THE ROBUST BAYESIAN VIEWPOINT[1]

by

James O. Berger
Purdue University

Technical Report #82-9

Department of Statistics
Purdue University

April, 1982

[1] Research supported by the National Science Foundation under Grant #MCS-8101670A1.

CONTENTS

1. INTRODUCTION

1.1 Introduction

Statistics needs a "foundation", by which I mean a framework of analysis within which any statistical investigation can theoretically be planned, performed, and meaningfully evaluated. The words "any" and "theoretically" are key, in that the framework should apply to any situation but may only theoretically be implementable. Practical difficulties or time limitations may prevent complete (or even partial) utilization of such a framework, but the direction in which "truth" could be found would at least be known.

To a large number of statisticians the above goal is deemed unattainable, with the attendant attitude being that one must "keep an open mind" and use "whatever works well for a given problem". Besides seeming unnecessarily pessimistic and somewhat unscientific, such a position seems almost meaningless in that without the desired foundational framework there would be no way of determining what works well in a given problem.

The main contender for the crown is Bayesian analysis. ("Classical" statisticians tend to be of the "there is no foundation" ilk.) The main justification for Bayesian analysis is a belief (for a variety of reasons) in

Assumption I. In any statistical investigation, one will ultimately be faced with making reports, inferences, or decisions which involve uncertainties. Of interest is the information available about these uncertainties after seeing the data, and the only trustworthy and sensible measures of this information are Bayesian posterior measures.

Belief in Assumption I leads many Bayesians to argue that the desired foundation is simply the usual Bayesian analysis in which one specifies a prior distribution for the unknowns and processes the data via Bayes rule.

This attitude is vigorously opposed by non-Bayesians, partly because of objections to Assumption I but, more often, because of a belief in

Assumption II. Prior distributions can never be quantified or elicited exactly (i.e., without error), especially in a finite amount of time.

Because of this distrust of prior distributions, many statisticians reject the Bayesian viewpoint out of hand.

A Bayesian viewpoint has long existed, however, which is based on belief in both Assumptions I and II. While Assumption I calls for a basically Bayesian outlook, Assumption II precludes the obvious Bayesian solution of writing down a single prior distribution and doing a Bayesian analysis. Instead, the viewpoint is essentially that one should strive for Bayesian behavior which is satisfactory for all prior distributions which remain plausible after the prior elicitation process has been terminated. I will call this the robust Bayesian viewpoint, and argue that it provides the desired foundational framework.

The robust Bayesian viewpoint is by no means new, of course, and virtually all Bayesians will ascribe to it to some degree. For example, deFinetti (as quoted by Dempster (1975)) stated

> "Subjectivists should feel obligated to recognize that any opinion (so much more the initial one) is only vaguely acceptable...So it is important not only to know the exact answer for an exactly specified initial problem, but what happens changing in a reasonable neighborhood the assumed initial opinion."

Most of the arguments and examples presented herein have undoubtedly been presented elsewhere. For instance, a very large proportion of the ideas can be found in the works of I.J. Good, even as early as Good (1950). (Indeed, I would have very few qualms about calling myself a Doogian.) Herman Rubin and Bruce Hill (among others) have also always espoused similar views. In some sense, therefore, this should be thought of as basically a review paper, with the goal of tying together the various elements of the robust Bayesian viewpoint in an attempt to present a convincing case. (To keep the account readable, I will defer most historical references to Section 5.)

This article is written more for the "non-robust" Bayesian, than for the non-Bayesian. In other words, little attempt will be made to justify Assumption I. Besides the sheer impossibility of adequately discussing Assumption I in a single paper, the rationale is that the Bayesian should have clean hands before he accuses someone else's hands of being dirty. Presenting the (enormously convincing) arguments for Assumption I seems to have little effect on non-Bayesians if they are able to come back with the complaint that Assumption II seems totally obvious to them and they refuse to operate in violation of it. Fully admitting (and even expounding on) the truth of Assumption II, while showing how Assumption I can still basically be followed, should greatly enhance the Bayesian argument. (See Berger (1982d).)

In reading the paper, keep in mind that the robust Bayesian viewpoint is being advocated as the framework for ultimately verifying the sensibility of an analysis, and is not necessarily being advocated as an applied methodology to do all of statistics. Comparatively little work has been done on robust Bayesian methods, so while it is a very illuminating viewpoint from which to understand things, it is not to be expected to provide easy answers to all our problems.

