CHAPTER 3

46656 Varieties of Bayesians (#765)

Some attacks and defenses of the Bayesian position assume that it is unique so it should be helpful to point out that there are at least 46656 different interpretations. This is shown by the following classification based on eleven facets. The count would be larger if I had not artificially made some of the facets discrete and my heading would have been "On the Infinite Variety of Bayesians."

All Bayesians, as I understand the term, believe that it is usually meaningful to talk about the probability of a hypothesis and they make some attempt to be consistent in their judgments. Thus von Mises (1942) would not count as a Bayesian, on this definition. For he considered that Bayes's theorem is applicable only when the prior is itself a physical probability distribution based on a large sample from a superpopulation. If he is counted as a Bayesian, then there are at least 46657 varieties, which happens to rhyme with the number of Heinz varieties. But no doubt both numbers will increase on a recount.

Here are the eleven facets:

1. *Type II rationality.* (a) Consciously recognized; (b) not. Here Type II rationality is defined as the recommendation to maximize expected utility allowing for the cost of theorizing (#290). It involves the recognition that judgments can be revised, leading at best to consistency of *mature* judgments.

2. *Kinds of judgments.* (a) Restricted to a specific class or classes, such as preferences between actions; (b) all kinds permitted, such as of probabilities and utilities, and any functions of them such as expected utilities, weights of evidence, likelihoods, and surprise indices (#82; Good, 1954). This facet could of course be broken up into a large number.

3. *Precision of judgments.* (a) Sharp; (b) based on inequalities, i.e. partially ordered, but sharp judgments often assumed for the sake of simplicity (in accordance with 1[a]).

4. *Extremeness.* (a) Formal Bayesian procedure recommended for all applications; (b) non-Bayesian methods used provided that some set of axioms of intuitive probability are not seen to be contradicted (the Bayes/non-Bayes compromise: Hegel and Marx would call it a synthesis); (c) non-Bayesian methods used only after they have been given a rough Bayesian justification.

5. *Utilities.* (a) Brought in from the start; (b) avoided, as by H. Jeffreys; (c) utilities introduced separately from intuitive probabilities.

6. *Quasiutilities.* (a) Only one kind of utility recognized; (b) explicit recognition that "quasiutilities" (##690A, 755) are worth using, such as amounts of information or "weights of evidence" (Peirce, 1978 [but see #1382]; #13); (c) using quasiutilities without noticing that they are substitutes for utilities. The *use* of quasiutilities is as old as the words "information" and "evidence," but I think the name "quasiutility" serves a useful purpose in focussing the issue.

7. *Physical probabilities.* (a) Assumed to exist; (b) denied; (c) used as if they exist but without philosophical commitment (#617).

8. *Intuitive probability.* (a) Subjective probabilities regarded as primary; (b) credibilities (logical probabilities) primary; (c) regarding it as mentally healthy to think of subjective probabilities as estimates of credibilities, without being sure that credibilities really exist; (d) credibilities in principle definable by an international body. . . .

9. *Device of imaginary results.* (a) Explicit use; (b) not. The device involves imaginary experimental results used for judging final or posterior probabilities from which are inferred discernments about the initial probabilities. For examples see ##13, 547.

10. *Axioms.* (a) As simple as possible; (b) incorporating Kolmogorov's axiom (complete additivity); (c) using Kolmogorov's axiom when mathematically convenient but regarding it as barely relevant to the *philosophy* of applied statistics.

11. *Probability "types."* (a) Considering that priors can have parameters with "Type III" distributions, as a convenient technique for making judgments; (b) not. Here (a) leads, by a compromise with non-Bayesian statistics, to such techniques as Type II maximum likelihood and Type II likelihood-ratio tests (#547).

Thus there are at least $2^4 \cdot 3^6 \cdot 4 = 46656$ categories. This is more than the number of professional statisticians so some of the categories must be empty. Thomas Bayes hardly wrote enough to be properly categorized; a partial attempt is b--aaa?-b--. My own category is abcbcbccaca. What's yours?

CHAPTER 4

The Bayesian Influence, or How to Sweep Subjectivism under the Carpet (#838)

ABSTRACT

On several previous occasions I have argued the need for a Bayes/non-Bayes compromise which I regard as an application of the "Type II" principle of rationality. By this is meant the maximization of expected utility when the labour and costs of the calculations are taken into account. Building on this theme, the present work indicates how some apparently objective statistical techniques emerge logically from subjective soil, and can be further improved if their subjective logical origins (if not always historical origins) are not ignored. There should in my opinion be a constant interplay between the subjective and objective points of view, and not a polarization separating them.

Among the topics discussed are, two types of rationality, 27 "Priggish Principles," 46656 varieties of Bayesians, the Black Box theory, consistency, the unobviousness of the obvious, probabilities of events that have never occurred (namely all events), the Device of Imaginary Results, graphing the likelihoods, the hierarchy of types of probability, Type II maximum likelihood and likelihood ratio, the statistician's utilities versus the client's, the experimenter's intentions, quasiutilities, tail-area probabilities, what is "more extreme"?, "deciding in advance," the harmonic mean rule of thumb for significance tests in parallel, density estimation and roughness penalities, evolving probability and pseudorandom numbers and a connection with statistical mechanics.

1. PREFACE

. . . There is one respect in which the title of this paper is deliberately ambiguous: it is not clear whether it refers to the *historical* or to the *logical* influence of "Bayesian" arguments. In fact it refers to both, but with more emphasis on the logical influence. Logical aspects are more fundamental to a science or philosophy than are the historical ones, although they each shed light on the other. The logical development is a candidate for being the historical development on *another* planet.

I have taken the expression the "Bayesian influence" from a series of lectures in mimeographed form (#750). In a way I am fighting a battle that has already been won to a large extent. For example, the excellent statisticians L. J. Savage, D. V. Lindley, G. E. P. Box (R. A. Fisher's son-in-law) and J. Cornfield were converted to the Bayesian fold years ago. For some years after World War II, I stood almost alone at meetings of the Royal Statistical Society in crusading for a Bayesian point of view. Many of the discussions are reported in the *Journal, series B*, but the most detailed and sometimes heated ones were held privately after the formal meetings in dinners at Berterolli's restaurant and elsewhere, especially with Anscombe, Barnard, Bartlett, Daniels, and Lindley. [Lindley was a non-Bayesian until 1954.] These protracted discussions were historically important but have never been mentioned in print before as far as I know. There is an unjustifiable convention in the writing of the history of science that science communication occurs only through the printed word. . . .

II. INTRODUCTION

On many previous occasions, and especially at the Waterloo conference of 1970, I have argued the desirability of a Bayes/non-Bayes compromise which, from one Bayesian point of view, can be regarded as the use of a "Type II" principle of rationality. By this is meant the maximization of expected utility when the labour and costs of calculations and thinking are taken into account. Building on this theme, the present paper will indicate how some apparently objective statistical techniques emerge logically from subjective soil, and can be further improved by taking into account their logical, if not always historical, subjective origins. There should be in my opinion a constant interplay between the subjective and objective points of view and not a polarization separating them.

Sometimes an orthodox statistician will say of one of his techniques that it has "intuitive appeal." This is I believe always a guarded way of saying that it has an informal approximate Bayesian justification.

Partly as a matter of faith, I believe that *all* sensible statistical procedures can be derived as approximations to Bayesian procedures. As I have said on previous occasions, "To the Bayesian all things are Bayesian."

Cookbook statisticians, taught by non-Bayesians, sometimes give the impression to *their* students that cookbooks are enough for all practical purposes. Any one who has been concerned with complex data analysis knows that they are wrong: that subjective judgment of probabilities cannot usually be avoided, even if this judgment can later be used for constructing apparently non-Bayesian procedures in the approved sweeping-under-the-carpet manner.

(a) What Is Swept under the Carpet?

I shall refer to "sweeping under the carpet" several times, so I shall use the abbreviations UTC and SUTC. One part of this paper deals with what is swept

under the carpet, and another part contains some examples of the SUTC process. (The past tense, etc., will be covered by the same abbreviation.)

Let us then consider what it is that is swept under the carpet. Maurice Bartlett once remarked, in a discussion at a Research Section meeting of the Royal Statistical Society, that the word "Bayesian" is ambiguous, that there are many varieties of Bayesians, and he mentioned for example, "Savage Bayesians and Good Bayesians," and in a letter in the *American Statistician* I classified 46656 varieties (#765). There are perhaps not that number of practicing Bayesian statisticians, but the number comes to 46656 when your cross-classify the Bayesians in a specific manner by eleven facets. Some of the categories are perhaps logically empty but the point I was making was that there is a large variety of possible interpretations and some of the arguments that one hears against *the* Bayesian position are valid only against *some* Bayesian positions. As so often in controversies "it depends what you mean." The particular form of Bayesian position that I adopt might be called non-Bayesian by some people and naturally it is my own views that I would like most to discuss. I speak for some of the Bayesians all the time and for all the Bayesians some of the time. In the spoken version of this paper I named my position after "the Tibetan Lama K. Caj Doog," and I called my position "Doogian." Although the joke wears thin, it is convenient to have a name for this viewpoint, but "Bayesian" is misleading, and "Goodian" or "Good" is absurd, so I shall continue with the joke even in print. (See also Smith, 1961, p. 18, line minus 15, word minus 2.)

Doogianism is mainly a mixture of the views of a few of my eminent pre-1940 predecessors. Many parts of it are therefore not original, but, taken as a whole I think it has some originality; and at any rate it is convenient here to have a name for it. It is intended to be a general philosophy for reasoning and for rationality in action and not just for statistics. It is a philosophy that applies to all activity, to statistics, to economics, to the practice and philosophy of science, to ordinary behavior, and, for example, to chess-playing. Of course each of these fields of study or activity has its own specialized problems, but, just as the theories of each of them should be consistent with ordinary logic, they should in my opinion be consistent also with the theory of rationality as presented here and in my previous publications, a theory that is a useful and practically necessary extension of ordinary logic. . . .

At the Waterloo conference (#679), I listed 27 Priggish Principles that summarize the Doogian philosophy, and perhaps the reader will consult the Proceedings and some of its bibliography for a more complete picture, and for historical information. Here it would take too long to work systematically through all 27 principles and instead I shall concentrate on the eleven facets of the Bayesian Varieties in the hope that this will give a fairly clear picture. I do not claim that any of these principles were "discovered last week" (to quote Oscar Kempthorne's off-the-cuff contribution to the spoken discussion), in fact I have developed, acquired or published them over a period of decades, and most of them were used by others before 1940, in one form or another, and with various degrees of bakedness or emphasis. The main merit that I claim for the Doogian

philosophy is that it codifies and exemplifies an adequately complete and simple theory of rationality, complete in the sense that it is I believe not subject to the criticisms that are usually directed at other forms of Bayesianism, and simple in the sense that it attains realism with the minimum of machinery. To pun somewhat, it is "minimal sufficient."

