Post-Bayesian Statistical Inference

(revised Summer 1981)

Arnold M. Faden

1. Introduction

The casual reader of literature on the foundation of statistics is struck with a curious anomaly. There is near consensus on the formal representation of the subject, and wild disagreement on the meaning of statements in it. A typical statement has the form $P_\theta(E) = X$, where E is a measurable set (an "event"), $X \in [0,1]$ and P_θ a probability operator indexed by a parameter θ.

To what in the real world does E refer? To events in the literal sense? To propositions? What about θ? And, these questions being resolved, what meaning attaches to the "mass" X that is assigned to E by P_θ?

We do not aim in this paper to review the voluminous literature on foundations. For our purposes it is sufficient to think in terms of two broad schools, called "bayesian" and "frequentist".

Our position is that all existing schools are inadequate, that their mutual criticisms are generally justified, and that the nature and purpose of foundational studies needs to be reexamined. In this paper our major goals are: (1) to outline the principles of a new approach, which we call "post-bayesian", and (2) to develop some specific statistical procedures that grow out of this approach. (The post-bayesian approach should lead ultimately to a revision of procedures across the entire spectrum of inference, so these examples should be looked upon as a sampler.)

2. Basic Principles

Statistics is the art of living successfully in an uncertain world. "Making inferences" is subordinate to this overall goal. Inferences are

guides for action. (Thus we side with Neyman in his long controversy with R. A. Fisher over "inductive behavior" vs. "inductive inference", and with Wald's decision-theoretic outlook.)

An adequate theory should be compatible with the way science develops (the accretions of "normal" science with occasional revolutions, the role of hypothesis. See Kuhn), and also with the way we accumulate knowledge casually in everyday life. It should also be "axiomatically satisfying", a vague desideratum including simplicity and straightforwardness.

The bayesian approach to statistical inference would be correct if people had unlimited and costless information-processing capacity. However, owing to human limitations, more-or-less serious departures from bayes are warranted. Hence the name "post-bayesian". Formally, the post-bayesian criterion for inference is to minimize expected loss (or cost). Thus it fits into the general framework of decision theory: inferences are "bayes decisions" with respect to some prior distribution. But major stress falls on two cost categories which hardly appear in the work of Wald or his successors.

a. First, the complexity costs associated with information processing: constructing models, gathering and storing data, solving models, communicating results, etc. (Some forms of complexity costs have been taken account of in the literature. Cost of observation enters into sample survey design, sequential analysis, selection of variables in regression analysis, etc. But the relative cost of alternative models themselves has been ignored*, and it is from this that the most radical consequences flow.)

---

*At least explicitly. As we discuss below, many, or even most, procedures of orthodox statistics may be interpreted as attempts to take account of this factor implicitly.

b. The second cost category is inaccuracy costs. We use the term "inaccuracy" in a special sense to be spelled out in detail below. Briefly, it refers to departures from the ideal pattern of bayesian inference, and here the meaning of "ideal" requires discussion. But first we conclude:

The best procedure is one minimizing the expected sum of complexity and inaccuracy costs.

3. Bayesian Inference and Models

The bayesian approach may be summarized as follows:

a. Probability statements assign numbers between 0 and 1 to propositions, the numbers representing in some sense the credibility of these propositions. The numbers assigned to a system of propositions satisfy the standard probability model. Specifically, propositions correspond to measurable sets, with the natural boolean correspondences (negation = complementation, conjunction = intersection, etc.)

A propositional range is a set of propositions that are exclusive and exhaustive (this corresponds to a measurable partition of sets). Quite often a range is indexed by a numerical parameter in a natural way, e.g. "The world population at Noon, 1 January 2000 A.D. is θ", $\theta = 0, 1, 2, \ldots$. Further, ranges often form indexed families in a natural way, e.g. the range above is one member of the family, "The world population at time t is θ," indexed by t real. It is convenient to represent the range by the parameter θ, and the family by an indexed set of parameters or random variables X_t*. Thus Prob $(X_t = 3)$ is the

