TEDDY SEIDENFELD

WHY I AM NOT AN OBJECTIVE BAYESIAN; SOME REFLECTIONS PROMPTED BY ROSENKRANTZ

I

Roger Rosenkrantz's recent work* gives the reader snapshots, taken through the eye of a Bayesian camera, of one philosopher's vision that major problems in the philosophy of science can be solved by the machinery of contemporary probability theory and statistics. It is my intent here to review the philosophical foundations of Rosenkrantz's project, but first let me give an overview of his book.

The text is divided into three principal sections: *Informative Inference; Scientific Method;* and *Statistical Decision.* The first, and in my opinion most important part of the book provides the basis for the rest of the work. Quoting from the dust jacket, I can summarize Rosenkrantz's positon by stating that, "The author holds that men in possession of the same data should agree in their probabilities, arguing that objective posterior distributions are obtainable by conditionalizing on an 'informationless' prior". It is this objective Bayesianism which is the subject of my examination in II.

The second part of the book has *simplicity* as its subject and Rosenkrantz focuses on his own measure of simplicity of a theory. He argues that this measure gives a (quantitative) scale for weighing the complexity of a theory against its supporting data, and that this comparison has value for judging the merits of one theory against another, i.e., it is an inter-theoretic measure as well. The final third of *Inference, Method and Decision* provides a somewhat hurried tour of a sample of topics which are part of the contemporary debate between Bayesians and non-Bayesians. Rosenkrantz reviews the standing of familiar orthodox statistical techniques and assesses their validity subject to the inductive principles set out in the earlier parts of the book. A closing chapter touches on acceptance and cognitive decisions.

Since it is my purpose in this article to review the major theme of the book, my emphasis will be on the opening section: *Informative Inference.*

Theory and Decision 11 (1979) 413–440. 0040–5833/79/0114–0413$02.80.

Let me summarize my findings. I claim that the twin inductive principles which form the core of objective Bayesianism are unacceptable. *Invariance* (due to H. Jeffreys) and the rule of *maximum entropy* (due to E. Jaynes) are each incompatible with conditionalization (Bayes' theorem). I argue that the former principle leads to inconsistent representations of 'ignorance', i.e., so-called informationless priors generated by the invariance principle are at odds with Bayes' theorem. I claim that Jaynes' rule of maximizing the entropy of a distribution to represent 'partial information' is likewise unacceptable. It leads to precise probability distributions that are excessively aprioristic, containing more information than the evidence generating them allows. Again, the conflict is with Bayes' theorem.

Before giving the supporting arguments to these claims, I wish to express my opinion about the latter parts of this book. I am troubled by Professor Rosenkrantz's treatment of simplicity. It is not just that his measure (average likelihood) produces numbers which do not square with my intuitions about simplicity (after all, simplicity is far from an intuitive notion), but I am unable to follow the arguments leading up to this choice of measure.

In particular, I do not understand Rosenkrantz' concept of 'sample coverage', which determines the range of possible experimental outcomes that support the theory in question. Loosely, sample coverage is thought to be inversely related to the complexity of a theory, e.g., a composite statistical hypothesis with an 'adjustable' parameter is thought to be more complex than an instance of it, that is more complex than a simple statistical hypothesis obtained by fixing the 'adjustable' parameter. But what is the sense of saying that the composite hypothesis has greater sample coverage?[1] How is it that the (simple) hypothesis that this is a *fair* coin has less sample coverage than the (composite) hypothesis that the coin is of unknown bias? In a long run of, say, $10^{10^{10}}$ flips any possible outcome is consistent with either posit.[2]

Of course, in a comparison between the two hypotheses based on a particular experimental outcome, the average likelihood of the composite hypothesis is an important quantity since, assuming the Bayesian principle of conditionalization (discussed in detail in the next section), we see that it summarizes the effect of the evidence on the belief state about the hypothesis.[3] However, I do not understand how the average likelihood reports the trade-off between the complexity of the theory and its coverage because: (i) I fail to see exactly what Rosenkrantz has in mind by 'sample coverage';

and (ii) the average likelihood is based on a *particular* experimental outcome, not on the *set of possible outcomes* (which seems to be the crucial ingredient in a discussion of coverage).

The final third of the book: *Statistical Decision* contains a wealth of numerical applications which introduce the reader to an impressive range of statistical and philosophical problems. Moreover, the pedagogical technique of showing applications and illustrations breathes life in the subject by giving us an understanding-through-doing instead of mere theoretical accounts of the differences in alternative approaches. The analysis is not without faults, however. A case in point is the discussion of the 'optional stopping' problem, an historically important topic which served as one of the first examples of serious conflict between Bayesian and orthodox statistics. Rosenkrantz attempts a novel interpretation which compromises the traditional Bayesian solution. Though his motive for reconciliation is noble indeed, a technical blunder reduces the import of Rosenkrantz's claims which, I feel, must be withdrawn.[4]

This book has wide scope. Professor Rosenkrantz's goal is the application of a version of Bayesianism, objective Bayesianism, to philosophy of science. His pursuit is aggressive and sincere; yet I criticize him because he has built his project on quicksand and I think he must share the responsibility for the adequacy of the foundations. It is to the discussion of the foundations of objective Bayesianism that I now turn.

II

There are three postulates which form the heart of all Bayesian programs. First is that an agents' beliefs (at a time) are represented by some probability function, p( ). This is usually called the *coherence* principle. Second is that the agent's commitments to changes in belief when new evidence is acquired are governed by Bayes' theorem. In particular, if $\mathrm{p}_K(\;)$ represents the current state (where 'K' denotes the consistent, deductively closed knowledge base of the state) and if $\mathrm{p}_{K^*}(\;)$ represents the hypothetical belief state arrived at by adding to K the new evidence d (that is, if 'K^*' denotes the deductive consequences of K & d), then $\mathrm{p}_{K^*}(\;) = \mathrm{p}_K(\;/d)$, which by Bayes' theorem satisfies:

$$\mathrm{p}_K(\;/d) \propto \mathrm{p}_K(d/\;) \cdot \mathrm{p}_K(\;).$$

Since the updated beliefs are represented by the original probability function conditionalized on the new evidence, i.e., $\mathrm{p}_{K^*}(\ ) = \mathrm{p}_K(\ /d)$, the principle is called *conditionalization*. Because of the central role Bayes' theorem plays in regulating conditional probabilities the program takes its name from this theorem. Finally, the *total evidence* principle assures that the probability function $\mathrm{p}_K(\ )$ takes all (and only) the agent's knowledge K for its base.

For the purposes of this discussion we may contrast two extremes of Bayesianism. At one pole is subjectivism (as defended by Savage and de Finetti), which insists that the inductive logic has *no* additional postulates limiting the acceptable belief states.[5] Any probability function satisfying these three postulates is admissible and the corresponding belief state reasonable. At the other pole is objectivism (as defended by Jeffreys and Jaynes), which argues for a *uniquely* admissible probability function, $\mathrm{p}_K(\ )$, given a knowledge state K.[6] The probability function is *objective* because it is agent-invariant once the knowledge base K is fixed. Of course, objective Bayesianism requires extra inductive postulates to identify the uniquely admissible probability function, $\mathrm{p}_K(\ )$, and our job is to investigate the compatibility of the additional inductive principles with the original three (above).

Before rushing to this task, let us consider whether the fundamental assumption of objective Bayesianism is philosophically plausible. There are, I find, two motivating sources of objectivism. First, if Bayesianism is to make sense of traditional statistical inference then only specific 'ignorance' probabilities will work. For example, if h is some interval of simple statistical hypotheses, i.e., if h is a composite interval hypothesis, then for a Bayesian to translate a traditional 95% confidence interval about h, given typical data d, into a posterior probability of the kind $\mathrm{p}(h/d) = 0.95$, then a *privileged* 'ignorance' probability $\mathrm{p}(h)$ is required. This 'informationless' probability captures the ignorance, prior to observing the evidence d, the agent professes about the truth of h. Of course, in setting out to reconstruct rational science with Bayesian tools, Rosenkrantz is committed to reconstructing a fair part of traditional statistical inference as well.