(b) Rationality, Probability, and the Black Box Theory

In some philosophies of rationality, a rational man is defined as one whose judgments of probabilities, utilities, and of functions of these, are all both consistent and sharp or precise. Rational men do not exist, but the concept is useful in the same way as the concept of a reasonable man in legal theory. A rational man can be regarded as an ideal to hold in mind when we ourselves wish to be rational. It is sometimes objected that rationality as defined here depends on betting behavior, and people sometimes claim they do not bet. But since their every decision is a bet I regard this objection as unsound: besides they could in principle be forced to bet in the usual monetary sense. It seems absurd to me to suppose that the *rational* judgment of probabilities would normally depend on whether you were forced to bet rather than betting by free choice.

There are of course people who argue (rationally?) against rationality, but presumably they would agree that it is sometimes desirable. For example, they would usually prefer that their doctor should make rational decisions, and, when they were fighting a legal case in which they were sure that the evidence "proved" their case, they would presumably want the judge to be rational. I believe that the dislike of rationality is often merely a dishonest way of maintaining an indefensible position. Irrationality is intellectual violence against which the pacifism of rationality may or may not be an adequate weapon.

In practice one's judgments are not sharp, so that to use the most familiar axioms it is necessary to work with judgments of inequalities. For example, these might be judgments of inequalities between probabilities, between utilities, expected utilities, weights of evidence (in a sense to be defined . . .), or any other convenient function of probabilities and utilities. We thus arrive at a theory that can be regarded as a combination of the theories espoused by F. P. Ramsey (1926/31/50/64), who produced a theory of precise subjective probability and utility, and of J. M. Keynes (1921), who emphasized the importance of inequalities (partial ordering) but regarded logical probability or credibility as the fundamental concept, at least until he wrote his obituary on Ramsey (Keynes, 1933).

To summarize then, the theory I have adopted since about 1938 is a theory of subjective (personal) probability and utility in which the judgments take the form of inequalities (but see Section III [iii] below). This theory can be formulated as the following "black box" theory. . . .[See pp. 75-76.]

To extend this theory to rationality, we need merely to allow judgments of preferences also, and to append the "principle of rationality," the recommendation to maximize expected utility. (##13, **26, 230**.)

The axioms, which are expressed in conditional form, have a familiar appearance (but see the reference to "evolving probability" below), and I shall not state them here.

There is emphasis on human judgment in this theory, based on a respect for the human brain. Even infrahuman brains are remarkable instruments that cannot yet be replaced by machines, and it seems unlikely to me that decision-making in general, and statistics in particular, can become independent of the human brain until the first ultraintelligent machine is built. Harold Jeffreys once remarked that the brain may not be a perfect reasoning machine but is the only one available. That is still true, though a little less so than when Jeffreys first said it. It [the brain] has been operating for millions of years in billions of individuals and it has developed a certain amount of magic. On the other hand I believe that some formalizing is useful and that the ultraintelligent machine will also use a subjectivistic theory as an aid to its reasoning.

So there is this respect for judgment in the theory. But there is also respect for logic. Judgments and logic must be combined, because, although the human brain is clever at perceiving facts, it is also cunning in the rationalization of falsity for the sake of its equilibrium. You *can* make bad judgments so you need a black box to check your subjectivism and to make it more objective. That then is the purpose of a subjective theory; to increase the objectivity of your judgments, to check them for consistency, to detect the inconsistencies and to remove them. Those who want their subjective judgments to be free and untrammeled by axioms regard themselves as objectivists: paradoxically, it is the subjectivists who are prepared to discipline their own judgments!

For a long time I have regarded this theory as almost intuitively obvious, partly perhaps because I have used it in many applications without inconsistencies arising, and I know of no other theory that I could personally adopt. It is the one and only True Religion. My main interest has been in developing and applying the theory, rather than finding *a priori* justification for it. But such a justification has been found by C. A. B. Smith (1961), based on a justification by Ramsey (1926/31/50/64), de Finetti (1937/64), and L. J. Savage (1954) of a slightly different theory (in which sharp values of the probabilities and utilities are assumed). These justifications show that, on the assumption of certain compelling desiderata, a rational person will hold beliefs, and preferences, *as if* he had a system of subjective probabilities and utilities satisfying a familiar set of axioms. He might as well be explicit about it: after all it is doubtful whether any of our knowledge is better than of the "as if" variety.

Another class of justifications, in which utilities are not mentioned, is exemplified by Bernstein (1921/22), Koopman (1940a, b), and R. T. Cox (1946, 1961). (See also pp. 105-106 of #13; also Good, 1962d.) A less convincing, but simpler justification is that the product and addition axioms are forced (up to a monotonic transformation of probabilities) when considering ideal games of chance, and it would be *surprising* if the same axioms did not apply more generally. Even to deny this seems to me to show poor or biased judgment.

Since the degrees of belief, concerning events over which he has no control, of a person with ideally good judgment, should surely not depend on whether he intends to use his beliefs in any specific manner, it seems desirable to have justifications that do not mention preferences or utilities. But utilities necessarily come in whenever the beliefs are to be used in a practical problem involving action.

(c) Consistency and the Unobviousness of the Obvious

Everybody knows, all scientists know, all mathematicians know, all players of chess know, that from a small number of sharp axioms you can develop a very rich theory in which the results are by no means obvious, even though they are, in a technical sense, tautological. This is an exciting feature of mathematics, especially since it suggests that it might be a feature of the universe itself. Thus the completion of the basic axioms of probability by Fermat and Pascal led to many interesting results, and the further near completion of the axioms for the *mathematical* theory of probability by Kolmogorov led to an even greater expansion of mathematical theory.

Mathematicians are I think often somewhat less aware that a system of rules and suggestions of an axiomatic system can also stimulate useful technical advances. The effect can be not necessarily nor even primarily to produce theorems of great logical depth, but sometimes more important to produce or to emphasize attitudes and techniques for reasoning and for statistics that seem obvious enough after they are suggested, but continue to be overlooked even then.

One reason why many theoretical statisticians prefer to prove mathematical theorems rather than to emphasize logical issues is that a theorem has a better chance of being indisputably novel. The person who proves the theorem can claim all the credit. But with general principles, however important, it is usually possible to find something in the past that to some extent foreshadows it. There is usually something old under the sun. Natural Selection was stated by Aristotle, but Darwin is not denied credit, for most scientists were still overlooking this almost obvious principle.

Let me mention a personal example of how the obvious can be overlooked.

In 1953, I was interested in estimating the physical probabilities for large contingency tables (using essentially a log-linear model) when the entries were very small, including zero. In the first draft of the write-up I wrote that I was concerned with the estimation of *probabilities of events that had never occurred before* (#83). Apparently this *concept* was itself an example of such an event, as far as the referee was concerned because he felt it was too provocative, and I deleted it in deference to him. Yet this apparently "pioneering" remark is obvious: every event in life is unique, and every real-life probability that we estimate in practice is that of an event that has never occurred before, provided that we do enough cross-classification. Yet there are many "frequentists" who still sweep this fact UTC.

A statistical problem where this point arises all the time is in the estimation of

physical probabilities corresponding to the cells of multidimensional contingency tables. Many cells will be empty for say a 2^{20} table. A Bayesian proposal for this problem was made in Good (p. #75 of #398), and I am hoping to get a student to look into it; and to compare it with the use of log-linear models which have been applied to this problem during the last few years. One example of the use of a log-linear model is, after taking logarithms of the relative frequencies, to apply a method of smoothing mentioned in #146 in relation to factorial experiments: namely to treat non-significant interactions as zero (or of course they could be "flattened" Bayesianwise instead for slightly greater accuracy).

Yet another problem where the probabilities of events that have never occurred before are of interest is the species sampling problem. One of its aspects is the estimation of the probability that the next animal or word sampled will be one that has not previously occurred. The answer turns out to be approximately equal to n_1/N, where n_1 is the number of species that have so far occurred just once, and N is the total sample size: see ##38 & 86; this work was originated with an idea of Turing's (1940) which anticipated the empirical Bayes method in a special case. (See also Robbins, 1968.) The method can be regarded as non-Bayesian but with a Bayesian influence underlying it. More generally, the probability that the next animal will be one that has so far been represented r times is approximately $(r+1)n_{r+1}/N$, where n_r is the "frequency of the frequency r," that is, the number of species each of which has already been represented r times. (In practice it is necessary to smooth the n_r's when applying this formula, to get adequate results, when $r>1$.) I shall here give a new proof of this result. Denote the event of obtaining such an animal by E_r. Since the order in which the N animals were sampled is assumed to be irrelevant (a Bayesian-type assumption of permutability), the required probability can be estimated by the probability that E_r would have occurred on the last occasion an animal was sampled if a random permutation were applied to the order in which the N animals were sampled. But E_r would have occurred if the last animal had belonged to a species represented $r+1$ times *altogether*. This gives the result, except that for greater accuracy we should remember that we are talking about the $(N+1)$st trial, so that a more accurate result is $(r+1)\mathfrak{E}_{N+1}(n_{r+1})/(N+1)$. Hence the expected physical probability q_r corresponding to those n_r species that have so far occured r times is

$$\mathfrak{E}(q_r) = \frac{r+1}{N+1}\cdot\frac{\mathfrak{E}_{N+1}(n_{r+1})}{\mathfrak{E}_N(n_r)}.$$

This is formula (15) of #38 which was obtained by a more Bayesian argument. The "variance" of q_r was also derived in that paper, and a "frequency" proof of it would be more difficult. There is an interplay here between Bayesian and frequency ideas.

One aspect of Doogianism which dates back at least to F. P. Ramsey (1926/31/50/64) is the emphasis on *consistency*: for example, the axioms of probability can provide only *relationships* between probabilities and cannot manufacture a

Fisher's fiducial argument. (This assumption is pinpointed in #659 on its p. 139 omitted herein. The *reason* Fisher overlooked this is also explained there.)

The idea of consistency seems weak enough, but it has the following immediate consequence which *is* often overlooked.