As a final caveat, although I will talk about various "classes" of Bayesians and non-Bayesians (such as "objective" Bayesians, "pure" subjective Bayesians, frequentists, etc.), these classes are to a large extent imaginary; most statisticians are a composite of a number of such classes. These distinctions will be made only for convenience in representing certain basic viewpoints.

In Section 2, justifications for Assumptions I and II will briefly be outlined. Since II implies that one must consider classes of plausible prior distributions, reasonable such classes will also be discussed, along with the problems of updating prior information. Section 3 is concerned with methods of measuring Bayesian robustness, and the somewhat surprising conclusion is reached that frequentist measures can be useful in measuring robustness. (This seems to conflict with Assumption I, and indeed behavior violating Assumption I can occur under this viewpoint, but <u>only</u> to the extent necessary to achieve robustness.) Section 4 deals with certain consequences of adopting this viewpoint, showing how certain features of many non-Bayesian techniques can be partially justified from the robust Bayesian viewpoint. Section 4 also presents an example, involving the Stein effect, which demonstrates that naive Bayesian intuition is not always trustworthy in the face of robustness considerations. Section 5 gives a brief survey of existing work related to Bayesian robustness, and contains some useful guidelines for achieving robustness. Section 6 consists of some conclusions and philosophical meanderings concerning the robust Bayesian viewpoint and objections to it.

1.2 Notation

In this paper it will be assumed that the data x is a realization of a random variable X with distribution $P_\theta(\cdot)$ on the sample space $\mathscr{X}$ for some unknown $\theta \in \Theta$. Although Θ will be referred to as the parameter space and $\{P_\theta, \theta \in \Theta\}$ will usually be a parametric family in the examples, the basic arguments hold for any index set Θ; thus the nonparametric situation would be included by letting Θ index any desired class of probability distributions. A prior distribution on Θ will be denoted π, $\pi(\cdot|x)$ will denote the posterior distribution of θ given the observation x, and $m(\cdot)=E^{\pi}[P_\theta(\cdot)]$ will denote the marginal (or unconditional or predictive) distribution of X. (E will stand for expectation, with superscripts indicating what the expectation is being taken over, and subscripts indicating fixed parameter values.)

1.3 Decision Theory

Many of the examples discussed will be presented from a decision theoretic viewpoint. The reason is mainly that, if a point is to be made, it can most clearly be done in a precisely quantifiable situation. It can, of course, be argued that, just as the robust Bayesian viewpoint seems necessary for understanding, so the robust decision theoretic viewpoint is also essential. ("Inference" problems would simply be problems where very little knowledge concerning the loss function was obtainable, and hence where robustness over a wide class of loss functions would be sought.) I certainly support this view, feeling that there are great dangers in refusing to at least think in decision theoretic terms. (Incidentially, it has always struck me as curious that there are violently antidecision-theoretic Bayesians and violently anti-Bayesian decision theorists. Is there really such a big difference between the two types of subjective inputs?) To keep the paper contained, however, the decision theoretic issue will not be explicitly considered; issues get clouded when too much is attempted.

When employing a decision theoretic viewpoint, the action space will be denoted $\mathcal{A}$, the loss in taking action $a \in \mathcal{A}$ when $\theta \in \Theta$ obtains will be denoted $L(\theta,a)$, and the posterior expected loss of action a with respect to the prior π and observation x will be denoted

$$(1.1) \qquad \rho(\pi,x,a) = E^{\pi(\cdot|x)}L(\theta,a) = \int_{\Theta} L(\theta,a)\pi(d\theta|x).$$

A decision rule (for simplicity assumed to be a nonrandomized function from $\mathcal{X}$ into $\mathcal{A}$) will be denoted $\delta(x)$. We will have cause to consider the (frequentist) risk function

$$R(\theta,\delta) = E_{\theta}L(\theta,\delta(X)) = \int_{\mathcal{X}} L(\theta,\delta(x))P_{\theta}(dx)$$

and the Bayes risk

$$(1.2) \qquad r(\pi,\delta) = E^{\pi}R(\theta,\delta) = \int_{\Theta} R(\theta,\delta)\pi(d\theta)$$

$$= E^{m}\rho(\pi,X,\delta(X)) = \int_{\mathcal{X}} \rho(\pi,x,\delta(x))m(dx).$$

2. THE ROBUST BAYESIAN VIEWPOINT

As the robust Bayesian viewpoint is founded on a belief in Assumptions I and II, these assumptions will be discussed in the first two subsections. Subsection 2.3 discusses reasonable classes of prior distributions which could be considered in light of Assumption II. Subsection 2.4 discusses the issue of updating uncertain prior information.

