to, medical diagnosis itself. However, if there are people in the medical profession who think – and even worse, teach their students – that Bayesian methods present us with dangers that some other method avoids, then perhaps I have a contribution to make after all.

If I understand the term "flow-chart method" correctly, this refers to the efficiency of the technical means of testing, and really concerns the design of experiments and not principles of inference. If a single test can eliminate several possibilities at once, so much the better on Bayesian or any other philosophy of inference. But I think that the kind of antiBayesian arguments that Professor Cohen quotes pose a quite different danger to medical diagnosis.

We have all heard of the proverbial medicine that cures the disease, but unfortunately turns out to have side-effects worse than the disease. We need to be on guard equally against an emotional argument that fills a psychological need of the person making it, but when translated into policy turns out to have disastrous consequences that a rational argument would have foreseen. Probably most of us have been in the following scenario.

More students than you can accept want admission to your University; and so you find yourself sitting in a committee reading grade records and letters of application and recommendation. The committee must decide today which ones we shall admit, and which ones we shall not. In every such committee there is always some person who raises the argument: "Well, now, if you admit only the promising students then there is a terrible possibility that someone who does not look good in these documents, is actually brilliant and you could be missing him." This is true, this is a risk we are taking, and if anyone can tell me how to remove that risk without incurring side-effects that are even worse, please let me know about it.

If you choose the promising applicant over the unpromising one, there is inevitably a small chance that you are doing an injustice; but with the opposite policy you are virtually certain to be doing an injustice. I have spent many hours expounding Bayesian decision theory to colleagues in such committees, and am pleased to report that those with medical training tend to appreciate this even more quickly than those with pure mathematical training.

If you test for the probable disease before the rare one, there is inevitably a small chance that you are doing the wrong thing. Testing for the rare one first does not diminish that chance but increases it. I would hold it to be a great merit of the Bayesian method that by requiring us to state our entire loss function – and not just the part of it referring to rare or emotionally charged contingencies – it foresees these side effects and protects us against them. No method can eliminate risk, but the Bayesian method minimizes it.

Cohen to Jaynes

The risk is, if you teach students the Bayesian approach to the subject, then they will adopt attitudes which will make them content with preferring the more probable to the less probable, under certain circumstances, without carrying out the further tests which might raise the probability of the rare disease's being present. The Bayesian approach does not force a search for high *weight* of evidence, as the flow-chart, or Baconian, method does.

Jaynes to Cohen II

My apologies – after that exhortation to state our full loss function, I should have followed my own advice. In my remarks, the supposed objective was not to get accurate statistics on rare diseases (although that may be a worthy objective in its own right).

I supposed implicitly, but failed to state explicitly, that our physician's objective was to help the current patient by diagnosing whatever disease he may have – rare or common – as quickly as possible. In a real situation, only a finite amount of facilities and time are available. How, then, in the light of all the information at hand, shall those facilities be best used to achieve that end?

Clearly, if our technology can test for only one disease at a time, the optimal strategy is the one that tests for the more probable diseases first. But this does not mean that the rare disease will fail to be diagnosed if the patient actually has it; at least, I would hope that medical students are taught to keep testing until a positive diagnosis is made. I cannot see how anything in Bayesian teaching would cause students to "adopt attitudes" predisposed to stop testing earlier, but would think that antiBayesian teaching might well do so, because it does not predispose one to think in terms of loss functions at all.

Of course, real medical diagnosis is vastly more complicated than these simple examples suggest, and other objectives may be called for. If our class C of possible diseases contains (a) a very common but mild one that poses no threat to life, and (b) a rare one that is fatal if not treated within 48 hours, then a prudent physician would test first for (b). But this is not an unBayesian procedure; it is a Bayesian one with a different loss function.

The issues that Professor Cohen brings up seem to be, not Bayesian vs nonBayesian principles of inference in medical diagnosis, but rather the optimal design of experiments and: "Which loss function should the physician use?" Probably no full agreement can be hoped for here, since the question raises matters of ethics as well as of medical fact.

It is, however, a theorem that a person who adopts a policy that is unBayesian with respect to his prior and loss function, is necessarily violating some very elementary desiderata of rational behavior. This has been shown by Bru-

no de Finetti, Abraham Wald, and Richard T. Cox. Therefore, while physicians should think very hard about their loss functions, it is appalling to me to think of medical students being taught to deviate from Bayesian principles in their inferences. I hope that neither Professor Cohen nor I ever comes under the care of such a physician.