Owing to the adjectives "initial" and "final" or "prior" and "posterior," it is usually assumed that initial probabilities must be assumed before final ones can be calculated. But there is nothing in the theory to prevent the implication being in the reverse direction: we can make judgments of initial probabilities and infer final ones, or we can equally make judgments of final ones and infer initial ones by *Bayes's theorem in reverse*. Moreover this can be done corresponding to entirely *imaginary* observations. This is what I mean by the Device of Imaginary Results for the judging of initial probabilities. (See, for example, Index of #13). I found this device extremely useful in connection with the choice of a prior for multinomial estimation and significance problems (#547) and I believe the device will be found to be of the utmost value in future Bayesian statistics. Hypothetical experiments have been familiar for a long time in physics, and in the arguments that led Ramsey to the axioms of subjective probability, but the use of Bayes's theorem in reverse is less familiar. "Ye priors shall be known by their posteriors" (p. 17). Even the slightly more obvious technique of imaginary bets is still disdained by many decision makers who like to say "That possibility is purely hypothetical." Anyone who disdains the hypothetical is a philistine.

III. THE ELEVENFOLD PATH OF DOOGIANISM

As I said before, I should now like to take up the 46656 varieties of Bayesians, in other words the eleven facets for their categorization. I would have discussed the 27-fold path of Doogianism if there had been space enough.

(i) Rationality of Types I and II

I have already referred to the first facet. Rationality of Type I is the recommendation to maximize expected utility, and Type II is the same except that it allows for the cost of theorizing. It means that in any practical situation you have to decide when to stop thinking. You can't allow the current to go on circulating round and round the black box or the cranium forever. You would like to reach a sufficient maturity of judgments, but you have eventually to reach some conclusion or to make some decision and so you must be prepared to sacrifice strict logical consistency. At best you can achieve consistency as far as you have seen to date (p. 49 of #13). There is a time element, as in chess, and this is realistic of most practice. It might not appeal to some of you who love ordinary logic, but it is a mirror of the true situation.

It may help to convince some readers if I recall a remark of Poincaré's that some antinomies in ordinary (non-probabilistic) logic can be resolved by bringing in a time element. ["Temporal," "evolving" or "dynamic" logic?]

controversies between the orthodox and Bayesian points of view, also involves a shifting of your probabilities. The subjective probabilities shift as a consequence of thinking. . . . [See p. 107.] The conscious recognition of Type II rationality, or not, constitutes the two aspects of the first facet.

Another name for the principle of Type II rationality might be the *Principle of Non-dogmatism.*

(ii) Kinds of Judgment

Inequalities between probabilities and between expected utilities are perhaps the most standard type of judgment, but other kinds are possible. Because of my respect for the human mind, I believe that one should allow any kind of judgments that are relevant. One kind that I believe will ultimately be regarded as vying in importance with the two just mentioned is a judgment of "weights of evidence" (defined later) a term introduced by Charles Sanders Peirce (1878) although I did not know this when I wrote my 1950 book. . . .

It will encourage a revival of reasoning if statisticians adopt this appealing terminology . . . . [But Peirce blew it. See #1382.]

One implication of the "suggestion" that all types of judgments can be used is to encourage you to compare your "overall" judgments with your detailed ones; for example, a judgment by a doctor that it is better to operate than to apply medical treatment, on the grounds perhaps that this would be standard practice in the given circumstances, can be "played off" against separate judgments of the probabilities and utilities of the outcomes of the various treatments.

(iii) Precision of Judgments

Most theories of subjective probability deal with numerically precise probabilities. These would be entirely appropriate if you could always state the lowest odds that you would be prepared to accept in a gamble, but in practice there is usually a degree of vagueness. Hence I assume that subjective probabilities are only partially ordered. In this I follow Keynes and Koopman, for example, except that Keynes dealt primarily with logical probabilities, and Koopman with "intuitive" ones (which means either logical or subjective). F. P. Ramsey (1926/31/50/64) dealt with subjective probabilities, but "sharp" ones, as mentioned before.

A theory of "partial ordering" (inequality judgments) for probabilities is a compromise between Bayesian and non-Bayesian ideas. For if a probability is judged merely to lie between 0 and 1, this is equivalent to making no judgment about it at all. The vaguer the probabilities the closer is this Bayesian viewpoint to a non-Bayesian one.

Often, in the interests of simplicity, I assume sharp probabilities, as an approximation, in accordance with Type II rationality.

(iv) Eclecticism

Many Bayesians take the extreme point of view that Bayesian methods should always be used in statistics. My view is that non-Bayesian methods are acceptable *provided that they are not seen to contradict your honest judgments, when combined with the axioms of rationality.* This facet number (iv) is an application of Type II rationality. I believe it is sometimes, but not by any means always, easier to use "orthodox" (non-Bayesian) methods, and that they are often *good enough.* It is always an application of Type II rationality to say that a method is good enough.

(v) Should Utilities Be Brought in from the Start in the Development of the Theory?

I have already stated my preference for trying to build up the theory of subjective probability without reference to utilities and to bring in utilities later. The way the axioms are introduced is not of great practical importance, provided that the same axioms are reached in the end, but it is of philosophical interest. Also there is practical interest in seeing how far one can go without making use of utilities, because one might wish to be an "armchair philosopher" or "fun scientist" who is more concerned with discovering facts about Nature than in applying them. ("Fun scientist" is not intended to be a derogatory expression.) Thus, for example, R. A. Fisher and Harold Jeffreys never used ordinary utilities in their statistical work as far as I know (and when Jeffreys chaired the meeting in Cambridge when I presented my paper **#26** he stated that he had never been concerned with economic problems in his work on probability). See also the following remarks concerned with quasiutilities.

(vi) Quasiutilities

Just as some schools of Bayesians regard subjective probabilities as having sharp (precise) values, some assume that utilities are also sharp. The Doogian believes that this is often not so. It is not merely that utility inequality judgments of course vary from one person to another, but that utilities for individuals can also often be judged by them only to lie in wide intervals. It consequently becomes useful and convenient to make use of substitutes for utility which may be called *quasiutilities* or *pseudoutilities.* Examples and applications of quasiutilities will be considered later in this paper. The conscious recognition or otherwise of quasiutilities constitutes the sixth facet.

(vii) Physical Probability

Different Bayesians have different attitudes to the question of physical probability. de Finetti regards it as a concept that can be defined in terms of subjective probability, and does not attribute any other "real existence" to it. My view, or that of my alter ego, is that it seems reasonable to suppose that

physical probabilities do exist, but that they can be measured only be means of a theory of subjective probability. For a fuller discussion of this point see de Finetti (1968/70) and #617. The question of the real existence of physical probabilities relates to the problem of determinism versus indeterminism and I shall have something more to say on this.

When I refer to physical probability I do not assume the long-run frequency definition: physical probability can be applied just as well to unique circumstances. Popper suggested the word "propensity" for it, which I think is a good term, although I think the suggestion of a word cannot by itself be regarded as the propounding of a "theory." [See also p. 405 of Feibleman, 1969.] As I have indicated before, I think good terminology is important in crystallizing out ideas. Language can easily mislead, but part of the philosopher's job is to find out where it can *lead*. Curiously enough Popper has also stated that the words you use do not matter much: what is important is what they mean in your context. Fair enough, but it can lead to Humpty-Dumpty-ism, such as Popper's interpretation of simplicity [or Carnap's usage of "confirmation" which has misled philosophers for decades].

(viii) Which is Primary, Logical Probability (Credibility) or Subjective Probability?

It seems to me that subjective probabilities are primary because they are the ones you have to use whether you like it or not. But I think it is mentally healthy to think of your subjective probabilities as estimates of credibilities, whether these really "exist" or not. Harold Jeffreys said that the credibilities should be laid down by an international body. He would undoubtedly be the chairman. As Henry Daniels once said (c. 1952) when I was arguing for subjectivism, "all statisticians would like their models to be adopted," meaning that in some sense everybody is a subjectivist.

(ix) Imaginary Results

This matter has already been discussed but I am mentioning it again because it distinguishes between some Bayesians in practice, and so forms part of the categorization under discussion. I shall give an example of it now because this will help to shed light on the tenth facet.

It is necessary to introduce some notation. Let us suppose that we throw a sample of N things into t pigeon holes, with statistically independent physical probabilities $p_1, p_2, \ldots, p_t$, these being unknown, and that you obtain frequencies $n_1, n_2, \ldots, n_t$ in the t categories or cells. This is a situation that has much interested philosophers of induction, but for some reason, presumably lack of familiarity, they do not usually call it multinomial sampling. In common with many people in the past, I was interested (##398, 547) in estimating the physical probabilities $p_1, p_2, \ldots, p_t, \ldots$. [See pp. 100-103.]

That then is an example of a philosophical attitude leading to a practical solution of a statistical problem. As a matter of fact, it wasn't just the estimation

of the p's that emerged from that work, but, more important, a significance test for whether the p's were all equal. The method has the pragmatic advantage that it can be used for all sample sizes, whereas the ordinary chi-squared test breaks down when the cell averages are less then 1. Once you have decided on a prior (the initial relative probabilities of the components of the non-null hypothesis), you can calculate the weight of evidence against the null hypothesis without using asymptotic theory. (This would be true for any prior that is a linear combination of Dirichlet distributions, even if they were not symmetric, because in this case the calculations involve only one-dimensional integrations.) That then was an example of the device of imaginary results, for the selection of a prior, worked out in detail.

The successful use of the device of imaginary results for this problem *makes it obvious that it can and will also be used effectively for many other statistical problems. I believe it will revolutionize multivariate Bayesian statistics.*

(x) Hierarchies of Probabilities

When you make a judgment about probabilities you might sit back and say "Is that judgment probable." This is how the mind works—it is natural to think that way, and this leads to a hierarchy of types of probabilities (#26) which in the example just mentioned, I found useful, as well as on other occasions. Now an objection immediately arises: There is nothing in principle to stop you integrating out the higher types of probability. But it remains a useful suggestion to help the mind in making judgments. It was used in #547 and has now been adopted by other Bayesians, using different terminology, such as priors of the second "order" (instead of "type" or "two-stage Bayesian models." A convenient term for a parameter in a prior is "hyperparameter." [See also #1230.]

New techniques arose out of the hierarchical suggestion, again apparently first in connection with the multinomial distribution (in the same paper), namely the concept of Type II maximum likelihood (maximization of the Bayes factor against the null hypothesis by allowing the hyperparameters to vary), and that of a Type II likelihood ratio for significance tests. I shall discuss these two concepts when discussing likelihood in general.

(xi) The Choice of Axioms

One distinction between different kinds of Bayesians is merely a mathematical one, whether the axioms should be taken as simple as possible, or whether, for example, they should include Kolmogorov's axiom, the axiom of complete additivity. I prefer the former course because I would want people to use the axioms even if they do not know what "enumerable" means, but I am prepared to use Kolmogorov's axiom whenever it seems to be sufficiently mathematically convenient. Its interest is mathematical rather than philosophical, except perhaps for the philosophy of mathematics. This last facet by the way is related to an excellent lecture by Jimmie Savage of about 1970, called "What kind of probability do you want?"