---

*We use the terms parameter and random variable synonymously, though with the usual connotation that the former do not and the latter do, form a family. Note that the entire family is often thought of in orthodox statistics as constituting one random variable, a usage hung over from frequentist interpretations that we find inconvenient. Another connotation that we reject is that parameters are not observable and random variables are.

probability that the proposition indexed by the number 3 in the t - th range is true.

b. Having chosen the boolean algebra of propositions of interest, the investigator is to assign probabilities to these according to his "prior" intensity of beliefs.

c. If observation F is made, probabilities are to be modified by conditioning on F: For all proposition G,

$$\text{Prob}''(G) = \text{Prob}'(G \mid F)$$

where ', " denote probabilities prior and posterior to the observation F, respectively.

d. If an action is to be taken at time t, it should be chosen to minimize expected loss, taken with respect to one's probability assignment at t.

A statistical <u>model</u> is a collection of random variables $X_1, X_2, \ldots, X_m$, $Y_1, Y_2, \ldots, Y_n$ and a conditional probability measure

$$\text{Prob}\left[(Y_1, \ldots, Y_n) \in B \mid X_1, \ldots, X_m\right] . \qquad (\dagger)$$

The X's are the <u>exogenous</u> variables, the Y's are the <u>endogenous</u> variables.

Note: Any of the X_i's or Y_j's may be parameters or may be family-member ranges. It is even possible for some ranges in a family to be exogenous and others endogenous (e.g. initial values vs. later values in time-indexed families.)

No variables are intrinsically exogenous or endogenous (cf. Leamer).

The typical economic model does not look like (†) but consists mostly of a system of equations. This may be regarded as a convenient way of representing (†) if one keeps in mind that the distribution of the residual terms is an integral part of the model, and that this distribution itself typically

depends on further parameters not appearing in the equations (e.g. residual covariances).

Critique of "Bayesian" Practice

The preceding outline of ideal bayesian inference is never followed in practice, even by "bayesians".

1. Choice of prior. Consider the following simple and typical set-up. There is an unknown parameter θ and an observable family X_t, $t = 1, \ldots, T$. The conditional probability is given by a likelihood function in density form

$$p(X \mid \theta) = p(X_1 \mid \theta) \ldots p(X_T \mid \theta)$$

and this is completed by a prior on θ, $p(\theta)$.

Comments: All schools of thought use likelihood functions. Controversy rages over the additional specification $p(\theta)$. This is a good example of straining at a gnat and swallowing a camel. For the major objections to the prior $p(\theta)$--subjectivity and the like--apply just as well to the likelihood function (Basu); further, the really powerful assumptions generally reside in the likelihood function, in particular in the assumption of conditional independence of the X_t given θ. This is the assumption that makes the conclusions of a bayesian analysis insensitive to the choice of $p(\theta)$. Widely different p's are typically canceled out by one or two extra observations.

Actually, <u>both</u> the likelihood function and the θ-prior* in practice are chosen in a conventional simplified manner. This is well-recognized for $p(\theta)$,

---

*The prior in bayesian analysis should be over <u>all</u> variables (θ, X_1, ..., X_T), but in practice the word is reserved for the marginal on θ, factoring off the allegedly more objective likelihood function.

where one uses diffuse priors, conjugate priors and the like. It is a little harder to document for likelihood functions, but the following may be mentioned: for successive observations, the assumption of conditional independence, of identical distributions, of distributions drawn from some standard family, of conditional expectations linear in exogenous variables, etc. Each of these simplifications may of course be abandoned, but at any point there always exists a reserve of further realistic complications that could be incorporated.

There is a simple informal test for the presence of these simplifications. Ideal bayesian analysis never abandons models, but merely incorporates evidence by conditioning within its models. Hence the question: would you keep your model in the face of any new evidence whatever? Usually the answer will be "no", that for highly unusual observations, or those that seem to fall into an aberrant pattern, you will abandon the model. But this contradicts the principles of bayesian inference. cf. the role of hypothesis in science.

Our conclusion is that most "bayesian" inference is actually post-bayesian in that the priors and likelihood functions are not honest reflections of the researcher's beliefs, but merely a simplified approximation to these beliefs.