Question Kuipers to Jaynes

I have a question of clarification, which could be answered by you or some of the other speakers of today. You have shown us a number of problem situations in which Bayesian statistics gives really impressive results. If I understood you correctly, you claim that classical statistics is not applicable in these types of situations. My question is whether you also claim to have an *explanation* of the apparent Bayesian successful applications or not: why does it work?

Jaynes to Kuipers

The rationale of the Bayesian approach in the problems we were just talking about is, like the principle of relativity, so simple that there is no way to explain it in terms of something still simpler. We looked at all the possibilities (the "scenes") that were compatible with our data, and asked, "In how many different ways W_1 could Nature have made scene 1; in how many ways W_2 could she have made scene 2?" and so on.

This is a question that orthodox statistics does not ask, because a scene is not a "random variable" and a multiplicity W is not a frequency in any "random experiment". But then the orthodoxian has no basis for choosing one possible scene over another, because they all have the same likelihood. And if all those possible scene also had about the same multiplicity W, then the Bayesian would be in just the same trouble. Such cases may indeed exist, but none has been found as yet.

The cases where Bayesian methods work beatifully are those in which multiplicities vary greatly from one possible scene to another, so that the Bayesian's question yields a great deal of cogent information that orthodox statistics does not recognize.

The surprisingly big effect that this has, is due to the fact that combinatorial factors (multinomial coefficients) mount up very rapidly to enormous numerical values. The multiplicity of a scene shown by Gull and Daniell was greater than that of a neighboring, very similar looking, scene by a factor of perhaps 10^{10}.

By definition, the scene of maximum entropy log W can be realized in more ways than can any other; but in fact the scenes with entropy close

to maximum can be realized in overwhelmingly more ways than can all others combined. This is the gist of the "entropy concentration theorem" given in my forthcoming collection of articles (E.T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, R.D. Rosenkrantz (ed.), D. Reidel Publishing Co., 1982).

So the explanation of the Bayesian success amounts to this: suppose there are two scenes compatible with your data. But for every way in which Nature could have created scene 1, there are 10^{10} ways (about four times the number of minutes since the Great Pyramid was built) in which she could have made scene 2. I shall, rather confidently, place my bets on scene 2 as being the right one, and nobody should be surprised to find that we have always got the right predictions from such reasoning.

This is by no means the same thing as asserting that all those ways were "equally likely". Unless Nature had for other reasons some strong predilection for scene 1 over scene 2 – and by more than a factor of 10^{10} – the result would not be different. Psychologically, if you show people the entropy numbers log W, they do not see why such a small difference should be so important. If you show them the multiplicity factors W, they see the point at once.

Question Hacking to Jaynes

Referring to your example of the photograph it seems to me that saying the number of ways that nature could have produced the photograph which you are going to reconstruct, is slightly misleading and it is just not nature, it is nature and a piece of apparatus about which you have a substantial body of physical knowledge. You know roughly speaking how it works. Now it seems to me that it is so often the case that a piece of Bayesian reasoning can also be reconstructed in an alternative way. Could you not tell a very different story of what is going on here, that you really do have approximate physical hypotheses about the relative frequency of the ways in which the blurred scene would have been produced; relative frequencies which are based on your views of how the photographic apparatus works in the first place?

Jaynes to Hacking

The knowledge about how the apparatus works, generating the relative frequency with which different blurred scenes would be produced, is what the orthodox statistician would call the "sampling distribution". That information is available to and used by both the Bayesian and the orthodoxian.

The extra information used by the Bayesian alone, is that multiplicity factor W, which is not the number of ways in which Nature could have made the known blurred scene, but the number of ways she could have made the

various possible true but unknown scenes. These multiplicity factors determine the enormous variations in prior probabilities of different scenes, on which the success of the Bayesian reconstructions depend. At the present time, orthodox ideology does not recognize the relevance of this information at all.

I do, however, agree most emphatically with your general remark that any given result can be rationalized in various ways. If you give twenty students the answer to their homework problem in advance, their hand-ins will all arrive at exactly that answer, and offer twenty different reasons for it.

Likewise, now that the Bayesian – Maximum Entropy arguments have shown us the right answers to the image reconstruction problem, I am confident that antiBayesians will find ways in which they can rationalize using that multiplicity information after all, and thus show how they could have got the same results.