2.1 Justification for Assumption I

There are at least seven basic reasons that have been advanced for being a Bayesian, these being:

(i) Prior information is too important to ignore or deal with in an adhoc fashion.

(ii) According to most "classical" criteria, the class of "optimal" procedures corresponds to the class of Bayes procedures, so one should select from among this class according to prior information.

(iii) The Bayesian viewpoint works better than any other in revealing the common sense features of a situation and producing reasonable procedures.

(iv) The goal of statistics is to communicate evidence about uncertainties, and the correct language of uncertainty is probability. Only subjective probability provides a broad enough framework to encompass the types of uncertainties encountered, and Bayes theorem tells how to process information in the language of subjective probability.

(v) Axioms of rational behavior imply that any "coherent" mode of behavior corresponds to Bayesian behavior with respect to some prior distribution.

(vi) The Likelihood Principle seems irrefutable, yet the only general way of implementing it seems to be through Bayesian analysis.

(vii) Bayesian posterior measures of accuracy seem to be the only meaningful measures of accuracy.

Many papers and books have been written about these reasons, and no attempt will be made to review or explain these reasons in detail. A few comments seem in order concerning the importance and effectiveness of each of these reasons, however.

Reasons (i), (ii), and (iii) do not bear directly on Assumption I, but do lend considerable support to the Bayesian position. Reason (i) is important, especially when it is realized that choice of such things as a model is really just a (perhaps rather extreme) use of prior information. Nevertheless, reason (i) is not very effective for "conversion" since it can always be argued (incorrectly or not) that in many problems no (or very little) prior information is available.

Reason (ii) is very suggestive, pointing out a frequently occurring one-to-one correspondence between "good" classical procedures or methods and Bayesian procedures. In testing between two simple hypotheses, for example (the "dichotomy" discussed from this perspective by Lindley and Savage in, for instance, Savage, et. al. (1962)), the classical Neyman-Pearson tests are the Bayes tests. In selecting a test, therefore, one can either make a grand intellectual leap to α and β, or can carefully consider the available prior information (and information about the loss or consequences of accepting and rejecting and cost of experimentation) and select the test on Bayesian (decision-theoretic) grounds. To me the essence of reasoning is to reduce a complicated problem to simple components, analyze the components separately, and recombine to get an answer. I distrust grand intellectual leaps.

Another example, involving current research, is the work on finding alternatives to the least squares estimator, due either to pursuance of the Stein phenomenon (that in three or more dimensions the usual estimator is inadmissible) or ridge regression ideas. Again there is a one-to-one correspondence between "good" procedures (say, as measured by mean squared error) and Bayesian procedures (as shown in the normal case by Brown (1971)). One can thus select an alternative to the usual estimator either by a (mystical to me) intuitive method, or by considering which θ are apriori most likely to occur and selecting a Bayes estimator designed to do well for these θ (while preserving mean squared error dominance if desired). Further discussion of this example is given in subsection 4.5.

Reason (iii) is certainly not very good for conversion, but is the reason Bayesians tend to become more and more Bayesian as time progresses. Application of Bayesian reasoning will time and again clear up mystifying situations, and easily arrived at Bayesian procedures (say, with respect to noninformative prior distributions) often perform much better than complicated and difficult to determine classical procedures. (It is a shame that the very simplicity of much of Bayesian analysis is considered an indictment of it; I may find very stimulating a difficult mathematical derivation of, say, a minimax rule, and not be so intellectually excited at the routine calculation of the corresponding noninformative prior Bayes rule, yet (if done sensibly) the latter rule will virtually always be better.)

The last four reasons all pertain to the validity of Assumption I, and indeed are very related. They correspond to essentially four different modes of argument for Assumption I, however, and hence are listed separately.

Reason (iv) has been eloquently argued by many scientists, philosophers, probabilists, and statisticians (c.f. Jeffreys (1961), deFinetti (1972, 1974, 1975), Good (1950), Jaynes (1981), and Lindley (1982)). We often work very hard in elementary statistics courses to suppress in students their natural instincts to talk about the "chance that θ is in the interval" or the "probability that the hypothesis is true", telling them that (although these are what they really want to know) we must be "objective" and create an artificial language of confidence statements and error probabilities. Such artificial languages do not seem able to withstand deep scrutiny.