Second, for these Bayesians who recognize a non-epistemic, observer independent probability, call it *chance*, special conditional probability (where knowledge of chances is part of the corpus of knowledge) for corresponding observable random variables is invariant. For example, given that this is a *fair* coin (a chance statement), the probability is 0.5 that the next flips lands head-up. This instance of a simple *direct* probability shows how the knowledge base may fix the admissible belief state through relations other than mere deductive consequences. As we note shortly, each of these heuristics is central to the constructive strategy employed by the objectivists for determining precisely which probability is *the* admissible one, and in particular for determining 'informationless' probability.

As early as 1939 Sir Harold Jeffreys argued for objectivism.[7] He proposed a major advance over his 1939 formulation in his 1948 theory of Invariants.[8] Let me review Jeffreys' program since it is adopted (with only minor alteration) by Jaynes and Rosenkrantz. Suppose we are interested in some quantity m which, given our current knowledge, is limited to some continuous range M of real values. We parameterize this set M of possibilities by choosing some sufficiently smooth real-valued function, f, of m, with range Θ, i.e., $f(m) = \theta \in \Theta$; where θ becomes the parameter of interest and Θ its associated parameter space of possible values. For instance f may be the trivial function which maps the quantity m into its dimension-free magnitude. How are we to represent our initial ignorance about m in a probability function if all we know is that $\theta \in \Theta$?

Jeffreys' earliest solution was to consider the parameter space Θ and to fix an 'ignorance' distribution for θ based on simple mathematical properties of Θ.[9] For example, if the parameter space is the whole real line, $(-\infty, +\infty)$, then an (improper) uniform distribution is adopted.[10] If the parameter space is the positive half of the real line, $(0, +\infty)$, then an (improper) density proportional to $1/\theta$ is adopted. Jeffreys argued, aprioristically, that the advantage of these selections is that they are consistent over an important family of alternative parameterizations. Thus, the rule for positive parameters leads to the same ignorance distribution regardless of which parametrization in the family θ^r (r a real) is chosen for the function f. Similarly, the rule for real-valued parameters (the uniform distribution) leads to the same probability function representing ignorance for any linear transformation of the parameter. Moreover, the two rules are mutually consistent when a positive-valued parameter θ is transformed to a real-valued one by taking logarithms. That is an (improper) density proportional to $1/\theta$ is equivalent to an (improper) distribution uniform over $\ln(\theta)$, for $\theta > 0$.

Certainly these minimal consistency results fail to justify Jeffreys' two rules for picking 'informationless' prior probability functions. Nor do they solve the more realistic problems that arise when M is bounded at both termini, or when M is more than one-dimensional. For instance, if m is limited to the unit interval (0, 1), are we to use the positive parameter rule by transforming to $\theta = m/(1-m)$, or the simpler uniform distribution rule leading to the (proper) density function dm, or some other transformation taking a real-valued parameterization? Equally problematic is the situation where we learn that a real-valued parameter is limited to the positive half of its parameter space. Then do we use the (improper) conditional distribution obtained by truncating the (improper) uniform distribution at the origin, or do we shift to the positive-valued parameter rule?

Jeffreys' response (prior to the development of Invariants) was to suggest that the dimensions associated with m (which usually is some physical quantity, not just a number) contain relevant information for determining the correct family of alternative parameterizations which are to count as equivalent, hence the dimensions contain information for selecting the right ignorance distribution. I think it is fair to say, however, that the convincing justification for Jeffreys' rules came with the applications he offered. By showing that many classic statistical tests had Bayesian models only when his rules were used to fix the prior 'ignorance' probability, Jeffreys served notice that his objective Bayesianism had the resources for reconstructing current statistical practice from, what he argued was, a philosophically sound base. His successes were especially noteworthy in cases of location or scale parameterization, e.g. estimation of a mean or variance from a normal distribution, and his account of fiducial inference remains a standard for comparison.[11]

A major advance in Jeffreys' program was his theory of Invariants.[12] The innovation was to consider not just the quantity of interest, m, but also to consider the statistical distribution of the observable random variable which is to provide the information about m. That is, to specify an 'ignorance' distribution about m one must take into account the statistical model for the data which are to be the evidence acquired. Typically the factor of interest, m, parameterizes this statistical model into *simple* statistical hypotheses, i.e., precise *direct* probabilities. For example, we may be interested in the location of the center of mass of a bent coin. Our experiment is to flip the coin and record the outcomes: heads-up or tails-up. Based on what we know about the flipping process we are prepared to accept a binomial statistical model to model the flipping process, and we accept a functional relation between the binomial parameter (of the model) and the factor of interest: the center of mass of the coin. That is, the magnitude of the binomial parameter corresponds to the location of the center of mass of the coin. Making use of the luxury of *chance*, we see that a posit about the center of mass is equivalent to an hypothesis about the chance of the coin's landing heads-up. By a simple direct inference we discover observer invariant probability for the distribution of the observed random variable: the outcome of the flip. For instance, the posit that the center of mass is the geometric center of the coin may correspond to the *chance* statement that the coin is *fair*; so given this simple statistical hypothesis, by direct inference the objective probability is 0.5 that the next flip lands heads (tails)-up. Thus, by an appeal to the data-to-be-acquired, Jeffreys establishes contact with the second heuristic argument (of page 416).

Invariants are mathematical measures that affix to statistical distributions. Their importance for Jeffreys' project is that they provide the ingredients in the *invariance* principle: a rule for identifying 'ignorance' distributions. As the name suggests, invariants are invariant over 1–1 differentiable transformations of the parameters or random variables appearing in a statistical distribution. Thus, once the factor of interest becomes tied to the data-to-be-acquired by a statistical model, it does not matter to the invariance principle which of the alternative but equivalent parameterizations of this model is chosen.

Applying invariance results in an 'ignorance' distribution that represents ignorance about the factor of interest, based on a particular source of information captured in the statistical model. I list a few basic applications of Jeffreys' invariance rule:[13]

(a) With a binomial parameter θ and unit interval parameter space, the Jeffreys' 'informationless' prior probability is given by the density

$$\mathrm{pd}(\theta) \propto [1/\pi\sqrt{\theta(1-\theta)}]\,d\theta;^{14}$$

(b) With a location parameter μ and parameter space consisting of the whole real line, the Jeffreys' 'informationless' prior probability is given by the (improper) uniform density $\mathrm{pd}(\mu) \propto d\mu$;

(c) With a scale parameter σ and parameter space consisting of the positive half of the real line, the Jeffreys' 'informationless' prior probability is given by the (improper) density

$$\mathrm{pd}(\sigma) \propto d\sigma/\sigma.$$

Thus, the earlier successes Jeffreys' scored with his simplified rules for generating 'ignorance' distributions are preserved by the invariance principle, and the first of our heuristic arguments (page 416) applies.[15]

Following Jaynes, Rosenkrantz adopts a modified invariance principle to identify 'informationless' probability.[16] Jeffreys' rule is altered by Jaynes as follows: instead of using the mathematical invariants to define an 'ignorance' distribution, he recognizes particular transformations of the random variable and parameter in the statistical model which are *experimentally meaningful* in the process for obtaining the data, and then he finds a rule for picking an 'ignorance' distribution that leads to the same probability function no matter which parameterization, from among the particular transformations deemed experimentally meaningful, is chosen. That is, the invariance is over parameterizations that are supposed to represent mathematically and *empirically* equivalent formulations of the experimental process captured in the statistical model.[17]

For example, our experiment may consist of weighing an object O on a scale S_b whose readings (on separate weighings) are approximately normally distributed about the true weight w_O of O plus the bias factor b (for scale S_b), with a known variance of, say, 1 unit. That is, repeated measurements of O on S_b are modelled by the statistical distribution which is $N(\theta_b, 1)$; where $\theta_b = w_O + b$. We suppose that the experimenter knows the bias factor for any given scale, but does not know the weight of O. What is the 'ignorance' distribution for the parameter θ_b?