Theories of "bayesian" estimation and hypothesis testing have also been developed (Jeffreys, Lindley, Zellner). Here the departure from ideal bayesian inference is patent. Estimation or hypothesis rejection involve assigning probabilities of <u>zero</u> to propositions whose bayesian probabilities typically are <u>positive</u>. E.g. an interval estimate $A = [\theta_1, \theta_2]$ of a parameter, or an acceptance of the corresponding hypothesis, amounts to assigning $\Pr[\theta \notin A] = 0$.

Decision Theory in Post-Bayesian Terms

It is customary since Wald to think of inferences such as estimation or hypothesis rejection as decisions to which loss functions may be applied. This is good, but the lumping-together of diverse kinds of decisions obscures the interpretation of the costs involved. We distinguish

a. policy decisions (acts) - e.g. accepting a shipment, replacing a part, making an investment, selecting a drug, passing a student.

b. observational decisions - e.g. experimental designs, sampling methods, stopping rules.

c. cognitive decisions - e.g. estimating parameters, forecasting, hypothesis rejection.

The loss functions associated with (a) and (b) are well-understood, and discussed in books on cost-benefit analysis (Keeney and Raiffa, Decisions with Multiple Objectives: Drake, Keeney, and Morse, Analysis of Public Systems), sample surveys, experimental designs, etc.

But (c) is a different matter, and it is precisely the misinterpretation (or noninterpretation) of loss functions in this case that has led to setting up the decision problem in a misleading way.

The two new cost categories in (c) are complexity and inaccuracy costs.

Complexity costs include observational costs, already well-understood and coming under (b) above. But they include also model complexity costs, depending on such things as--number of variables and parameters, whether the model is deterministic or probabilistic, linear or nonlinear, whether observations must be processed or not, the number of qualitatively distinct hypotheses involved. All these factors influence the cost of model-construction, processing, recording and communication.

The importance of taking account of nonobservational complexity costs is illustrated by Lindley's theory of regression model selection (Lindley 1968). Adding additional regressors involves extra observational costs, but adding more complicated functions of the same regressors do not. Lindley considers only observational costs, and so cannot understand why polynomial regression of arbitrarily high degree should not be used.

Inaccuracy costs raise much deeper issues. Consider point estimation. Typically, loss functions here are of the form $L(\hat{\theta}, \theta)$, depending on estimate $\hat{\theta}$ and (unknown) true value θ.

But this is wrong. If bayesian inference is correct (absent complexity costs), then "inaccuracy" should refer to departures from bayesian inference. Now, at any time, in any observational state, knowledge of θ is summarized in a probability distribution $p(\theta)$. Inaccuracy arises if another distribution $p(\hat{\theta})$ is substituted for $p(\theta)$. Hence the loss function should have the form $L(\hat{p}, p)$, with pairs of distributions as arguments.

What is $L(\hat{p},p)$? Somewhere down the line we will make an inefficient policy decision from using $\hat{p}$ in place of p. The simplest case is a once-and-for-all immediate decision. Let $U(\delta,\theta)$ be utility; then

$$L(\hat{p},p) = \max_{\delta} \int U(\delta,\theta)\ p(d\theta) - \int U(\delta^1,\theta)\ p(d\theta),$$

where δ^1 maximizes $\int U(\delta,\theta)\ \hat{p}(d\theta)$.

(The trouble here, of course, apart from the gross over-simplication, is that calculating $L(\hat{p},p)$ requires using the complex p. In fact we can and must estimate L itself by various simplifying devices.)

We shall not discuss in detail the proper form of L. The following properties of $L(\hat{p},p)$ seem highly plausible.

1. $L \geq 0$.

2. $L(p,p) = 0$.

3. L is jointly continuous in say, the weak p-topology.

4. L increases as $\hat{p}$,p become "farther apart" in some sense.

5. L is convex in p.

Now among distributions the degenerate ones are especially simple, which suggests looking at losses of the form $L(\hat{\theta},p)$, $\hat{\theta}$ standing for the distribution

$$\Pr(\theta = \hat{\theta}) = 1$$

(We act "as if" we were sure that $\theta = \hat{\theta}$.)

The following example is instructive, because it shows how the cognitive decisions to test an hypothesis or not, or to estimate or not, can themselves be formally incorporated into the overall decision procedure and not set a priori.