Reason (v) is compelling to many, but is perhaps a touch overemphasized. The axioms of rationality are, for the most part, very believable, and it is interesting to know that any coherent method of behavior corresponds to Bayesian behavior with respect to some prior distribution, but this does not say that the right way to behave is to write down a prior distribution and perform a Bayesian analysis. Indeed I would term this latter behavior incoherent (in a broad sense), in that the prior distribution used can only be an approximation to true prior beliefs (see the next subsection). The value of rationality and coherence is that they indicate that my "optimal" analysis will correspond to a Bayesian analysis with respect to my "true" prior distribution (admittedly circularly defined here), and hence indicate the direction in which I should look to determine my optimal analysis. See Section 3.3 and Berger (1982d) for further discussion and references.

Reasons (vi) and (vii) are often the most convincing to non-Bayesians. They bring out the key point that many Bayesians became Bayesians, not because they were infatuated with prior information, but because they could see no other meaningful solution to the conditional inference problems besetting classical statistics.

The Likelihood Principle is wonderful, in that so much follows from so little. The Likelihood Principle essentially says that if the family of distributions $\{P_\theta\}$ has densities $\{p_\theta\}$ with respect to some dominating measure and the observation from the experiment is x, then the evidence about θ obtainable from the experiment is contained in the likelihood function $\ell_x(\theta) = p_\theta(x)$ (considered as a function of θ). The appeal of the principle is partly the fact that (as shown by Birnbaum (1962)) it follows from the Principles of Sufficiency and Conditionality; indeed all that is needed of the Conditionality Principle is that if one chooses between two experiments based on the flip of an (independent) fair coin, then the evidence about θ obtained is precisely the evidence obtained from the experiment actually performed. These latter principles seem so self-evident that it is hard to disagree with the Likelihood Principle, yet belief in the Likelihood Principle forces a complete revolution in thought; one must then think conditionally on the actual observation x. Further investigation (c.f. Basu (1975) and Berger and Wolpert (1982b)) leads to the conclusion that $\ell_x(\theta)$ can be meaningfully used only by considering it as a probability density with respect to a measure π, which should reflect prior beliefs about θ. Hence the result of this line of reasoning is that one must view things in a Bayesian fashion.

Reason (vii) is related to the Likelihood Principle, in that it argues that only conditional measures (based on the posterior distribution) given x are sensible for evaluating the evidence about θ, but it is less foundational and more of a "proof by counterexample". For instance, consider

Example 1. Suppose $X = \theta + 1$ or $\theta - 1$ with probability $\frac{1}{2}$ each $(\theta \in R^1)$, and that a 75% confidence interval of smallest size for θ, based on independent observations X_1 and X_2, is desired. This is obviously given by

$$C(x_1,x_2) = \begin{cases} \text{the point } \frac{1}{2}(x_1+x_2) & \text{if} \quad |x_1-x_2| = 2 \\ \text{the point } (x_1+1) & \text{if} \quad |x_1-x_2| = 0 \end{cases}$$

(or we could have chosen (x_1-1) if $|x_1-x_2| = 0$). But if $|x_1-x_2| = 2$, we are absolutely certain that $\theta = \frac{1}{2}(x_1+x_2)$, while if $|x_1-x_2| = 0$ we are equally uncertain whether $\theta = X_1 + 1$ or $\theta = X_1 - 1$ (barring specific prior information). In either case, it seems absurd to report $C(X_1,X_2)$ as being a 75% confidence interval. The point, of course, is that frequentist measures such as "confidence" can be totally misleading for given data x. The frequentist can protest that such measures as "confidence" are not to be interpreted conditionally, but what is the sense in proposing a measure of accuracy which clearly presents a false image of the information about θ contained in the data. (The Bayesian posterior credible regions for this situation are, of course, very sensible.)

Examples are available for essentially any non-Bayesian measure of accuracy (or at least any frequency measure of accuracy), showing that the measure can very inaccurately portray the information about θ contained in the observation x. After seeing enough of these examples, posterior measures start to look very attractive.

All sorts of classical defences and objections to reasons (vi) and (vii) can, of course, be raised, such as bringing in questions of design, stopping rules in sequential analysis, analysis in nonparametric situations (where a likelihood function may not exist),allowing "conditional" frequentist statements, etc., but they all seem to be answerable. Further discussion here would be inappropriate and can be found in Basu (1975) and Berger and Wolpert (1982b), which also contain earlier references.