In order to apply the modified invariance rule we examine data that might be obtained by weighing O on scales of differing biases. Imagine the experimenter uses an unbiased scale, $b = 0$, and reads a measurement w. On a scale with a bias $b' = 5$ this measurement corresponds to a reading of $w + 5$. Invariance requires:

(i) that we have a rule for obtaining an 'ignorance' distribution for any parameter θ_b, for any bias b, so that this distribution is *identical* for each b; and

(ii) an 'ignorance' distribution for each parameter θ_b so that with equivalent data, e.g., w with S_0 and $w + 5$ with S_5, the resulting probability distributions for the unknown weight w_O are the same. The solution for this problem is an (improper) uniform 'ignorance' with density:

$$\text{pd}(\theta_b) \propto d\theta;$$

which agrees with Jeffreys' invariance principle applied to a location parameter.

In my opening remarks I claimed that the invariance principle (in either form) is unacceptable because it conflicts with the second (of the three basic) Bayesian postulate(s): conditionalization. Let me outline the reasons for this statement, after which an illustration clinches the point.

The invariance principle, in any of its guises, is a restricted form of the ancient Laplacean principle of *insufficient reason*. That old saw runs: where one's knowledge merely limits uncertainty to a set of alternative possibilities, by symmetry, each is awarded equal probability. Equally old is the rebuke that different partitions of the same set of alternatives leads, by insufficient reason, to incompatible probability assignments. For example, in the case of a continuously variable parameter, θ, a uniform 'ignorance' distribution over θ is inconsistent with a uniform distribution over the re-parameterized θ^3.

Invariance looks to the data-to-be-acquired and finds a special family of parameterizations, picked out by mathematical symmetries in the statistical distribution for these data, to which insufficient reason can be applied consistently. As discussed before, we note that in a location parameter problem symmetries with respect to linear transformations lead to a uniform distribution over the alternatives parameterized in location form.

Now, conditionalization determines the commitments that arise when the knowledge base of a belief state is hypothetically enlarged by learning new facts. It is a simple point that conditionalization entails an invariance of posterior beliefs over alternative sequences of data acquisition. That is, with composite evidence, d_1 & d_2, by conditionalization it follows that conditioning first on d_1 and then on d_2 (in two steps), leads to the same distributions as are obtained when the data are accepted in the reverse order (or for that matter, when the data are accepted in one fell swoop). However, if the bits of evidence convey the outcomes of different experiments, and not merely repetition of the same experiment, application of the invariance rule may lead to violations of conditionalization because the symmetries of the first experimental process may, when fed through the invariance principle, lead to an 'ignorance' distribution that is inconsistent with the 'ignorance' distribution that follows if, instead, invariance is applied to the second experiment. Thus, depending upon which datum is considered first, the belief state after all the evidence is reported has two, precise but mutually incompatible probabilistic representations. Of course, it is an important part of scientific methodology that hypotheses be subjected to confirmation by a variety of tests. Thus, this inconsistency is a real, and not imagined danger.

An illustration is readily available to us. Suppose we want to investigate the unknown volume v of a hollow cube. Our procedures for testing are twofold. We may fill the cube with a liquid of known density, say 1 unit weight/unit volume, and then weigh this quantity of liquid on one of our familiar scales, say, the unbiassed one, $b = 0$. Then readings from this experiment are modelled by a simple statistical model, the normal distribution with mean v and unit variance: $N(v, 1)$. Invariance requires us to adopt the (improper) 'ignorance' distribution, uniform for weights of the liquid; hence, the result is an 'informationless' prior probability uniform over possible values of v, since the true weight of the liquid equals the magnitude of the unknown volume.[18]

The second experiment consists in cutting a rigid rod of known density, say 1 unit weight/unit length, to the length of an edge of the cube and then weighing the rod segment on the scale. Invariance applied to this datum leads to an 'ignorance' distribution which is uniform over possible weights of the rod; hence the result is an 'informationless' prior probability uniform over possible lengths of the cube's edge, since the true weight of the rod segment equals the magnitude of the unknown edge length. But the volume of the cube v is functionally related to its edge length l as $v = l^3$. So the invariance rule applied to the second experiment leads to an 'ignorance' distribution which is uniform over $\sqrt[3]{v}$. Depending upon which report is accepted first a different 'informationless' probability is used and, after all the evidence is reported, the upshot is a pair of distinct posterior probability functions concerning v.

The moral of this story is the invariance principle is *not* successful at avoiding the kind of inconsistencies that plague the naive principle of insufficient reason, even when rather sophisticated mathematical tools are employed to identify the 'relevant' symmetries. The fatal flaw is, I suggest, the requirement that ignorance is to be represented by some *precise* probability function. But since this is the basic tenet by which Jeffreys' Bayesian project becomes *objective* Bayesianism, my evaluation is that it is the entire program that founders and not just some incidental feature of the plan.[19]

Jaynes' and Rosenkrantz's work on developing objective Bayesianism does not stop with the modified version of Jeffreys' invariance rule. A second supporting beam (though one which rests on invariance in problems with continuously variable unknowns) is Jaynes' strategy for picking out 'partial information' distributions through an entropy based measure of uncertainty. That is, in rough outline, for problems where there exist appropriate statistical constraints on some unknown variable (constraints expressed as expected values of select functions of the unknown variable), Jaynes argues there is a *precise* probability function for that unknown variable which accurately represents the informational content of the constraints. In other terms, given just the information that constraints $c_1, \ldots, c_n$ hold of the (distribution for the) unknown factor x, then (if all goes well, mathematically) Jaynes' rule picks out that probability function which *maximizes the entropy* of the distribution for x, subject to $c_1, \ldots, c_n$, as the conditional probability $\mathrm{p}(x/c_1, \ldots, c_n)$.[20]

With a discrete variable x, and a finite sample space of m possible states, the entropy H associated with a distribution $\mathrm{p}[x]$ is given by the formula:

$$\mathrm{H}(\mathrm{p}[x]) = -\sum_i \{\mathrm{p}_i[x] \log(\mathrm{p}_i[x])\}, \quad 1 \leqslant i \leqslant m.$$

As many readers will notice, H is Shannon's measure of uncertainty of a distribution (from communications theory), or the familiar entropy measure (from statistical mechanics). Jaynes' rule, then, is to select that distribution $\mathrm{p}^*[x]$ (if one exists) which maximizes $\mathrm{H}(\mathrm{p}[x])$ subject to the constraints $c_1, \ldots, c_n$, and to call this the *objective* probability for x, given the (partial information) $c_1, \ldots, c_n$.[21]

Shannon's measure is elegantly characterized by three plausible properties of uncertainty:

S_1. H is a continuous function of the p_i's.

S_2. When $\mathrm{p}_i[x] = 1/n$ (all i), H is monotonically increasing in n, the sample space.