Let θ be an unknown real-valued parameter, with distribution $p(\theta)$. (This distribution itself may be posterior to preceding observations. We assume here only that it is the best honest summary of evidence to date, if any.) The option of point-estimating refers to the replacement of p by a distribution degenerate at some point $\hat{\theta}$ to be chosen. We need the complexity and inaccuracy costs for this problem.

For <u>complexity</u> costs, assume for the moment that all degenerate $\hat{\theta}$ are equally complex. Let C be the extra complexity involved in carrying the full distribution p rather than some degenerate distribution.

For inaccuracy costs, a simple and not implausible formula is

$$L(\hat{\theta},p), = E_p(\hat{\theta} - \theta)^2 = \int (\hat{\theta} - \theta)^2 \, p(\theta) \, d\theta =$$

expected mean square deviation from the true value θ.

First conclusion: if a point estimate is made, it should be the mean:

$$\hat{\theta}_0 = E_p(\theta),$$

and then

$$L(\hat{\theta}_0, p) = \mathrm{Var}_p(\theta).$$

For, complexity costs are fixed, and the optimal $\hat{\theta}$ is then the one minimizing inaccuracy costs, which is well known to be $E_p(\theta)$.

Second conclusion: the full distribution should be kept if $C < \mathrm{Var}_p(\theta)$; the point estimate $\hat{\theta}_0$ should be made if $C > \mathrm{Var}_p(\theta)$. (If $C = \mathrm{Var}_p(\theta)$, the two decisions are indifferent.)

For this balances the extra complexity of p against the inaccuracy of $\hat{\theta}$.

Next, drop the assumption that all degenerate $\hat{\theta}$ are equally complex. Suppose, for example, that $\hat{\theta} = 0$ is simpler than any $\hat{\theta} \neq 0$, these others all being equally complex. (This is plausible when θ is a regression coefficient; $\hat{\theta} = 0$ then corresponds to dropping a variable.) Inaccuracy cost is still calculated as above. We then have the table

		Decision		
		I full distribution	II point estimate $\hat{\theta}_0 = E\theta$	III $\hat{\theta}_0 = 0$
Costs	Complexity	C_2	C_1	0
	Inaccuracy	0	Var θ	Var θ + $(E\theta)^2$

Here $C_2 > C_1 > 0$. The inaccuracy cost for III arises from $E(\theta - 0)^2 = \mathrm{Var}\ \theta + (E\theta)^2$.

The decision for I, II, or III depends on the least total cost, and is expressible in terms of the first two moment of the p distribution, as in Figure 1.

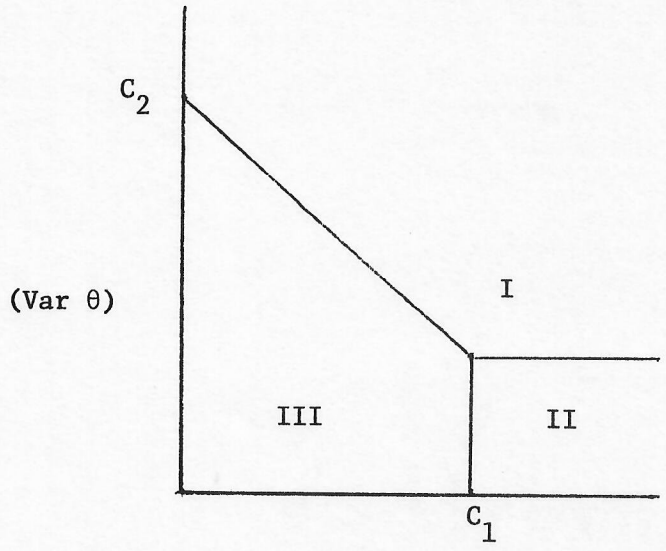

Fig. 1

We now make a radical suggestion: Decision III should read--"accept the hypothesis that $\theta = 0$". The testing notion here is quite different from that of any school in orthodox statistics (Fisher or Neyman-Pearson) and also from the various bayesian approaches to hypothesis testing (Jeffreys or Lindley). These approaches all say in effect: "Accept a hypothesis when it is likely to be true." The post-bayesian approach says: "Accept a hypothesis when it is simple and likely to be approximately true."