So much for the eleven facets. Numbers (i) to (vii) and number (ix) all involve a compromise with non-Bayesian methods; and number (xiii) a compromise with the "credibilists."

IV. EXAMPLES OF THE BAYESIAN INFLUENCE AND OF SUTC

(a) The Logical and Historical Origins of Likelihood

One aspect of utility is communicating with other people. There are many situations where you are interested in making a decision without communicating. But there are also many situations, especially in much statistical and scientific practice where you do wish to communicate. One suggestion, "obvious," and often overlooked as usual, is that you should make your assumptions clear and you should try to separate out the part that is disputable from the part that is less so. One immediate consequence of this suggestion is an emphasis on likelihood, because, as you all know, in Bayes's theorem you have the initial probabilities, and then you have the likelihoods which are the probabilities of the event, given the various hypotheses, and then you multiply the likelihoods by the probabilities and that gives you results proportional to the final probabilities. That is Bayes's theorem expressed neatly, the way Harold Jeffreys (1939/61) expressed it. Now the initial probability of the null hypothesis is often highly disputable. One person might judge it to be between 10^{-3} and 10^{-1} whereas another might judge it to be between 0.9 and 0.99. There is much less dispute about likelihoods. There is no dispute about the numerical values of likelihoods if your basic parametric model is accepted. Of course you usually have to use subjective judgment in laying down your parametric model. Now the *hidebound* objectivist tends to hide that fact; he will not volunteer the information that he uses judgment at all, but if pressed he will say "I do, in fact, have good judgment." So there are good and bad subjectivists, the bad subjectivists are the people with bad or dishonest judgment and also the people who do not make their assumptions clear when communicating with other people. But, on the other hand, there are no good 100% (hidebound) objectivists; they are all bad because they sweep their judgments UTC.

> *Aside*: In the spoken discussion the following beautiful interchanges took place. *Kempthorne* (who also made some complimentary comments): Now, on the likelihood business, the Bayesians discovered likelihood Goddamit! Fisher knew all this stuff. Now look Jack, you are an educated guy. Now please don't pull this stuff. This really drives me up the wall! *Lindley*: If Fisher understood the likelihood principle why did he violate it? *Kempthorne*: I'm not saying he understood it and I'm not saying you do or you—nobody understands it. But likelihood ideas, so to speak, have some relevance to the data. That's a completely non-Bayesian argument. *Good*: It dates back to the 18th century. *Kempthorne*: Oh it dates back; but there are a lot of things being (?) Doogian, you know. They started with this guy Doog. Who is this bugger? Doog is the guy who spells everything backwards.

In reply to this entertaining harangue, which was provoked by a misunderstanding that was perhaps my fault, although I did refer to Fisherian information, I mention the following points. Bayes's theorem (Bayes, 1763/65, 1940/58; Laplace, 1774) cannot be stated without introducing likelihoods; *therefore likelihood dates back at least to 1774.* Again, *maximum* likelihood was used by Daniel Bernoulli (1774/78/1961); see, for example, Todhunter (1865, p. 236) or Eisenhart (1964, p. 29). Fisher introduced the name *likelihood* and emphasized the method of maximum likelihood. Such emphasis is important and of course merits recognition. The fact that he was anticipated in its use does not deprive him of the major part of the credit or of the blame especially as the notion of defining [his kind of] amount of information in terms of likelihood was his brilliant idea and it led to the Aitken-Silverstone information inequality (the minimum-variance bound). [Perhaps *not* due to Aitken and Silverstone.]

Gauss (1798/1809/57/1963) according to Eisenhart, used inverse probability combined with a Bayes *postulate* (uniform initial distribution) and an assumption of normal error, to give one of the interpretations of the method of least squares. He could have used maximum likelihood in this context but apparently did not, so perhaps Daniel Bernoulli's use of maximum likelihood had failed to convince him or to be noticed by him. Further historical research might be required to settle this last question if it is possible to settle it at all.

So likelihood is important as all statisticians agree now-a-days, and it *takes sharper values* than initial probabilities. But some people have gone to extremes and say that initial probabilities don't mean anything. Now I think one reason for their saying so is trade unionism of a certain kind. It is very nice for a statistician to be able to give his customer absolutely clear-cut results. It is unfortunate that he can't do it so he is tempted to cover up, to pretend he has not had to use any judgment. Those Bayesians who *insist* on sharp initial probabilities are I think also guilty of "trade unionism," unless they are careful to point out that these are intended only as crude approximations, for I do not believe that sharp initial probabilities usually correspond to their honest introspection. If, on the other hand, they agree that they are using only approximations we might need more information about the degree of the approximations, and then they would be forced to use inequality judgments, thus bringing them closer to the True Religion. (I believe Dr. Kyburg's dislike of the Bayesian position, as expressed by him later in this conference, depended on his interpreting a Bayesian as one who uses sharp initial probabilities.) The use of "vague" initial probabilities (inequality judgments) does not prevent Bayes's theorem from establishing the likelihood principle. For Dr. Kempthorne's benefit, and perhaps for some others, I mention that to me the likelihood principle means that the likelihood function exhausts all the information about the parameters that can be obtained from an experiment or observation, provided of course that there is an undisputed set of exhaustive simple statistical hypotheses such as is provided, for example, by a

parametric model. (In practice, such assumptions are often undisputed but are never indisputable. This is the main reason why significance tests, such as the chi-squared test, robust to changes in the model, are of value. Even here there is a Doogian interpretation that can be based on beliefs about the distribution of the test statistic when it is assumed that the null hypothesis is false. I leave this point on one side for the moment.) Given the likelihood, the inferences that can be drawn from the observations would, for example, be unaffected if the statistician arbitrarily and falsely calimed that he had a train to catch, although he really had decided to stop sampling because his favorite hypothesis was ahead of the game. (This might cause you to distrust the statistician, but if you believe his observations, this distrust would be immaterial.) On the other hand, the "Fisherian" tail-area method for significance testing violates the likelihood principle because the statistician who is prepared to pretend he has a train to catch (optional stopping of sampling) can reach arbitrarily high significance levels, given enough time, even when the null hypothesis is true. For example, see Good (1956).

(b) Weight of Evidence

Closely related to the concept of likelihood is that of weight of evidence, which I mentioned before and promised to define.

Let us suppose that we have only two hypotheses under consideration, which might be because we have decided to consider hypotheses two at a time. Denote them by H and $\bar{H}$, where the bar over the second H denotes negation. (These need not be "simple statistical hypotheses," as defined in a moment.) Suppose further that we have an event, experimental result, or observation denoted by E. The conditional probability of E is either $P(E|H)$ or $P(E|\bar{H})$, depending on whether H or $\bar{H}$ is assumed. If H and $\bar{H}$ are "simple statistical hypotheses," then these two probabilities have sharp uncontroversial values given tautologically by the meanings of H and $\bar{H}$. Even if they are *composite* hypothesis, not "simple" ones, the Bayesian will still be prepared to talk about these two probabilities. In either case we can see, by four applications of the product axiom, or by two applications of Bayes's theorem, that

$$\frac{P(E|H)}{P(E|\bar{H})} = \frac{O(H|E)}{O(H)}$$

where O denotes *odds*. (The odds corresponding to a probability p are defined as $p/(1-p)$.) Turing (1941) called the right side of this equation *the factor in favor of the hypothesis H provided by the evidence E*, for obvious reasons. Its logarithm is the *weight of evidence* in favor of H, as defined independently by Peirce (1878), #13, and Minsky and Selfridge (1961). [But see #1382.] It was much used by Harold Jeffreys (1939/61), except that in that book he identified it with the final log-odds because his initial probabilities were taken as 1/2. He had previously (1936) used the general form of weight of evidence and had called it "support." The non-Bayesian uses the left side of the equation, and calls it the probability ratio, provided that H and $\bar{H}$ are simple statistical hypotheses. He SUTC the right side, because he does not talk about the probability of a hypothesis. The Bayesian, the doctor, the judge and the jury can appreciate the importance of the right side even with only the vaguest estimates of the initial odds of H. For example, the Bayesian (or at least the Doogian) can logically argue in the following manner (p. 70 of #13): If we assume that it was sensible to start a sampling experiment in the first place, and if it has provided appreciable weight of evidence in favor of some hypothesis, and it is felt that the hypothesis is not yet convincing enough, then it is sensible to enlarge the sample since we know that the final odds of the hypothesis have increased whatever they are. Such conclusions can be reached even though judgments of the relevant initial probability and of the utilities have never been announced. Thus, even when the initial probability is extremely vague, the axioms of subjective probability (and weight of evidence) can be applied.

When one or both of H and $\bar{H}$ are composite, the Bayesian has to assume relative initial probabilities for the simple components of the composite hypothesis. Although these are subjective, they typically seem to be less subjective than the initial probability of H itself. To put the matter more quantitatively, although this is not easy in so general a context, I should say that the judgment of the factor in favor of a hypothesis might typically differ from one person to another by up to about 5, while the initial odds of H might differ by a factor of 10 or 100 or 1000. Thus the separation of the estimation of the weight of evidence from the initial or final probability of H serves a useful purpose, especially for communication with other people, just as it is often advisable to separate the judgments of initial probabilities and likelihoods.

It often happens that the weight of evidence is so great that a hypothesis seems convincing almost irrespective of the initial probability. For example, in quantum mechanics, it seems convincing that the Schrödinger equation is approximately true (subject to some limitations), given the rest of some standard formulation of quantum mechanics, because of great quantities of evidence from a variety of experiments,such as the measurements of the frequencies of spectral lines to several places of decimals. The large weight of evidence makes it seem, to people who do not stop to think, that the initial probability of the equation, conditional on the rest of the theory, is irrelevant; but really there has to be an implicit judgment that the initial probability is not too low; for example, not less than 10^{-50}. (In a fuller discussion I would prefer to talk of the relative odds of *two* equations in competition.) How we judge such inequalities, whether explicitly or implicitly, is not clear: if we knew how we made judgments we would not call them judgments (#183). It must be something to do with the length of the equation (just as the total length of [the "meaningful" nonredundant parts of the] chromosomes in a cell could be used as a measure of complexity of an organism) and with its analogy with the classical wave equation and heat equation. (The latter has even suggested to some people, for example, Weizel [1953], that there is some underlying random motion that will be found to "explain" the equation.) At any rate the large weight of evidence permits the initial

probability to be SUTC and it leads to an apparent objectivism (the reliance on the likelihoods alone) that is really multisubjectivism. The same happens in many affairs of ordinary life, in perception (p. 68 of #13), in the law, and in medical diagnosis (for example, #755).