S_3. H is additive under decomposition of the sample space. That is, if the set of n possible outcomes with distribution p[x] is partitioned into m groups, $m \leqslant n$, with a derived probability distribution q[x] over the m states, then the original uncertainty H(p) equals the sum of (i) the uncertainty over the m states H(q), and (ii) the weighted uncertainty (with weightings $q_j[x]$ for the jth of the m partitions) for an outcome within any of the m groups, i.e., $\Sigma_j\, q_j[x] \cdot$ H(p[x/jth state]), $1 \leqslant j \leqslant m$.[22]

Maximizing uncertainty by maximizing entropy leads to a version of the principle of insufficient reason. For example, the maximum entropy distribution for a sample space of n possible outcomes is merely the uniform distribution, $p^*[x] = 1/n$. To repeat a useful illustration, given by Rosenkrantz (p 57), imagine we are faced with a six sided die, of unknown bias, and we must identify the probability function that represents our beliefs about the outcome of a roll of the die. Insufficient reason stipulates the simple, uniform distribution: p(side$_i$) = 1/6, $1 \leqslant i \leqslant 6$. If we add the constraint that the expected value of a roll of the die is 3.5 (the average of the six outcomes), Jaynes' rule for maximizing entropy also leads to the simple uniform distribution over the sample space $\{1, \ldots, 6\}$.[23]

The generalization of Jaynes' rule to problems with continuous distributions (and probability densities pd[ ]) is not at all trivial and points out the primacy of the invariance rule. The straightforward move to extend the uncertainty equation to:

$$\mathrm{H}^*(\mathrm{pd}[x]) = -\int \mathrm{pd}[x] \log(\mathrm{pd}[x])\, dx.$$

is unsatisfactory since H* is not stable under trivial changes of variables $x \rightarrow y = f(x)$, when y is an equivalent random variable, i.e., when the transformation f is 1–1 and smooth. Jaynes' answer is to derive a measure of uncertainty, H_c, for continuous distributions, which is relativized to an 'ignorance' distribution m[ ] obtained by invariance:[24]

$$\mathrm{H}_c(\mathrm{pd}[x]) = -\int \mathrm{pd}[x] \log(\mathrm{pd}[x]/\mathrm{m}[x])\, dx.$$

Because (pd[x]/m[x]) is stable over just those transformations which are thought to be mathematically and evidentially relevant to the problem at hand (remember that m[ ] is obtained by the invariance principle), H_c (unlike

H*) satisfies important consistency requirements subject to transformations of the random variable x, when those transformations are members of the privileged set of experimentally appropriate ones.

When defending Jaynes' entropy rule, Rosenkrantz acknowledges several objections (p. 77), to which he supplies partial responses and concludes, "The methods . . . which I have sketched seem fully capable of handling the difficulties traditionally raised" (p. 81). I believe that his own challenges to his position, as presented in the text, fail to strike at the most serious fault of the program. In advocating precise probability functions, either through the invariance principle or Jaynes' entropy rule (based on invariance), the objective Bayesian is driven to violate his own principle of conditionalization. We have seen (above) how invariance fails conditionalization. Next, I want to show that Jaynes' strategy for using entropy as a measure of uncertainty, which is to be maximized subject to known constraints, fares no better.

I suspect that an evenhanded evaluation of objective Bayesianism must concede, at a minimum, that it leads to surprisingly 'informative' distributions based on 'limited' evidence. For instance, from the impoverished 'data' that a random variable x has an expected value of 3.5 over a sample space of six possible outcomes: $S = \{1, \ldots, 6\}$, the entropy rule dictates the uniform distribution, with constant value 1/6th for each member of S. If x models the outcome of a roll of a die, then (subject to the information given) the entropy rule directs us to act as though we believe the die to be *fair*. But this selection is one of a continuum of distributions consistent with the two constraints: (a) a sample space S, and (b) an expectation of 3.5 over S. Even if we insist on a distribution symmetric about the expected value (thereby paralleling Carnap's symmetry requirement), there remains a continuum of solutions.

The *a priorism* of the solutions is just as striking in problems with continuous distributions, where H_c is the uncertainty measure to be maximized. Let θ be the parameter of interest and let us assume an (improper) 'ignorance' distribution uniform over the real line, $(-\infty, +\infty)$, as might be derived by the invariance principle. Hence, $m[\theta]$ is a constant and the entropy of the probability density for θ, $pd[\theta]$ is:

$$[*] \qquad H_c(pd[\theta]) = -\int pd[\theta] \log(pd[\theta])\, d\theta.$$

Amazingly, this measure has a unique maximum subject to the following three constraints.

Theorem:[25]

(i) Let pd[θ] be defined over the whole parameter space ($-\infty, +\infty$) and let it be a proper density, i.e.,

$$\int_{-\infty}^{+\infty} \mathrm{pd}[\theta]\ d\theta = 1.$$

(ii) Let pd[θ] have a first moment, i.e.,

$$\int_{-\infty}^{+\infty} \theta \mathrm{pd}[\theta]\ d\theta = \mu.$$

(iii) Let pd[θ] have a second moment, i.e.,

$$\int_{-\infty}^{+\infty} (\theta - \mu)^2 \mathrm{pd}[\theta]\ d\theta = \sigma^2.$$

Then [*] is maximized by the normal distribution with mean μ and variance σ^2, i.e., the N(μ, σ^2) distribution.

Thus, if invariance selects the uniform 'ignorance' distribution for the factor θ, and additional evidence E amounts to the constraints (i)–(iii) on the distribution pd[θ], then Jaynes' rule to maximize the entropy H_c picks the N(μ, σ^2) distribution as *the* probability function representing beliefs about θ, given E.

My claim is that the entropy rule is unsatisfactory since it directs us to act as though we had more information than in fact we do. To follow Jaynes' rule we must be prepared to violate conditionalization. I offer the following fictional account of an exchange between two Bayesians, J. (a Jaynesian) and B. (a non-objective-Bayesian) as my analysis of this claim. The dialogue opens with a brief discussion of a Bayesian technique (the device of imaginary experiments, due, I believe, to I. J. Good) that fills the gap for *non*-objective-Bayesians who must look outside the postulates of their inductive logic for an answer to the question of how to completely identify an agent's belief state once his knowledge base is determined. There follows my criticism of the maximum entropy rule.

Our story takes place in the dining room of the renowned consulting firm

Significant Significance Inc., whose motto "It's safe to accept when we don't reject!" (printed boldly at the top of company stationery) is a strong reassurance to the many satisfied clients. Two colleagues, B. and J. meet for lunch and . . .

B. Congratulations *J*! I saw your long article in the latest issue of the *JOURNAL*. I hear you excited all the top administrators with your new 'maximizing entropy' techniques. Is is true that you're to give us some lectures on it next month?

J. Well, I'm happy that at last the firm can adopt sound Bayesian methods, for now we can explain to the clients that the 'priors' are *objectively* valid. You remember all the fuss when one of the junior consultants told the client that our analysis (I think it was on the purity of some chemicals) might not stand up in a courtroom test because the opposition could show that by using a different prior to represent *their* opinion about some nuisance factor (I think it was a question of the age of some containers), the very data we obtained would support their case. We don't have to worry about that anymore! My method determines just which probability is objective.

B. And, if I understand your article, the techniques appeal to factual matters only.

J. Yes. That's the importance of it all. Once the problem has been correctly formulated (and here I'm thinking of the requirements for using the invariance principle), there is nothing to debate except the accuracy of some calculations. No more need of those old fashioned, awkward 'imaginary experiments' methods for extracting a representation of the client's beliefs.[26]

Do you remember those hours spent explaining coherent/incoherent betting systems to the poor chaps? I'm surprised they all didn't demand their money back when we tried to show them that we could identify a probability function for representing their current beliefs by asking a battery of questions like, "Would you bet 2:1 that such-and-such happens if the imaginary experiment turns out so-and-so?" By the time we finished I wonder how many thought the data printed in our reports were just the hypothetical outcomes of the imaginary experiments?

B. I remember the case where the client called us back after we sent him our analysis and told us that once he saw the actual experimental outcomes he realized that he hadn't given us the right answers to the questions about his views on the imaginary experiments. Of course, he wanted to change his responses and have us 'correct' our report. You can guess the to-do that followed when I told him that that couldn't be done. There was quite a row until he understood the point.