We shall first elucidate the differences in outlook, and then argue that the post-bayesian concept is the right one for science and for the art of living.

Suppose θ is known to be equal to a number $\theta_0 \neq 0$, where $\theta_0^2 < C_1$. Then $(E\theta)^2 < C_1$, and $\mathrm{Var}\theta = 0$, so the hypothesis $\theta = 0$ is <u>accepted</u>, even though known to be false. The point is that 0 is <u>close</u> to the true value, so that the extra simplicity of $\theta = 0$ more than compensates the slight loss in accuracy. We make the same implicit judgment every time we round off an estimate.

It is an article of folk wisdom among statisticians that any null hypothesis will eventually be rejected with a big enough sample.

(J. Berkson, JASA, 1938). Indeed, we know a priori that almost all our models involve idealizations that are, strictly speaking, false. There are no perfectly fair coins, etc. Any hypothesis specifying that a parameter θ lies in a lower dimensional surface of the full parameter space Θ may be assumed a priori to be false. Now conventional significance tests do test the literal truth of hypotheses and are therefore bound to falsify them eventually, leading to Berkson's paradox. The disenchantment with significance tests in the literature--the contrast drawn between "statistical significance" and "economic significance" (Dillon and Officer; cf. Morrison, The Significance Test Controversy)--attests to the growing awareness of this very unsatisfactory situation. In short, standard tests appear to be asking the wrong question.

The post-bayesian approach avoids Berkson's paradox by asking a different question.*

Turning to scientific practice, we find everywhere theories being applied which are known to be false--e.g. classical mechanics. And even of theories not known to be falsified, any of these can be embedded in a continuum of "neighboring" theories (perturbations) all of which are compatible

---

*A quick comparison with two alternative approaches. Hodges and Lehmann avoid Berkson's paradox by substituting H' for the original hypothesis H, where H' is all θ-values within some distance δ of points in H. This raises the question of finding the appropriate metric, finding the proper level δ , and relating these choices to appropriate loss functions. Should δ depend on sample size, for example? Thus, while the Hodges-Lehmann approach is an improvement over more conventional ones, it has an ad hoc character which needs framing in terms of a more general approach--the post-bayesian.

Jeffreys (Theory of Probability) believes that simple hypotheses have positive prior probability, and therefore denies that Berkson's paradox must occur for properly designed (bayesian) tests. But even if we grant Jeffreys his prior for the sake of argument, the post-bayesian argument above still applies: the posterior probability of an hypothesis may be very close to zero, and yet the hypothesis still be perfectly acceptable if it is simple and "close" to being true.

with all known facts (E.g. add a term with coefficient ε, where ε is sufficiently small). The specific theory accepted by the scientific community is typically the simplest one in the neighborhood. To summarize, the post-bayesian approach is (implicitly) the one accepted in scientific research.

Unknown Normal Mean, Variance Known

We give a post-bayesian analysis of the following standard problem. $X_1, \ldots X_n$ are iid $N(\theta,\sigma^2)$, σ^2 known. Make an inference concerning θ on the basis of $X_1, \ldots X_n$.

For simplicity (!), let θ have a uniform prior; then $(\theta | X_1, \ldots, X_n) \sim N(\bar{X}, \sigma^2/n)$. A strictly bayesian analysis ends at this point. But consider the alternatives (a) make a point estimate (b) test the hypothesis $\theta = 0$. We use the cost functions suggested above, so that inaccuracy loss is quadratic, C_1 is the cost of a non-zero point estimate over $\hat{\theta} = 0$, and $C_2 > C_1$ is the cost of the full distribution. Then, if a point estimate is to be made, it should be $\hat{\theta} = \bar{X}$. From Fig. 1, we see that a point estimate should be made if $\bar{X}^2 > C_1$ and $\sigma^2/n < C_2 - C_1$. Hypothesis $\theta = 0$ should be accepted if $\bar{X}^2 < C_1$ and $\bar{X}^2 + \sigma^2/n < C_2$.