On a point of terminology, the factor in favor of a hypothesis is equal to the likelihood ratio, in the sense of Neyman, Pearson, and Wilks, only when both H and $\bar{H}$ are simple statistical hypotheses. This is another justification for using Turing's and Peirce's expressions, apart from their almost self-explanatory nature, which provides their potential for improving the reasoning powers of all people. Certainly the expression "weight of evidence" captures one of the meanings that was intended by ordinary language. It is not surprising that it was an outstanding philosopher who first noticed this: for one of the functions of philosophy is to make such captures. [It is a pity that Peirce's discussion contained an error.]

George Barnard, who is one of the Likelihood Brethren, has rightly emphasized the merits of graphing the likelihood function. A Bayesian should support this technique because the initial probability density can be combined with the likelihood afterwards. If the Bayesian is a subjectivist he will know that the initial probability density varies from person to person and so he will see the value of graphing of the likelihood function for communication. A Doogian will consider that even his own initial probability density is not unique so he should approve even more. Difficulties arise in general if the parameter space has more than two dimensions, both in picturing the likelihood hypersurface or the posterior density hypersurface. The problem is less acute when the hypersurfaces are quadratic in the neighborhood of the maximum. In any case the Bayesian can in addition reduce the data by using such quantities as expected utilities. Thus he has all the advantages claimed by the likelihood brotherhood, but has additional flexibility. [See also #862, p. 711 and #1444.]

(c) Maximum Likelihood, Invariance, Quasiutilities, and Quasilosses

Let us now consider the relationship between Bayesian methods and *maximum* likelihood.

In a "full-dress" Bayesian estimation of parameters, allowing for utilities, you compute their final distribution and use it, combined with a loss function, to find a single recommended value, if a point estimate is wanted. When the loss function is quadratic this implies that the point estimate should be the final expectation of the parameter (even for vector parameters if the quadratic is nonsingular). The final expectation is also appropriate if the parameter is a physical probability because the subjective expectation of a physical probability of an event is equal to the current subjective probability of that event.

If you do not wish to appear to assume a loss function, you can adopt the argument of Jeffreys (1939/61, Section 4.0). He points out that for a sample of size n (n observations), the final probability density is concentrated in a range of order $n^{-1/2}$, and that the difference between the maximum-likelihood value of the parameter and the mode of the final probability density is of the order $1/n$. (I call this last method, the choice of this mode, a Bayesian method "in mufti.") "Hence if the number of observations is large, the error committed by taking the maximum likelihood solution as the estimate is less than the uncertainty inevitable in any case. . . . The above argument shows that in the great bulk of cases its results are indistinguishable from those given by the principle of inverse probability, which supplies a justification for it." It also will not usually make much difference if the parameter is assumed to have a uniform initial distribution. (Jeffreys, 1939/61, p. 145; p. 55 of #13. L. J. Savage, 1959/62, p. 23, named estimation that depends on this last point "stable estimation.")

By a slight extension of Jeffreys's argument, we can see that a point estimate based on a loss function, whether it is the expectation of the parameter or some other value (which will be a kind of average) induced by the loss function, will also be approximated by using the Bayes method in mufti, and by the maximum-likelihood estimate, when the number of observations is large. Thus the large-sample properties of the maximum-likelihood method cannot be used for distinguishing it from a wide class of Bayesian methods, whether full-dress or in mufti. This is true whether we are dealing with point estimates or interval estimates. Interval estimates and posterior distributions are generally more useful, but point estimates are easier to talk about and we shall concentrate on them for the sake of simplicity.

One may also regard the matter from a more geometrical point of view. If the graph of the likelihood function is sharply peaked, then the final density will also usually be sharply peaked at nearly the same place. This again makes it clear that there is often not much difference between Bayesian estimation and maximum-likelihood estimation, provided that the sample is large. This argument applies provided that the number of parameters is itself not large.

All this is on the assumption that the Bayesian assumptions are not dogmatic in the sense of ascribing zero initial probability to some range of values of the parameter; though "provisional dogmatism" is often justifiable to save time, where you hold at the back of your mind that it might be necessary to make an adjustment in the light of the evidence. Thus I do not agree with the often-given dogmatic advice that significance tests *must* be chosen before looking at the results of an experiment, although of course I appreciate the point of the advice. It is appropriate advice for people of bad judgment.

It is perhaps significant that Daniel Bernoulli introduced the method of maximum likelihood, in a special case, at almost the same time as the papers by Bayes and Laplace on inverse probability were published. But, as I said before, it is the logical rather than the historical connections that I wish to emphasize most. I merely state my belief that the influence of informal Bayesian thinking on apparently non-Bayesian methods has been considerable at both a conscious and a less conscious level, ever since 1763, and even from 1925 to 1950 when non-Bayesian methods were at their zenith relative to Bayesian ones.

Let us consider loss functions in more detail. In practice, many statisticians

who do not think of themselves as Bayesians make use of "squared-error loss," and regard it as Gauss-given, without usually being fully aware that a loss is a negative utility and smacks of Bayes. The method of least squares is not always regarded as an example of minimizing squared loss (see Eisenhart, 1964), but it *can* be thought of that way. It measures the value of a putative regression line for fitting given observations. Since statisticians might not always be happy to concur with this interpretation, perhaps a better term for "loss" when used in this conventional way would be "quasiloss" or "pseudoloss." We might use it, *faute de mieux*, when we are not sure what the utility is, although posterior distributions for the parameters of the regression line would preserve more of the information.

If the loss is an analytic function it has to be quadratic in the neighborhood of the correct answer, but it would be more realistic in most applications to assume the loss to be asymptotic to some value when the estimate is far away from the correct value. Thus a curve or surface having the shape of an upside-down normal or multinormal density would be theoretically better than "squared loss" (a parabola or paraboloid). But when the samples are large enough the "tails of the loss function" perhaps do not usually affect the estimates much, except when there are outliers.

Once the minimization of the sum of squared residuals is regarded as an attempt to maximize a utility, it leads us to ask what other substitutes for utility might be used, *quasiutilities* if you like. This question dates back over a quarter of a millennium in estimation problems. Like quasilosses, which are merely quasiutilities with a change of sign, they are introduced because it is often difficult to decide what the real utilities are. This difficulty especially occurs when the applications of your work are not all known in advance, as in much pure science (the "knowledge business" to use Kempthorne's term). A quasiutility might be somewhat *ad hoc*, used partly for its mathematical convenience in accordance with Type II rationality. It is fairly clear that this was an important reason historically for the adoption of the method of least squares on a wide scale.

Fisher once expressed scorn for economic applications of statistics, but he introduced his ingenious concept of amount of information in connection with the estimation of parameters, and it can be regarded as another quasiutility. It measures the expected value of an experiment for estimating a parameter. Then again Turing made use of expected weight of evidence for a particular application in 1941. It measures the expected value of an experiment for discriminating between two hypotheses. The idea of using the closely related Shannon information in the design of statistical experiments has been proposed a number of times (Cronbach, 1953; Lindley, 1956; #77), and is especially pertinent for problems of search such as in dendroidal medical diagnosis (for example, #592). It measures the expected value of an experiment for distinguishing between several hypotheses. *In this medical example the doctor should switch to more "real" utilities for his decisions when he comes close enough to the end of the search to be able to "backtrack."* A number of other possibile quasiutilities are suggested

in ##592 & 755, some of which are invariant with respect to transformations of the parameter space.

In all these cases, it seems to me that the various concepts are introduced essentially because of the difficulty of making use of utility in its more standard economic sense. I believe the term "quasiutility" might be useful in helping to crystallize this fact, and thus help to clarify and unify the logic of statistical inference. The quasiutilities mentioned so far are all defined in terms of the probability model alone, but I do not regard this feature as part of the definition of a quasiutility.

Even in the law, the concept of weight of evidence (in its ordinary linguistic sense, which I think is usually the same as its technical sense though not formalized) helps to put the emphasis on the search for the truth, leaving utilities to be taken into account later. One might even conjecture that the expressions "amount of information" and "weight of evidence" entered the *English language* because utilities cannot always be sharply and uncontroversially estimated. Both these expressions can be given useful quantitative meanings defined in terms of probabilities alone, and so are relevant to the "knowledge business."

These various forms of quasiutility were not all suggested with the conscious idea of replacing utility by something else, but it is illuminating to think of them in this light, and, if the *word* "quasiutility" had been as old as quasiutilities themselves, the connection could not have been overlooked. It shows how influential Bayesian ideas can be in the logic if not always in the history of statistics. The history is difficult to trace because of the tendency of many writers (i) to cover up their tracks, (ii) to forget the origins of their ideas, deliberately or otherwise, and (iii) not to be much concerned with "kudology," the fair attributions of credit, other than the credit of those they happen to like such as themselves.

The word "quasiutility" provokes one to consider whether there are other features of utility theory that might be interesting to apply to quasiutilities, apart from the maximization of their expectations for decision purposes. One such feature is the use of minimax procedures, that is, cautious decision procedures that minimize the maximum expected loss (or quasiloss here). Although minimax procedures are controversial, they have something to be said for them. They can be used when all priors are regarded as possible, or more generally when there is a class of possible priors (Hurwicz, 1951; **#26**, where this generalized minimax procedure was independently proposed: "Type II minimax"), so that there is no unique Bayesian decision: then the minimax procedure corresponds to the selection of the "least favorable" prior, in accordance with a theorem of Wald (1950). When the quasiutility is invariant with respect to transformations of the parameter space, then so is the corresponding minimax procedure and it therefore has the merit of decreasing arbitrariness. When the quasiutility is Shannon (or Szilard) information, the minimax procedure involves choosing the prior of maximum entropy (##618, 622), a suggestion made for other reasons by Jaynes (1957). The maximum-entropy method was reinterpreted as a method for *selecting null hypothes*

emphasize this interpretation because the formulation of hypotheses is often said to lie outside the statistician's domain of formal activity, *qua* statistician. It has been pointed out that Jeffreys's invariant prior (Jeffreys, 1946) can be regarded as a minimax choice when quasiutility is measured by weight of evidence (##618, 622). Thus other invariant priors could be obtained from other invariant quasiutilities (of which there is a one-parameter family mentioned later).

Jeffreys's invariant prior is equal to the square root of the determinant of Fisher's information matrix, although Jeffreys (1946) did not express it this way explicitly. Thus there can be a logical influence from non-Bayesian to Bayesian methods, and of course many other examples of influence in this direction could be listed.

Let us return to the discussion of Maximum Likelihood (ML) estimation. Since nearly all methods lead to ROME (Roughly Optimal Mantic Estimation) when samples are large, the real justification for choosing one method rather than another one must be based on samples that are not large.