J. It is hard to see why, after learning the actual results, you can't remember your prior opinions. And those who understand that point still have to be shown why we can't just substitute their current opinions instead.

B. That would amount to using the same experimental data twice: once in the prior probability as part of their current beliefs, and then again as the outcome to be added in by conditionalizing. But that reminds me *J.*, I had a question to ask about your paper. I couldn't find a theorem to the effect that maximizing entropy and conditionalization always lead to the same posterior probability, no matter which procedure is used. For instance, if I calculate a distribution p($/d_1$) by your rule (where d_1 reports a constraint) and then use conditionalization to arrive at a final distribution p($/d_1$ & d_2), updating my beliefs to respond to the new evidence d_2, will that agree with result of applying your entropy rule to the composite evidence d_1 & d_2? That is, will the maximum entropy distribution, subject to constraints d_1 & d_2, be the same as the posterior probability p($/d_1$ & d_2)?

J. I don't think the two procedures can conflict in the way you suggest they might. Anyway, I haven't an example where they do. You see *B.*, the reason is that my entropy rule is designed to be used with information that *isn't* appropriate for conditionalization; that is, it can be used with information that doesn't fit Bayes' theorem. For instance, how can you use Bayes' theorem with evidence that, say, a distribution has a first and second moment? What sort of prior probability is there that a distribution has a second moment?[27] Information like that is akin to specifying a statistical model, not like data from an experiment. It is the sort of background evidence that you might use to determine the nature of an experiment, not the kind of evidence you could call 'observable', or assign a probability to for conditioning. That's why I say my method works with 'partial information'.

If you look at my paper, though, you'll find that I gave an illustration where, by suitably rewording the conditioning events, my entropy rule led to the same posterior probability as the standard conditionalization principle.

B. I can't remember that now, but please rehearse it for me.

J. The example I gave is very simple.[28] Imagine we are told that a die is loaded so that the expected value of a roll is 4.0, instead of the usual 3.5 for a *fair* die. The maximum entropy rule can be used to obtain the following distribution over the six outcomes:

outcome	1	2	3	4	5	6
probability	0.103	0.123	0.146	0.174	0.205	0.247

Next, suppose we are told that the die has been rolled and landed with a side *other than* the one-spot showing. We may obtain a conditional distribution by using Bayes' theorem to renormalize the first distribution over the restricted sample space of five outcomes, yielding:

(**)	*outcome*	2	3	4	5	6
	probability	0.137	0.163	0.194	0.321	0.275

If, instead of reporting the evidence that a one-spot failed to show, I told you that the distribution is of a five-sided die, numbered 2 through 6, with an expectation of 4.344 (the expected value of the distribution (**)), then the maximum entropy distribution is just (**) again.

Obviously, I've substituted different versions of the evidence for each argument. For conditionalization the new datum is that the one-spot failed to show up, which is evidence that is observable. For maximum entropy the new constraints are the restricted sample space (1 is deleted) and the increase by 0.344 in expected value. This last condition is definitely *not* observable since the bias of the die is theoretical. So you see, *B.*, there is little reason to worry about the sort of conflict you posited. Nonetheless, it would be nice to have a result establishing the compatibility of the two rules.

B. We're old friends *J.*, and you certainly know how skeptical I am of programs that insist on particular distributions to represent ignorance. My worries extend to your methods of treating 'partial information', quite apart from your use of the invariance rule. Let me give you an example of how I suspect your program gets into trouble.

Last week I had an assignment to determine the weight of some moonrock; you remember we're involved in a study to determine the feasibility of lunar mining. I used the new scales we received from *True-Weight* manufacturers.

True-Weight made an error with the packing slips for the scales and jumbled the instrument numbers with the accuracy ratings, so all I really knew about the scale I used was that it was unbiassed with some normal error, σ^2. All of True-Weights scales share this feature.

J. Of course you knew *something* about the accuracy of the scale. It wasn't designed for weighing elephants! But I'll grant you that the 'ignorance' distribution you're about to suggest is a good one for approximating the belief state we face here.

J. Yes, I did work with the familiar 'ignorance' distribution, expecting to correct for the 'unknown' accuracy after a phone call to *True-Weight*. Anyway, your invariance principle leads to the very same 'informationless' probability function. That is, assuming the data-to-be-acquired are from a normal distribution with unknown mean μ (corresponding to the unknown weight of the sample of moonrock) and unknown variance σ^2 (corresponding to the unknown precision of the scale), we have a location/scale distribution. Invariance directs us to the (improper) uniform density $d\mu$ for μ and the (improper) $d\sigma/\sigma$ for σ and, since the factors are unrelated, the joint 'ignorance' density is merely the product $d\mu \cdot d\sigma/\sigma$. We are interested in μ and temporarily σ^2 is a nuisance parameter (which will be eliminated after the phone call to True-Weight when I learn the accuracy of the scale used).

J. Thus far the problem is quite ordinary. What difficulties did you encounter with the maximum entropy rule? I've used it in this situation very often myself.

B. Here's my question *J.* Let's imagine I told you that my posterior probability distribution for μ satisfies these three constraints:

(i) it is defined over the whole real line;

(ii) it has a first moment of 10.123; and

(iii) it has a second moment of 10^{-4}.

Can't I use your maximum entropy rule to obtain a 'partial information' distribution for μ, subject to (i)–(iii)?

J. Yes. In fact the solution is simple. Since the 'ignorance' distribution for μ is the (improper) uniform one (note: no other factors appear in the 'informationless' probability function for μ), subject to (i)–(iii), the 'partial information' distribution for μ is $N(10.123, 10^{-4})$: the normal distribution with first and second moments equal to 10.123 and 10^{-4}, respectively.

B. Next, *J.*, I'll tell you that my posterior distribution for μ is based on 10 separate (statistically independent) measurements of the moonrock sample, using the same scale manufactured by *True-Weight*. The average of the 10 readings is 10.123.[29]

J. Then, you're telling me that your posterior distribution for μ was obtained by conditionalization from the joint 'ignorance' distribution using data from $N(\mu, \sigma^2)$, and that the sample average $\bar{x}$ of these ten values is 10.123?

B. That's right. But there is one additional fact I've yet to tell you and which you must know before you have all the relevant evidence about μ. You see . . .

J. Don't tell me. Let's see whether I can anticipate your little surprise. Since the maximum entropy rule identifies the $N(10.123, 10^{-4})$ distribution as the 'partial information' probability function for μ, when you also inform me that your posterior distribution was obtained by conditionalizing on data from a $N(\mu, \sigma^2)$ distribution, specifically: you had ten observations with an average of 10.123, then I can tell you that the missing information on which your posterior probability is based is the value of the nuisance factor σ^2. You must have called *True-Weight* and they told you that the scale has an accuracy $\sigma^2 = 10^{-3}$.

Let me explain. The 'partial information' you gave me is enough to identify your posterior distribution for μ. That is, the three conditions (i)–(iii) determine your posterior probability for μ, if my maximum entropy rule is used. Then when you tell me that this posterior probability is derived by conditionalizing on data from a *normal* statistical model (10 observations, in fact), you fix the likelihood function for the data. Finally, if the sample average $\bar{x}$ equals 10.123, I can readily calculate that the normal posterior probability $N(10.123, 10^{-4})$ was obtained by Bayes' theorem starting from the joint 'ignorance' distribution if and only if the total evidence includes the missing nuisance value, i.e., the two methods agree just in case $\sigma^2 = 10^{-3}$.

So you see, *B.*, by relying on the mutual compatibility of the two procedures: maximum entropy and conditionalization, I am able to work backwards from your posterior distribution, knowing the prior 'informationless' probability and part of the evidence, to fill out the missing data that were available to you.