2.2 Justification for Assumption II

Assumption II seems almost transparently obvious, yet there is considerable resistance to it among many Bayesians. Hence a brief discussion seems in order.

In the first place, there are situations in which it seems simply unreasonable to expect that beliefs can even be modeled by a single prior distribution. Consider the following simple (though admittedly artificial) example, essentially given in Zaman (1982).

Example 2. Consider 3 boxes labelled A, B, and 2B, one of which contains a ball. The only information you have is that box 2B is twice as likely to contain the ball as box B. You are to determine your subjective probabilities p_A, p_B, and p_{2B} of the ball being in the indicated box. Clearly you should have $p_{2B} = 2p_B$, but it is not clear what else can be said. Since nothing is known comparatively about A and B, it seems that one should have $p_A = p_B$, but by the same reasoning one would say $p_A = p_{2B}$, and both cannot hold. It does seem reasonable to suppose that one's prior probabilities should satisfy the constraints $p_B \leq p_A \leq p_{2B}$ and $p_{2B} = 2p_B$, but it is unreasonable to expect anything more precise to be concluded.

Even if in a situation where it is reasonable to expect beliefs to be expressible in terms of a single prior distribution π_T, can this actually be done? Consider, for instance, any of the axiomatic systems which guarantee the existence of π_T. (In the situation of Example 2, at least one of the axioms in any system will be violated.) The prior π_T is obtained by various betting or comparison schemes, but is exactly nailed down only after an infinite process of elicitation. This is clearly the case when Θ is infinite (or when the associated σ-field of events is infinite) since there are then simply an infinite set of probabilities to determine. Even when Θ is finite, the axiomatic systems formally call for considering an infinite number of bets or comparisons. In the betting schemes, one must compare all possible wagers, and indeed should really base the bets on a utility function which itself takes an infinite amount of time to perfect; and in the comparison schemes one must compare events with the infinite set of measurable events from some auxilliary - say, uniform - distribution. And all this assumes that Θ is known, whereas in many situations the possible states of nature are only vaguely comprehended (c.f. Shafer (1979, 1981a, 1981b) and Barnard (1982)).

From a strictly intuitive viewpoint it is also clear that the single prior axiom systems are, in a sense, inapplicable, since there is obviously a lower limit to the accuracy of prior elicitation. I cannot believe that anyone could ever distinguish between $P(A)=.25$ and $P(A) = .250001$ (or $P(A) = .25 + 10^{-100}$ if an extreme case is needed) in terms of subjective elicitation. Thus Savage (1961) says

> "No matter how neat modern operational definitions of personal probability may look, it is usually possible to determine the personal probabilities of important events only very crudely."

Similar views can be found in Koopman (1940), Good (1950, 1962a, 1973 (priggish principle 3)), Savage (1954), Smith (1961), Dempster (1967, 1968), Fine (1973), Kyburg (1974, 1976), Suppes (1975), Levi (1980), Rios and Girón (1980), DeRobertis and Hartigan (1981), and Zaman (1982).

Some Bayesians argue that the concept of a "true" prior π_T is meaningless, in that the approximate prior π_A that one arrives at after a finite amount of time is your true prior at the moment, and should hence be used as such. In the face of infinite Θ this is clearly not very reasonable, since in a finite amount of time an infinite set of probabilities cannot be specified without introducing a large degree of arbitrariness. Even if only a finite Θ is involved, it seems unreasonable to look upon π_A as any form of truth, since further thought would likely cause further refinement and there is always a considerable fuzziness in subjective elicitation. The distinction between π_T and π_A is made very succinctly by Dickey (1976a, 1976b), who calls them the "actual prior distribution" and the "operational prior distribution", respectively, and point out situations in which π_A can be known to be a good approximation to π_T. (It is possible to argue philosophically that π_T is essentially an imaginary quantity itself -- c.f. parts of Levi (1980) -- but it is often a useful imaginary quantity to consider.)

As an aside, it is interesting to observe that, in the above light, the subjectivist Bayesian objections to the objective Bayesian use of "noninformative" priors seem less forceful. In a situation where there is very little prior information about θ, a noninformative prior may be a better approximation to π_T than any hastily derived proper subjective approximation π_A.