One interesting feature of ML estimation, a partial justification for it, is its invariance property. That is, if the ML estimate of a parameter θ is denoted by $\hat{\theta}$, then the ML estimate of $f(\theta)$, for any monotonic function f, even a discontinuous one, is simply $f(\hat{\theta})$. Certainly invariant procedures have the attraction of decreasing arbitrariness to some extent, and it is a desideratum for an *ideal* procedure. But there are other invariant procedures of a more Bayesian tone to which I shall soon return: of course a completely Bayesian method would be invariant if the prior probabilities and utilities were indisputable. Invariance, like patriotism, is not enough. An example of a very bad invariant method is to choose as the estimate the least upper bound of all possible values of the parameter if it is a scalar. This method is invariant under all increasing monotonic transformations of the parameter!

Let us consider what happens to ML estimation for the physical probabilities of a multinomial distribution, which has been used as a proving ground for many philosophical ideas.

In the notation used earlier, let the frequencies in the cells be $n_1, n_2, \ldots, n_t$, with total sample size N. Then the ML estimates of the physical probabilities are n_i/N, $i = 1, 2, \ldots, t$. Now I suppose many people would say that a sample size of $n = 1{,}000$ is large, but even with this size it could easily happen that one of the n_i's is zero, for example, the letter Z could well be absent in a sample of 1,000 letters of English text. Thus a sample might be large in one sense but effectively small in another (##38, 83, 398). If one of the letters is absent ($n_i = 0$), then the maximum-likelihood estimate of p_i is zero. This is an appallingly bad estimate if it is used in a gamble, because if you believed it (which you wouldn't) it would cause you to give arbitrarily large odds against that letter occurring on the next trial, or perhaps ever. Surely even the Laplace-Lidstone estimate $(n_i + 1)/(N + t)$ would be better, although it is not optimal. The estimate of Jeffreys (1946), $(n_i + 1/2)/(N + t/2)$, which is based on his "invariant prior," is also better (in the same sense) than the ML estimate. Still better methods are available which

are connected with reasonable "Bayesian significance tests" for multinomial distributions (##398, 547).

Utility and quasiutility functions are often invariant in some sense, although "squared loss" is invariant only under *linear* transformations. For example, if the utility in estimating a vector parameter θ as ϕ is $u(\theta,\phi)$, and if the parameter space undergoes some one-one transformation $\theta^* = \psi(\theta)$ we must have, for consistency, $\phi^* = \psi(\phi)$ and $u^*(\theta^*,\phi) = u(\theta,\phi)$, where u^* denotes the utility function in the transformed parameter space.

The principle of selecting the least favorable prior when it exists, in accordance with the minimax strategy, may be called *the principle of least utility*, or, when appropriate, *the principle of least quasiutility*. Since the minimax procedure must be invariant with respect to transformations of the problem into other equivalent languages, it follows that the principle of least utility leads to an invariant prior. This point was made in ##618, 622. It was also pointed out there (see also ##699, 701, 810 and App. C of #815) that there is a class of invariant quasiutilities *for distributions*. Namely, the quasiutility of assuming a distribution of density $g(x)$, when the true distribution of x if $F(x)$, was taken as

$$\int \log\left\{g(\mathrm{x})\,[\det \Delta\,(\mathrm{x})]^{-1/2}\right\} dF(\mathrm{x})$$

where

$$\Delta(\theta) = \left\{ -\left.\frac{\partial^2 u(\theta,\phi)}{\partial\phi_i \partial\phi_j}\right|_{\phi\,=\,0} \right\} \quad i, j = 1,2,\ldots.$$

From this it follows further that

$$[\det\Delta(\mathrm{x})]^{1/2}$$

is an invariant prior, though it might be "improper" (have an infinite integral). In practice improper priors can always be "shaded off" or truncated to give them propriety (p. 56 of #13).

If θ is the vector parameter in a distribution function $F(\mathrm{x}|\theta)$ of a random variable x, and θ is not to be used for any other purpose, then in logic we must identify $u(\theta,\phi)$ with the utility of taking the distribution to be $F(\mathrm{x}|\phi)$ instead of $F(\mathrm{x}|\theta)$. One splendid example of an invariant utility is expected weight of evidence per observation for discriminating between θ and ϕ or "dinegentropy,"

$$u_0^0(\theta,\phi) = \int \log \frac{dF(\mathrm{x}|\theta)}{dF(\mathrm{x}|\phi)}\, dF(\mathrm{x}|\theta),$$

which is invariant under non-singular transformations both of the *random* variable and of the parameter space. (Its use in statistical mechanics dates back to Gibbs.) Moreover it is additive for entirely independent problems, as a utility function should be. With this quasituility, $\Delta(\theta)$ reduces to Fisher's information matrix, and the square root of the determinant of $\Delta(\theta)$ reduces to Jeffreys's invariant prior. The dinegentropy was used by Jeffreys (1946) as a measure

of distance between two distributions. The distance of a distribution from a correct one *can* be regarded as a kind of loss function. Another additive invariant quasiutility is (#82; Rényi, 1961; p. 180 of #755) the "generalized dinegentropy,"

$$u_c(\theta,\phi) = \frac{1}{c}\log\left[\frac{dF(\mathbf{x}|\theta)}{dF(\mathbf{x}|\phi)}\right]^c dF(\mathbf{x}|\theta) \quad (c>0),$$

the limit of which as $c \to 0$ is the expected weight of evidence, $u_0(\theta,\phi)$, somewhat surprising at first sight. The square root of the determinant of the absolute value of the Hessian of this utility at $\phi = \theta$ is then an invariant prior indexed by the non-negative number c. Thus there is a continuum of additive invariant priors of which Jeffreys's is an extreme case. For example, for the mean of a univariate normal distribution the invariant prior is uniform, mathematically independent of c. The invariant prior for the variance ϕ is $\sigma^{-1}\sqrt{\{2(1+c)\}}$, which is proportional to σ^{-1} and so is again mathematically independent of c.

In more general situations the invariant prior will depend on c and will therefore not be unique. In principle it might be worth while to assume a ("type III") distribution for c, to obtain an average of the various additive invariant priors. It might be best to give extra weight to the value $c = 0$ since weight of evidence seems to be the best general-purpose measure of corroboration (##211, 599).

It is interesting that Jeffreys's invariant prior, and its generalizations, and also the principles of maximum entropy and of minimum discriminaability (Kullback, 1959) can all be regarded as applications of the principle of least quasiutility. This principle thus unifies more methods than has commonly been recognized. The existing criticisms of minimax procedures thus apply to these special cases.

The term "invariance" can be misleading if the class of transformations under which invariance holds is forgotten. For the invariant priors, although this class of transformations is large, it does not include transformations to a different *application* of the parameters. For example, if θ has a physical meaning, such as height of a person, it might occur as a parameter in the distribution of her waist measurement or her bust measurement, and the invariance will not apply between these two applications. This in my opinion (and L. J. Savage's, July 1959) is a logical objection to the use of invariant priors when the parameters have clear physical meaning. To overcome this objection completely it would perhaps be necessary to consider the joint distribution of all the random variables of potential interest. In the example this would mean that the joint distribution of at least the "vital statistics," given θ, should be used in constructing the invariant prior.

There is another argument that gives a partial justification for the use of the invariant priors in spite of Savage's objection just mentioned. It is based on the notion of "marginalism" in the sense defined by Good (pp. 808-809 of #174; p. 61 of #603B; p. 15 of #732). I quote from the last named. "It is only in marginal cases that the choice of the prior makes much difference (when it is chosen to give the non-null hypothesis a reasonable chance of winning on the size of

sample we have available). Hence the name marginalism. It is a trick that does not give accurate final probabilities, but it protects you from missing what the data is trying to say owing to a careless choice of prior distribution." In accordance with this principle one might argue, as do Box and Tiao (1973, p. 44) that a prior should, at least on some occasions, be uninformative relative to the experiment being performed. From this idea they derive the Jeffreys invariant prior.

It is sometimes said that the aim in estimation is not necessarily to minimize loss but merely to obtain estimates close to the truth. But there is an implicit assumption here that it is better to be closer than further away, and this is equivalent to the assumption that the loss function is monotonic and has a minimum (which can be taken as zero) when the estimate is equal to the true value. This assumption of monotonicity is not enough to determine a unique estimate nor a unique interval estimate having an assigned probability of covering the true value (where the probability might be based on information before or after the observations are taken). But for large enough samples (*effectively* large, for the purpose in hand), as I said, all reasonable methods of estimation lead to Rome, if Rome is not too small.

(d) A Bayes/Non-Bayes Compromise for Probability Density Estimation

Up to a few years ago, the only nonparametric methods for estimating probability densities, from observations $x_1, x_2, \ldots, x_N$, were non-Bayesian. These, methods, on which perhaps a hundred papers have been written, are known as *window methods*. The basic idea, for estimating the density at a point x, was to see how many of the N observations lie in some interval or region around x, where the number ν of such observations tends to infinity while $\nu/N \to 0$ when $N \to \infty$. Also less weight is given to observations far from x than to those close to x, this weighting being determined by the shape of the window.

Although the window methods have some intuituve appeal it is not clear in what way they relate to the likelihood principle. On the other hand, if the method of ML is used it leads to an unsatisfactory estimate of the density function, namely a collection of fractions $1/N$ of Dirac delta functions, one at each of the observations. (A discussant: Go all the way to infinity if they are Dirac functions. Don't be lazy! IJG: Well I drew them a little wide so they are less high to make up for it.) There is more than one objection to this estimate; partly it states that the next observation will certainly take a value that it almost certainly will not, and partly it is not smooth enough to satisfy your subjective judgment of what a density function should look like. It occurred to me that it should make sense to apply a "muftian" Bayesian method, which in this application means finding some formula giving a posterior density in the function space of all density functions for the random variable X, and then maximizing this posterior density so as to obtain a single density function (single "point in function space") as the "best" estimate of the whole density function for X. But this means that

from the log-likelihood $\Sigma \log f(x_i)$ we should subtract a "roughness penalty" before maximizing. (##733, 699, 701, 810, 1200.) There is some arbitrariness in the selection of this roughness penalty (which is a functional of the required density function f), which was reduced to the estimation of a single hyperparameter, but I omit the details. The point I would like to make here is that the method can be interpreted in a non-Bayesian manner, although it was suggested for Bayesian reasons. Moreover, in the present state of the art, only the Bayesian interpretation allows us to make a comparison between two hypothetical density functions. The weight of evidence by itself is not an adequate guide for this problem. Then again the non-Bayesian could examine the operational characteristics of the Bayesian interpretation. The Doogian should do this because it might lead him to a modification of the roughness penalty. The ball travels backwards and forwards between the Bayesian and non-Bayesian courts, the ball-game as a whole forming a plank of the Doogian platform.