B. That is a pretty account, *J.*, but I still have a shock for you, which confirms my initial suspicions about your program. My posterior probability function for μ is *not* the N(10.123, 10^{-4}) distribution you calculated, but instead is given by the Students t-distribution, for $t = \sqrt{n}(\bar{x} - \mu)/s$, on 9 degrees of freedom and first and second moments related to the quantities I gave you. The missing information I was about to tell you is the sample variance of the 10 observations, s^2. The phone call to True-Weight wasn't completed until today and I just learned that the scale has a rating of only 5×10^{-3}.[30]

Let's see if we can figure out where your analysis went wrong. We began with the familiar 'ignorance' density $d\mu \cdot d\sigma/\sigma$, where μ is the parameter of interest and σ^2 is the nuisance factor. Then I told you that my posterior distribution for μ satisfied the three conditions (i)–(iii), which it did. You applied your entropy rule and identified this posterior probability as the normal probability function N(10.123, 10^{-4}). Finally, when I said that the evidence includes a sample of 10 observations from the statistical model N(μ, σ^2) and average $\bar{x} = 10.123$ you deduced the missing value for the nuisance factor, a precision of 10^{-3} for the scale.

J. That is all correct. Clearly, my rule selected the normal distribution as the 'partial information' distribution and that is why I arrived at a different posterior probability. There is no mystery here. You must have told me *less* about your posterior probability function than I need to fix it by maximum entropy considerations.

B. No, *J.* It's not that simple. Your 'partial information' distribution not only differed from my posterior probability, but it contained much more information than mine, even though you derived yours from only a part of the total evidence I had.

If the t-distribution (my posterior probability for t) is the 'partial information' solution based on *more* than the three conditions (i)–(iii) I gave (and I don't know how to identify Student's distribution by *adding* constraints to the three given), then for fixed first and second moments, your program ranks the normal distribution as *less* informative (greater uncertainty) than the t-distribution. But as you see, I can obtain the normal distribution from the

t-distribution, in this problem, by *adding* (not deleting) information, to wit: adding the value of the nuisance factor. That you were able to calculate a value for σ^2 shows this is so. Using conditionalization as the standard, we realize that in this problem the t-distribution is based on *less* information than the normal distribution, since I can move from the former to the latter by increasing evidence and, I would say, by decreasing uncertainty about μ. Your uncertainty measure is at fault here. It ranks the normal and t-distributions in reverse order of informational content as they would be ranked by conditionalization.

J. That sounds like a real problem. But how is it that the measure H_c conflicts with conditionalization like that?

B. I think the answer rests on your assumption that H_c provides a ranking for the 'informativeness' of any distribution, regardless of whether that distribution depends on other than the factor of interest. My posterior probability function for the parameter μ, given by the t-distribution, was obtained by conditioning the 'ignorance' distribution on the new data, i.e., the sample of 10 weighings, summarized by the jointly sufficient statistics $(\bar{x}, s^2)$. However, the t-distribution is a *marginal* probability function (from the joint posterior probability for both factors μ and σ^2) which involves the nuisance factor, σ^2, non-trivially. That is, my posterior probability function is not independent between μ and σ^2, even though the factors were independently represented in the 'informationless' prior probability function. Your maximum entropy rule treats the distribution for μ as one free of all nuisance parameters since it was free of them in the initial belief state. That is how the 'ignorance' density m[ ], which is constant, was determined and, therefore, how your rule led to the identification of the *normal* distribution. But by maximizing the entropy of a distribution, subject to constraints (i)–(iii), I neglect the uncertainty that results from the *interaction between* the uncertainty I hold about the factor of interest and the uncertainty I hold about the related nuisance factor. In effect, H_c disregards all but the parameter of interest. The upshot is a measure of uncertainty that commits the agent to more information than he/she may be entitled to. Conditionalization does not suppress concern about the nuisance factor and, as a result, the two procedures conflict.

J. Well, *B.*, that does seem to nail down the problem. I guess my project has some serious difficulties with several parameters. It's a shame, but I'll have to dust-off those questionaires for the imaginary experiments, at least until I come up with a response to your objection.

III. SUMMARY

The objective Bayesian program has as its fundamental tenet (in addition to the three Bayesian postulates) the requirement that, from a given knowledge base a particular probability function is uniquely appropriate. This amounts to fixing initial probabilities, based on relatively little information, because Bayes' theorem (conditionalization) then determines the posterior probabilities when the belief state is altered by enlarging the knowledge base. Moreover, in order to reconstruct orthodox statistical procedures within a Bayesian framework, only privileged 'ignorance' probability functions will work.

To serve all these ends objective Bayesianism seeks additional principles for specifying 'ignorance' and 'partial information' probabilities. H. Jeffreys' method of invariance (or Jaynes' modification thereof) is used to solve the former problem, and E. Jaynes' rule of maximizing entropy (subject to invariance for continuous distributions) has recently been thought to solve the latter. I have argued that neither policy is acceptable to a Bayesian since each is inconsistent with conditionalization. Invariance fails to give a consistent representation to the state of ignorance professed. The difficulties here parallel familiar weaknesses in the old Laplacean principle of insufficient reason. Maximizing entropy is unsatisfactory because the 'partial information' it works with fails to capture the effect of uncertainty about related nuisance factors. The result is a probability function that represents a state richer in empirical content than the belief state targeted for representation. Alternatively, by conditionalizing on information about a nuisance parameter one may move from a distribution of lower to higher entropy, despite the obvious *increase* in information available.

Each of these two complaints appear to me to be symptoms of the program's inability to formulate rules for picking privileged probability distributions that serve to represent ignorance or near ignorance. Certainly the methods advocated by Jeffreys, Jaynes and Rosenkrantz are mathematically convenient idealizations wherein specified distributions are elevated to the roles of 'ignorance' and 'partial information' distributions. But the cost that goes with the idealization is a violation of conditionalization, and if *that* is the ante that we must put up to back objective Bayesianism then I propose we look for a different candidate to earn our support.[31]

University of Pittsburgh *August, 1978*

NOTES

* Roger D. Rosenkrantz, *Inference, Method and Decision*, Reidel, Dordrecht, 1977.

[1] Rosenkrantz, pp. 93–94. On page 94 he writes,

> Now the one thing we always do when we pass from a special case to a parametric extension thereof is increase the sheer number of possible experimental findings which the theory can accommodate (by the lights of a given criterion of fit). We take this feature as definatory and measure simplicity by the paucity of possible experimental findings which the theory fits. More precisely, by the *sample coverage* of a theory *T* for an experiment *X*, I mean the chance probability that the outcome of the experiment will fit the theory, a criterion of fit being presupposed. The smaller its sample coverage over the range of contemplated experiments, the simpler the theory.

My worries about sample coverage include these problems.

(A) From a Bayesian perspective, tests of Goodness of Fit of a theory (a kind of Significance testing) are difficult to understand. See, for example, M. De Groot, 'Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio', *J.A.S.A.* 68, No. 344 (1973), 966–969; and S. Spielman, 'The Logic of Tests of Significance', *Philosophy of Science* 41, No. 3 (1974), 211–226.

(B) In testing a composite hypothesis against an *instance* of it, there is the question of how to represent the composite hypothesis so that the simple one *is* an instance of it. For example, with an 'adjustable' parameter, a Bayesian will be required to admit a distribution over the alternative values of this parameter. If the distribution has a *chance* or *statistical* probability basis, then the simple hypothesis may not be an instance of the composite hypothesis, but a contrary to it, e.g., this is a coin of unknown bias selected from an urn with a uniform distribution of biases, versus, this is a *fair* coin selected from an urn of fair coins. If the distribution for the 'adjustable' parameter is merely a credal probability, e.g., this is a coin of unknown bias with, say, a uniform 'ignorance' probability over possible biases (capturing the ignorance about the origins of the coin), then the simple statistical hypothesis, e.g., this is a fair coin, can't be evaluated against the composite hypothesis. Bayes' theorem prohibits a comparison since the alternatives represent different credal states, whose difference is not limited to a choice between empirical hypotheses, started from a common belief state.