It is easy to explain why the method of ML breaks down here. It was not designed for cases where there are very many parameters, and in this problem there is an infinite number of them, since the problem is nonparametric. (A nonparametric problem is one where the class of distribution functions cannot be specified in terms of a finite number of parameters, but of course any distribution can be specified in terms of an infinite number of parameters. My method of doing so is to regard the square root of the density function as a point in Hilbert space.)

To select a roughness penalty for multidimensional density functions, I find consistency appealing, in the sense that the estimate of densities that are known to factorize, such as $f(x)g(y)$ in two dimensions, should be the same whether f and g are estimated together or separately. This idea enabled me to propose a multidimensional roughness penalty but numerical examples of it have not yet been tried. [See also #1341.]

An interesting feature of the *subtractive roughness-penalty method* of density estimation, just described, is that it can be made invariant with respect to transformations of the x axes, even though such transformations could make the true density function arbitrarily rough. The method proposed for achieving invariance was to make use of the tensor calculus, by noticing that the elements of the matrix $\Delta(\theta)$ form a covariant tensor, which could be taken as the "fundamental tensor" g_{ij} analogous to that occurring in General Relativity. For "quadratic loss" this tensor becomes a constant, and, as in Special Relativity, it is then not necessary to use tensors. The same thing happens more generally if $u(\theta,\phi)$ is any function (with continuous second derivatives) of a quadratic.

(e) Type II Maximum Likelihood and the Type II Likelihood Ratio

The notion of a hierarchy of probabilities, mentioned earlier, can be used to produce a compromise between Bayesian and non-Bayesian methods, by treating hyperparameters in some respects as if they were ordinary parameters. In particular, a Bayes factor can be maximized with respect to the hyperparameters,

and the hyperparameters so chosen (their "Type II ML" values) thereby fix the ordinary prior, and therefore the posterior distribution of the ordinary parameters. This *Type II ML method* could also be called the *Max Factor method*. This technique was well illustrated in #547. It ignores only judgments you might have about the Type III distributions, but I have complete confidence that this will do far less damage than ignoring all your judgments about Type II distributions as in the ordinary method of ML. Certainly in the reference just mentioned the Type II ML estimates of the physical probabilities were far better than the Type I ML estimates.

The same reference exemplified the *Type II likelihood Ratio*. The ordinary (Neyman-Pearson-Wilks) Likelihood Ratio (LR) is defined as the ratio of two maximum likelihoods, where the maxima are taken within two spaces corresponding to two hypotheses (one space embedded in the other). The ratio is then used as a test statistic, its logarithm to base $1/\sqrt{e}$ having asymptotically (for large samples) a chi-squared distribution with a number of degrees of freedom equal to the difference of the dimensionalities of the two spaces. The Type II Likelihood Ratio is defined analogously as

$$\max_{\theta\epsilon\omega} P\{E|H(\theta)\} / \max_{\theta\epsilon\Omega} P\{E|H(\theta)\}$$

where θ is now a hyperparameter in a prior $H(\theta)$, Ω is the set of all values of θ and ω is a subset of Ω. In the application to multinomial distributions this led to a new statistic called G having asymptotically a chi-squared distribution with one degree of freedom (corresponding to a single hyperparameter, namely the parameter of a symmetric Dirichlet distribution). Later calculations showed that this asymptotic distribution was accurate down to fantastically small tail-area probabilities such as 10^{-16}, see #862. In this work it was found that if the Bayes factor F, based on the prior selected in #547 [see also #1199] were used as a non-Bayesian statistic, in accordance with the Bayes/non-Bayes compromise, it was almost equivalent to the use of G in the sense of giving nearly the same significance levels (tail-area probabilities) to samples. It was also found that the Bayes factor based on the (less reasonable) Bayes postulate was roughly equivalent in the same sense, thus supporting my claims for the Bayes/non-Bayes compromise.

(f) The Non-Uniqueness of Utilities

For some decision problems the utility function can be readily measured in monetary terms; for example, in a gamble. In a moderate gamble the utility can reasonably be taken as proportional to the relevant quantities of money. Large insurance companies often take such "linear" gambles. But in many other decision problems the utility is not readily expressible in monetary terms, and can also vary greatly from one person to another. In such cases the Doogian, and many a statistician who is not Doogian or does not know that he is, will often wish to keep the utilities separate from the rest of the statistical analysis

if he can. There are exceptions because, for example, many people might assume a squared loss function, but with different matrices, yet they will all find expected values to be the optimal estimates of the parameters.

One implication of the recognition that utilities vary from one person to another is that the expected benefit of a client is not necessarily the same, *nor even of the same sign*, as that of the statistical consultant. This can produce ethical problems for the statistician, although it may be possible to reward him in a manner that alleviates the problems. (See, for example, ##**26**, 690a.)

One example of this conflict of interests relates to the use of confidence-interval estimation. This technique enables the statistician to ensure that his interval estimates (*asserted* without reference to probability) will be correct say 95% of the time in the long run. If he is not careful he might measure his utility gain by this fact alone (especially if he learns his statistics from cookbooks) and it can easily happen that it won't bear much relation to his client's utility on a specific occasion. The client is apt to be more concerned with the final probability that the interval will contain the true value of the parameter.

Neyman has warned against dogmatism but his followers do not often give nor heed the warning. Notice further that there are *degrees* of dogmatism and that greater degrees can be justified when the principles involved are the more certain. For example, it seems more reasonable to be dogmatic that 7 times 9 is 63 than that witches exist and should be caused not to exist. Similarly it is more justifiable to be dogmatic about the axioms of subjective probability than to insist that the probabilities can be sharply judged or that confidence intervals should be used in preference to Bayesian posterior intervals. (Please don't call them "Bayesian confidence intervals," which is a contradiction in terms.)

Utilities are implicit in some circumstances even when many statisticians are unaware of it. Interval estimation provides an example of this; for it is often taken as a criterion of choice between two confidence intervals, both having the same confidence coefficient, that the shorter interval is better. Presumably this is because the shorter interval is regarded as leading to a more economical search or as being in general more informative. In either case this is equivalent to the use of an informal utility or quasiutility criterion. It will often be possible to improve the interval estimate by taking into account the customer's utility function more explicitly.

An example of this is when a confidence interval is stated for the position of a ship, in the light of direction finding. If an admiral is presented with say an elliptical confidence region, I suspect he would reinterpret it as a posterior probability density surface, with its mode in the center. (#618; Good, 1951.) The admiral would rationally give up the search when the expense per hour sank below the expected utility of locating the ship. In other words, the client would sensibly ignore the official meaning of the statistician's assertion. If the statistician knows this, it might be better, at least for his client, if he went Bayesian (in some sense) and gave the client what he wanted.

(g) Tail-Area Probabilities

Null hypotheses are usually known in advance to be false, and the point of significance tests is usually to find out whether they are nevertheless approximately true (p. 90 of #13). In other words *a null hypothesis is usually composite even if only just*. But for the sake of simplicity I shall here regard the null hypothesis as a simple statistical hypothesis, as an approximation to the usual real-life situation.

I have heard it said that the notion of tail-area probabilities, for the significance test of a null hypothesis H_0 (assumed to be a simple statistical hypothesis), can be treated as a primitive notion, not requiring further analysis. But this cannot be true irrespective of the test criterion and of the plausible alternatives to the null hypothesis, as was perhaps originally pointed out by Neyman and E. S. Pearson. A value X_1 of the test criterion X should be regarded as "more extreme" than another one X_2 only if the observation of X_1 gives "more evidence" against the null hypothesis. To give an interpretation of "more evidence" it is necessary to give up the idea that tail-areas are primitive notions, as will soon be clear. One good interpretation of "more evidence" is that the weight of evidence against H_0 provided by X_1 is greater than that provided by X_2, that is

$$\log \frac{\text{P.D.}\,(X_1|H_1)}{\text{P.D.}\,(X_1|H_0)} > \log \frac{\text{P.D.}\,(X_2|H_1)}{\text{P.D.}\,(X_2|H_0)},$$

where H_1 is the negation of H_0 and is a composite statistical hypothesis, and P.D. stands for "probability density." (When H_0 and H_1 are both simple statistical hypotheses there is little reason to use "tail-area" significance tests.) This interpretation of "more extreme" in particular provides a solution to the following logical difficulty, as also does the Neyman-Pearson technique if all the simple statistical hypotheses belonging to H_1 make the simple likelihood ratio monotonic increasing as x increases.

Suppose that the probability density of a test statistic X, given H_0, has a known shape, such as that in Figure 1a. We can transform the x axis so that the density function becomes any density function we like, such as that illustrated in Figure 1b. We then might not know whether the x's "more extreme" than the observed one should be interpreted as all the shaded part of 1(b), where the ordinates are smaller than the one observed. Just as the tail-area probability wallah points out that the Bayes postulate is not invariant with respect to transformations of the x axis, the Bayesian can say *tu quoque*. (Compare, for example, p. 53, of #750; Kalbfleisch, 1971, §7, 1-8.) Of course Doogians and many other modern Bayesians are not at all committed to the Bayes postulate, though they often use it as an approximation to their honest judgment, or marginalistically.

When tail-areas are used for significance testing, we need to specify what is meant by a "more extreme" value of the criterion. A smaller ordinate might

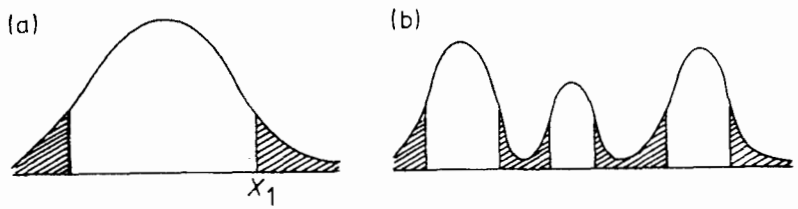

Figure 1.

not be appropriate, as we have just seen. I believe it is a question of ordering the values of the ordinate according to the weight of evidence against the null hypothesis, as just suggested. (Sometimes this ordering is mathematically independent of the relative initial probabilities of the simple statistical hypotheses that make up the composite non-null hypothesis H_1. In this case the interpretation of "more extreme" is maximally robust modulo the Bayesian assumptions.) This or similar fact is often swept UTC, although a special case of it is often implicit when it is pointed out that sometimes a single tail should be used and sometimes a double tail, depending on the nature of the non-null hypotheses.

For some problems it would not be appropriate to interpret "more extreme" to mean "further to the right" nor "either further to the right of one point or further to the left of another" (i.e. for "double tails"). For example, the null hypothesis might be a bimodal distribution with mean zero, the rivals being unmodal also with mean zero. Then we might need to regard values of the random variable close to the origin as significant, in addition to large positive and negative values. We'd be using a "triple tail" so to speak. All this comes out in the wash when "more extreme" is interpreted in terms of weight of evidence.