(C) A measure of the paucity of experimental outcomes supportive of a theory presupposes a standard of possibility uncontaminated by any of the hypotheses tested. What is the nature of this standard and what is presumed by its existence?

[2] Not every parametric extension of a simple hypothesis increases the "sheer number of possible experimental findings which the theory can accommodate (by the lights of a given criterion of fit". Depending upon the exact distribution for the alternative biases of the coin, we see that outcomes supportive of the hypothesis that it is fair may *not* fit the composite hypothesis.

[3] That is, changes in the probability of the composite hypothesis H are governed by the average likelihood (in Rosenkrantz's sense of that concept, p. 97) if conditionalization holds.

[4] Rosenkrantz, pp. 196–200. The optional stopping question is: does the evaluation of competing statistical hypotheses depend on the rule used to terminate experimentation, once the observed outcome is fully reported? Since orthodox statistical techniques are very sensitive to changes in the sample space of possible outcomes (leaving fixed the outcome attained), whereas Bayesian inference depends only on the actual result (and not what else might have happened), the answer is affirmative for the orthodox statistician but negative for the Bayesian.

Savage's example (used by Rosenkrantz and taken from Savage, L. J. *et al.*, *The Foundations of Statistical Inference*, Methuen, London, 1962, p. 72) is a simple binomial sampling problem, with an unknown bias of either 3/4 or 1/4 for 'success'. The stopping rule proposed is one that takes the likelihood ratio between the two simple hypotheses as the deciding factor: stop sampling when we see three more successes than failures. Say this happens on the ninth trial. Is it plausible to dismiss the stopping rule and act as though we decided to draw nine times and stop no matter what the outcomes?

The Bayesian solution gives the same result in both cases. If we assume a simple prior probability of 0.5 for each hypothesis, then the posterior probability is in the ratio 27:1 that the process is loaded for 'success', i.e., the parameter equals 3/4.

Rosenkrantz objects to this analysis as being incomplete. His account is that if we 'condition upon' the possible outcomes of the experiment then we see that, no matter which result occurs, the posterior probability is the same as the prior probability, i.e., the experiment is irrelevant when the perverse stopping rule is used, p. 199.

The technical blunder committed by Rosenkrantz is to assume that the set of possible outcomes (up to probability zero) is the same for each hypothesis tested. In fact, with the perverse rule: stop when three more successes occur, there is a non-zero probability that the experiment will run forever. (The stopping problem is a simple random walk exercise. One formulation is as a discrete Markov process with absorbing barrier. In the general case, we have a one dimensional step, beginning at the origin, with a chance of p of moving in the positive direction and a chance of q of moving in the negative direction, and an absorbing barrier at $+n$. The probability that the process continues without absorption is: $(p/q)^n$ if $p \leqslant q$, and 0 otherwise.) Hence, in our version, there is a probability of $(1/3)^3 = 1/27$ that the experiment keeps on going, if the hypothesis that the process is loaded for 'failures' is true, i.e., if $p = 1/4$.

By conditioning upon the termination of the experiment, Rosenkrantz suppresses all the relevant data, since it then does *not* matter on which trial the process ends.

I mention this error only because it strikes me as a familiar one. For example, in the so-called Tram Car (or Tank) problem – where one sees a car numbered n and wonders how many N there are, the assumption being that the chance of observing car numbered n is $1/N$, $n \leqslant N$, and 0 otherwise – the usual analysis fails to take into account that a car has been observed, or the waiting time to the observation. When this part of the total evidence is included the anomalous conclusion that the likelihood is greatest for the hypothesis $N = n$ (and decreases monotonically for increasing N) is no longer correct. A family of related problems can be found in A. P. Dawid and J. M. Dickey, 'Likelihood and Bayesian Inference from Selectively Reported Data', *J.A.S.A.* 72, No. 360 (1977), 845–850.

[5] B. DeFinetti, 'Foresight: Its Logical Laws, Its Subjective Sources', (1937), reprinted in H. E. Kyburg and H. E. Smokler, *Studies in Subjective Probability*, John Wiley and Sons, Inc., New York, 1964. L. J. Savage, *The Foundations of Statistics*, 2nd ed., Dover Publications, Inc., New York, 1972.

[6] H. Jeffreys, *Theory of Probability*, Oxford University Press, Oxford, 1st ed. 1939, 3rd ed. 1961. An earlier work by Jeffreys also reveals his commitments to objective Bayesianism, *Scientific Inference*, Cambridge University Press, Cambridge, 1st ed. 1931, 3rd ed. 1973. Jaynes' work is found in a series of articles, some of which are E. T. Jaynes, 'New Engineering Applications of Information Theory', in J. L. Bogdanoff and F. Kozin (eds.), *Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability*, John Wiley and Sons, Inc., New York, 1963; 'Foundations of Probability Theory and Statistical Mechanics', in M. Bunge (ed.), *Studies in the Foundations Methodology and Philosophy of Science*, Vol. 1, Springer-Verlag, New York, 1967; 'Prior Probabilities', *IEEE Transactions on Systems Science and Cybernetics*, SSC-4, No. 3 (1968), 227–241.

[7] Jeffreys, *op. cit.*

[8] Jeffreys, *Theory of Probability*, 2nd ed., 1948. Also, see his 'An Invariant Form for the Prior Probability in Estimation Problems', *Proc. Roy Soc.* A 186 (1946), 453.

[9] Jeffreys, *Theory of Probability*, 1st ed., 1939, p. 97.

[10] Throughout this paper the question of how to avoid recent marginalization paradoxes, attributed to *improper* distributions, will be sidestepped. I choose to follow the spirit of A. Renyi's analysis, in which improper distributions are seen as limits of proper distributions and, following suggestions made by I. Levi, to respond to the recent paradoxes (and older ones as well) by recognizing the failure of countable additivity with improper distributions. The crucial link between additivity and the paradoxes is deFinetti's results on 'conglomerability', found in his *Probability Induction and Statistics*, John Wiley and Sons, London, 1972.

[11] Jeffreys' influence on Ian Hacking's noteworthy reconstruction of Fisher's fiducial argument is acknowledged in his, *Logic of Statistical Inference*, Cambridge University Press, Cambridge, 1965, pp. 140 and 147.

[12] An excellent statement of Jeffreys' theory of Invariants is found in V. S. Huzurbazar, *Sufficient Statistics*, Marcel Kekker, Inc., New York, 1976, especially part 1, 'Invariance Theory of Prior Probabilities'.

[13] Jeffreys, *op. cit.*, Chapter 3.

[14] This 'ignorance' distribution lies between the (improper) density $[1/\theta(1-\theta)]\,d\theta$, and the simple (Bayes) uniform density $d\theta$. For additional comments about Jeffreys' solution here see R. A. Fisher, *Statistical Methods and Scientific Inference*, 3rd ed., Hafner Press, New York, 1973, pp. 16–17.

[15] Actually, Jeffreys' rule does not work well with multiparameter distributions. He is, of course, aware of this difficulty. Huzurbazar has constructed a modified version of Jeffreys' invariance rule that avoids the multiparameter problems faced by Jeffreys, however, at the expense of uniqueness of his solutions. See Huzurbazar, *op. cit.*

[16] Rosenkrantz, pp. 62–77. Unlike his presentation, here I discuss invariance before Jaynes' entropy rule. Since, for one, Jaynes' policy of maximizing entropy presupposes an 'ignorance' distribution obtained by invariance methods, if the distribution involves continuous variables, I have ordered by presentation with the evaluation of invariance first.

[17] I do not find Jaynes' account of this notion of equivalent formulations perspicacious. In particular, I cannot find criteria for judging the claims about experimental equivalence. In contrast, I point out that D. A. S. Frazer's theory of group theoretic invariance does go a long way towards solving this problem. His answer is, in short, that transformations with group invariance properties determine the relevant family of equivalent experimental procedures. See his *The Structure of Inference*, John Wiley and Sons, New York, 1968.

[18] Since weights are non-negative, this distribution is truncated at the origin. That is, the 'informationless' prior probability function is the uniform (improper) distribution truncated at the origin. This aspect of the counterexample is inconsequential and can be eliminated if desired.

[19] I have relied on this same illustration to point out the incoherence of fiducial inference. Again, the problem is the representation of a state of ignorance. See my 'Direct Inference, Inverse Inference, and Bayes' Theorem', *Journal of Philosophy*, **LXXV** (1978), 709–730.

[20] For the result, as stated, the constraints *must* be interpreted as holding of statistical (or chance) probabilities about the variable x. If the constraints are on the *credal* probabilities about x then the resulting distribution cannot be thought of as a *conditional probability* (as written in the text). This is because, were the constraints on *credal* probabilities, the conditional probability would be conditional on a statement of another *credal* probability; in effect, the upshot would be to iterate probabilities of probabilities; and that is nonsense. (See deFinetti, *op. cit.*, pp. 189–193.) In such cases (and I presume Jaynes' and Rosenkrantz are anxious to use the entropy rule even when the constraints are on credal probabilities), the entropy rule operates at the meta-level – for it is in the meta-language that the constraints are formulated. The result would be a probability in the object language, but *not* one in the conditional mode: $p(x/c_1, \ldots, c_n)$.

[21] When no maximum exists, Rosenkrantz calls the problem 'overdetermined', e.g., inconsistent constraints, and when several solutions exist the problem is 'underdetermined', e.g., too few constraints. Examples of these problems are considered in T. Fine's, *Theories of Probability*, Academic Press, New York, 1973, p. 168. I shall not be concerned with this problem here.

[22] C. E. Shannon, *Bell System Tech. J.* 27 (1948), 623. See, also, A. Hobson and B.-K. Cheng, 'A Comparison of the Shannon and Kullback Information Measures', *J. Stat. Physics* 7, No. 4 (1973), 301–310. Condition S_3 is suggestive of the multiplication rule for probability. It falls short of Bayes' theorem, however, as shown in the text.

[23] There is extra information in the second version of this example. That is, according to Jaynes' prescription, the information about the magnitudes of the outcomes within the sample space, as well as the expected value of the variable over this set of possible outcomes, is significant. A simplified version of the entropy rule applies if all that is known is the cardinality of the sample space. If all that is known is that the sample space has six members, then maximum entropy dictates a uniform distribution. But with the added constraints, the problem is non-trivial. In this illustration, it turns out that the new information is irrelevant, i.e., the distribution is the same as if only the cardinality of the sample space is known.

An important mathematical technique for solving such maximization problems is by Lagrange Multipliers. See Courant and Hilbert, *Methods of Mathematical Physics*, vol. 1, Interscience Publishers, New York, 4th printing, 1963, pp. 164–174.

[24] It is worthwhile to note that H_c is better characterized as a measure of uncertainty in Kullback's theory, instead of the association with Shannon's theory. Following analysis due to A. Hobson (Hobson and Cheng, *op. cit.*) we see that Kullback's 1951 measure of information for discriminating between two distributions p^0 and p^1, given by:

$$I_k(p^1, p^0) = \sum_i p_i^1 [x] \log (p_i^1 [x] / p_i^0 [x]),$$

can be characterized by five simple properties, the first three of which parallel $S_1 - S_3$ of Shannon's theory. Moreover, Hobson shows that with discrete distributions Shannon's uncertainty H is definable in terms of Kullback's information I_k. If we interpret $I_k(p^1, p^0)$ as the *decrease* in uncertainty in going from p^0 to p^1, then $H(p^1)$ is just the difference between $I_k(p^*, p^0)$ and $I_k(p^1, p^0)$; where p^* is the minimum entropy distribution concentrated on a point of the sample space, and p^0 is the maximum entropy distribution uniform over the sample space of p^1. That is:

$$H(p^1) = I_k(p^*, p^0) - I_k(p^1, p^0).$$

That H_c is properly characterized by I_k is justified since:

(a) the difficulties with Shannon's measure for continuous distributions are avoided by Kullback's measure;

(b) Shannon's measure is definable in terms of Kullback's for the discrete case;

(c) Jaynes' H_c is Kullback's measure when m[ ] (of Jaynes' rule) is p^0 is the 'ignorance' distribution obtained by invariance. For an additional reference on information measures like Kullback's see Chapter IX (Appendix) to A. Renyi, *Probability Theory*, American Elsevier Publishing, New York, 1970.

Rosenkrantz's opening chapter, *Information*, provides an introduction to the topic of information measures. A point of concern, however, is the shift he makes from Shannon information to Fisher information (Section 5 of Chapter 1). I do not think the transition is as simple as the book suggests. In fact, I think the two topics are only distantly related. See, for example, the discussions by Kullback and Leibler, 'On Information and Sufficiency', *Annals of Math. Stat.* 22 (1951), 79–86; or G. Barnard, 'The Theory of Information', *J. Roy. Stat. Soc.* B13 (1951), 46–64 (with discussion).

Lastly, I thank Prof. F. Keffer (Physics, Univ. of Pittsburgh) for bringing Hobson's work to my attention.

[25] See, C. R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley and Sons, New York, 1965, p. 131; 2nd edition, 1973, pp. 162–163.

[26] I. J. Good, *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, The M.I.T. Press, Cambridge, 1965, pp. 19, 20, and 45.

[27] Where the constraints are statistical (as opposed to credal – see footnote 20) Bayes' theorem may be difficult to apply nonetheless. If the statistical constraints are general enough, the 'prior' probability required for Bayes' may be equivalent to a distribution over different statistical models. Of course, if the constraints are over credal probabilities, Bayes' rules does not apply.

In the discussion that follows, it is *irrelevant* which reading is taken for the maximum entropy rule: a rule for fixing conditional probabilities when the constraints are interpreted as holding of *chance* distributions; or a meta-rule for prescribing probabilities over object language variables, given constraints holding at the meta-language level.

[28] Rosenkrantz, pp. 57–58. As formulated in the text and in the book, this problem is easier to follow when the constraints are thought of as holding on statistical distributions for an observable random variable. See footnotes 20 and 27 for clarification.

[29] As formulated in the text, this problem is easier to follow when the constraints are thought of as holding on credal probabilities. *B.* reports constraints that hold of his posterior probability function for hypotheses about μ.

Since constructing my objection to Jaynes' rule, which the reader will find in the analysis given by character *B.*, I have been alerted to the recent work of A. Shimony and K. Friedman, 'Jaynes' Maximum Entropy Prescription and Probability Theory', *J. Stat. Physics* 3, No. 4 (1971), 381–384; and Shimony, 'Comment on the Interpretation of Inductive Probabilities', *J. Stat. Physics* 9, No. 2 (1973), 187–191 (for which I am indebted to I. Levi). It is my impression that the objection raised here is a variant of the criticism voiced by Friedman and Shimony.

[30] By the way, s^2 must equal 7.8×10^{-4}. The entropy, according to H_C, of the posterior distribution for μ will *increase* after learning the value of σ^2 if (at least) $\sigma^2 \geqslant 10^{-3}$.

[31] Two interesting alternatives are available. For those willing to part with conditionalization, a more reasonable treatment of ignorance can be found in H. E. Kyburg's program of epistemological probability; see his *The Logical Foundations of Statistical Inference*, Reidel, Boston, 1974. For others seeking a better characterization of ignorance than is found in the Jeffreys/Jaynes program but who (like myself) are unwilling to forgo conditionalization, see I. Levi's, 'On Indeterminate Probabilities', *J. of Phil.* LXXI, 13 (1974), 391–418.