It is stimulating to consider what is "more extreme" in multivariate problems. It will be adequate to think of bivariate problems which are enough to bring out all the philosophical [or logical] aspects, which are more important than the mathematical ones. We might first ask what is the analogue of being "further to the right." One analogue is being "further to the north and east." This analogue is often dubious (unless the two independent varaiables are like chalk and cheese, or like oil and water) even without reference to any Bayesian or Neymanian-Pearsonian ideas. For under a linear transformation of the independent variables, such as an orthogonal transformation, there are a continuous infinity of different regions that are further to the north and east. The corresponding ambiguity in one dimension refers merely to the question of whether a single tail is more or less appropriate than a double tail.

The previously mentioned elucidation of "more extreme" in terms of weight of evidence applies just as much to multivariate problems as to univariate ones, and provides and answer to this "north-east" difficulty.

Even when a sensible meaning is ascribed to the expression "more extreme," my impression is that small tail-areas, such as 1/10000, are by no means as strong evidence against the null hypothesis as is often supposed, and this is one reason why I believe that Bayesian methods are important in applications where small tail areas occur, such as medical trials, and even more in ESP, radar, cryptanalysis, and ordinary life. It would be unfortunate if a radar signal were misinterpreted through overlooking this point, thus leading to the end of life on earth! *The more important a decision the more "Bayesian" it is apt to be.*

The question has frequently been raised of how the use of tail-area significance tests can be made comformable with a Bayesian philosophy. (See, for example, Anscombe [1968/69].) An answer had already appeared on p. 94 of #13, and I say something more about it here. (See also p. 61 of **#603B**.)

A reasonable informal Bayesian interpretation of tail-area probabilities can be given in some circumstances by treating the criterion X as if it were the whole of the evidence (even if it is not a sufficient statistic). Suppose that the probability density f_0 of X given H_0 is known, and that you can make a rough subjective estimate of the density f_1 given $\bar{H}_0$. (If you cannot do this at all then the tail area method is I think counterintuitive.) Then we can calculate the Bayes factor against H_0 as a ratio of ordinates $f_1(X)/f_0(X)$. It turns out that this is often the order of magnitude of $(1/\sqrt{N})\int_X^\infty f_1(x)dx/\int_X^\infty f_0(x)dx$, where N is the sample size, and this in its turn will be somewhat less than $1/(P\sqrt{N})$ where P is the right-hand tail-area probability on the null hypothesis. (See p. 863 of #127; improved on p. 416 of #547; and still further in #862.) Moreover, this argument suggests that, for a fixed sample size, there should be a roughly monotonic relationship and a *very* rough proportionality between the Bayes factor F against the null hypothesis and the reciprocal of the tail-area probability, P, provided of course that the non-null hypothesis is not at all specific. (See also p. 94 of #13; #547.)

Many elementary textbooks recommend that test criteria should be chosen before observations are made. Unfortunately this could lead to a data analyst's missing some unexpected and therefore probably important feature of the data. There is no existing substitute for examining the original observations with care. This is often more valuable than the application of formal significance tests. If it is easy and inexpensive to obtain new data then there is little objection to the usual advice, since the original data can be used to formulate hypotheses to be tested on later sample. But often a further sample is expensive or virtually impossible to obtain.

The point of the usual advice is to protect the statistician against his own poor judgment.

A person with bad judgment might produce many far-fetched hypotheses on the basis of the first sample. Thinking that they were worth testing, if he were non-Bayesian he would decide to apply standard significance tests to these hypotheses on the basis of a second sample. Sometimes these would pass the test

but some one with good judgment might be able to see that they were still improbable. It seems to me that the ordinary method of significance tests makes some sense because experimenters often have reasonable judgment in the formulation of hypotheses, so that the initial probabilities of these hypotheses are not usually entirely negligible. A statistician who believes his client is sensible might assume that the hypotheses formulated in advance by the client are plausible, without trying to produce an independent judgment of their initial probabilities.

Let us suppose that data are expensive and that a variety of different non-null hypotheses have been formulated on the basis of a sample. Then the Bayesian analyst would try, in conjunction with his client, to judge the initial probabilities $q_1, q_2, \ldots$ of these hypotheses. Each separate non-null hypothesis might be associated with a significance test if the Bayesian is Doogian. These tests might give rise to tail-area probabilities $P_1, P_1, P_3, \ldots$. How can these be combined into a single tail-area probability? (#174.)

Let us suppose that the previous informal argument is applicable and that we can interpret these tail-area probabilities as approximate Bayes factors $C/P_1, C/P_2, C/P_3, \ldots$ against the null hypothesis, these being in turn based on the assumption of the various rival non-null hypotheses. ("Significance tests in parallel.") By a theorem of weighted averages of Bayes factors, it follows that the resulting factor is a weighted average of these, so that the equivalent tail-area probability is about equal to a weighted harmonic mean of $P_1, P_2, P_3, \ldots$, with weights $q_1, q_2, q_3, \ldots$. This result is not much affected if C is a slowly decreasing function of P instead of being constant, which I believe is often the case. Nevertheless the harmonic-mean rule is only a rule of thumb.

But we could now apply the Bayes/non-Bayes compromise for the invention of test criteria, and use this weighted harmonic mean as a non-Bayes test criterion (p. 863 of #127; ##547, 862).

The basic idea of the Bayes/non-Bayes compromise for the invention of test criteria is that you can take a Bayesian model, which need not be an especially good one, come up with a Bayes factor on the basis of this model, but then use it as if it were a non-Bayesian test criterion. That is, try to work out or "Monte Carlo" its distribution based on the null hypothesis, and also its power relative to various non-null hypotheses.

An example of the Bayes/non-Bayes compromise arises in connection with discrimination between two approximately multinomial distributions. A crude Bayesian model would assume that the two distributions were precisely multinomial and this would lead to a linear discriminant function. This could then be used in a non-Bayesian manner or it might lead to the suggestion of using a linear discriminant function optimized by some other, possibly non-Bayesian, method. Similarly an approximate assumption of multinormality for two hypotheses leads to a quadratic discriminant function with a Bayesian interpretation but which can then be interpreted non-Bayesianwise. (See pp. 49-50 of #397 where there are further references.)

Let us now consider an example of an experimental design. I take this example from Finney (1953, p. 90) who adopts an orthodox (non-Bayesian) line. Finney emphasizes that, in his opinion, you should decide in advance how you are going to analyze the experimental results of a designed experiment. He considered an experimental design laid out as shown in Figure 2. The design consists of ten plots, consisting of five blocks each divided into two plots. We decide to apply treatment A and treatment B in a random order within each block, and we happen to get the design shown. Now this design could have arisen by another process: namely by selecting equiprobably the five plots for the application of treatment A from the $10!/(5!)^2 = 252$ possibilities. Finney then says, "The form of analysis depends not on the particular arrangement of plots and varieties in

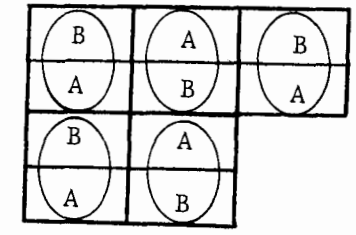

Figure 2. An agricultural experiment.

the field [I have been talking about *treatments* instead here but it does not affect the argument] but on the process of randomization from which the particular one was selected." (Perhaps one should talk of a stochastic or random design *procedure* and a *realization of the procedure.*) For one design procedure we would perhaps use the comparison within the five pairs, and for the other procedure we would compare the set of five yields from treatment A with the set of five yields from treatment B. Leaving aside the analysis of variance, we might find that every plot A did better than every plot B, thus bringing off a distribution-free chance of 1/252; but we are "permitted" to say merely that the chance is 1/32 if the design procedure was based on the five blocks. Suppose the statistician hadn't said which was his design and then he'd dropped dead after the experiment and suppose this is an important experiment organized by the government to decide whether a certain big expensive and urgent food production method was to be put into effect. Would it be reasonable to search the statistician's papers carefully to find out what his intentions had been? Or would it on the other hand be reasonable to call in agriculturalists to look at the plots in the field in order to try to decide which design would have been more reasonable? There are of course reasons for choosing one design rather than another one. So, if you entirely accept the Fisherian logic (as exemplified by Finney) you are whole-heartedly trusting the original judgment of choice of design: this is what the mystique recommends. My own feeling is that you would do better to judge

the prior probabilities that each of the two designs is to be preferred, and then use these probabilities as weights in a procedure for combining significance tests (#174 and p. 83 of #750).

A living agriculturalist might examine the field and say that the design corresponding to the tail-area probability of 1/32 deserved twice as much weight as the other design. Then the harmonic-mean rule of thumb would suggest that the equivalent tail-area probabity from the observations is

$$\frac{1}{\frac{2}{3} \times 32 + \frac{1}{3} \times 252} \approx \frac{1}{105}.$$

Of course we might do better by using the analysis of variance in a similar manner. I have used a distribution-free approach for the sake of simplicity. This imprecise result is better than either of the precise ones, 1/32 and 1/252. I predict that lovers of the "precisio fallacy" will ignore all this.

It is often said that non-Bayesian methods have the advantage of conveying the evidence in an experiment in a self-contained manner. But we see from the example just discussed that they depend on a previous judgment; which in the special case of the dead-dropping of the statistician, has to be a posterior judgment. So it's misleading to tell the student he must decide on his significance test in advance, although it is correct according to the Fisherian technique.

(h) Randomness, and Subjectivism in the Philosophy of Physics

I would have included a detailed discussion on the use of random sampling and random numbers, but have decided not to do so because my views on the subject are explained, for example, on p. 255 of #85A *and* on pp. 83-90. The use of random sampling is a device for obtaining apparently precise objectivity but this precise objectivity is attainable, *as always*, only at the price of throwing away some information (by using a *Statistician's Stooge* who knows the random numbers but does not disclose them). But the use of sampling without randomization involves the pure Bayesian in such difficult judgments that, at least if he is at all Doogian, he might decide, by Type II rationality, to use random sampling to save time. A Cornfield (1968/70, p. 108) points out, this can be covered within the Bayesian framework.

Since this conference is concerned with physics as well as with statistics I should like to mention a connection between something I have been saying and a point that is of interest in the philosophy of physics. (This point is also discussed in #815.)

When discussing the probability that the millionth digit of π is a 7, I could have pointed out that similar statements can be made about pseudorandom numbers. These are deterministic sequences that are complicated enough so that they appear random at last superficially. It would be easy to make them so complicated that it would be practically impossible to find the law of generation when you do know it. Pseudorandom numbers are of value in computer applications of the so-called Monte Carlo method. They are better than random