Another situation, in which working with a class of priors is clearly unavoidable, is when group conclusions or decisions must be made and the priors of all members of the group must be considered. (See Weerhandi and Zidek (1981) and Zidek (1982) for discussion and earlier references.) The issue of scientific communication is related to this, the (often unattainable) ideal being that of presenting a conclusion which would be the conclusion for any reasonable prior that a user of the information might have. (Among the works bearing on this issue are Hildreth (1963), Dickey (1973), and Jackson, Novick, and DeKeyrel (1980).) Although ideas in these areas must bear a strong relationship to those discussed in this paper, we will not be formally considering such group situations.

The above arguments do not, of course, establish that a serious problem exists with standard (i.e. single prior) Bayesian analysis. Indeed I am very sympathetic to the claim that single prior Bayesian analysis is the ideal goal and that the major problem remaining is that of developing good prior elicitation techniques. There is a very substantial and growing literature on the subject of prior elicitation (c.f. Kadane, Dickey, Winkler, Smith and Peters (1980) for discussion and other references), and as better elicitation methods become available it is natural to expect the need for consideration of Bayesian robustness to decline. The validity of Assumption II from a philosophical viewpoint seems clear, however.

2.3 Reasonable Classes of Prior Distributions

In subsection 2.2 it was argued that quantification of prior beliefs can never be done without error, and hence that one is left, at the end of the elicitation process, with a set Γ of prior distributions which reflect true prior beliefs; i.e., π_T is an unknown element of Γ. Some comments are in order concerning the specification of Γ.

The first and most crucial realization is that, in quantification of prior beliefs, only prior probabilities and relative likelihoods can accurately be elicited. In other words, such features of the prior distribution as percentiles and shape features (unimodality, monotonicity, symmetry, smoothness, etc.) can be elicited with some confidence, while features such as moments and exact functional form are much harder to accurately determine. The reason for this is simply that assessment of probabilities of events (and hence of percentiles of the prior distribution) is certainly feasible, as likewise is intuitive comparison of the "likelihood" of the various θ (at least to the extent of leading to reasonably certain structural information about the prior density). To specify a prior moment, on the other hand, demands very accurate specification of the "tail" of the prior distribution, which will almost never be feasible. Consider the following example.

Example 3. Suppose θ is an unknown normal mean, and that a necessarily brief period of prior assessment results in the conclusions that

$$p^{\pi}(\theta \leq -1) = p^{\pi}(-1 < \theta \leq 0) = p^{\pi}(0 < \theta \leq 1) = p^{\pi}(1 < \theta) \cong \frac{1}{4} ,$$

and that π has a symmetric unimodal density. Thus Γ could reasonably be chosen to consist of all priors with symmetric unimodal densities having median 0 and quartiles ± 1. (To be certain of robustness, it would probably be better to choose Γ to consist of all priors with medians within

ε_1 of zero and quartiles within ε_2 of ±1, with similar leeway for error allowed in the specification of symmetry. This will not make much difference in this situation, however.)

In this example, Γ contains both the conjugate prior $\mathcal{N}(0,2.19)$ (normal with mean zero and variance 2.19) and the $\mathcal{C}(0,1)$ prior (Cauchy with median zero and scale parameter 1). The normal prior has all moments, while the Cauchy prior has no moments whatsoever. It seems very unlikely in this situation to expect detailed knowledge of the tail of the distribution (i.e., detailed knowledge of a set of very small prior probability), so any attempt to specify Γ by prior moments seems fraught with peril.

An alternative reasonable approach to specifying Γ is to approximate π_T by a specific assessed approximation π_A, and then let Γ consist of all priors "close" to π_A. Again, "close" should be measured in terms of close probabilities, such as in the class

$$(2.1) \qquad \Gamma = \{\pi\colon \pi(\cdot)=(1-\varepsilon)\pi_A(\cdot)+\varepsilon P(\cdot),\ P \text{ an arbitrary probability distribution}\}\ ,$$

where ε reflects the believed accuracy of the prior assessment. This class Γ was first considered in Schneeweiss (1964), Blum and Rosenblatt (1967), and Huber (1973). Other reasonable classes can be found in Berger (1980b).

Much of the literature involving classes of priors chooses Γ to be either the set of priors with certain moments in specified ranges or a set of priors of a particular functional form with parameters in specified ranges. While these tend to be much easier to work with than are Γ such as in Example 3 or (2.1), they are unsuitable, as discussed earlier.

Easily specified classes, such as (2.1), are often somewhat too large, in that they do not incorporate probably available smoothness information about π. In (2.1), for example, it may well be felt that π definitely has a unimodal continuous density, which would place severe restrictions on the contaminations P allowed. Thus if, in doing a robustness analysis with respect to a Γ as in (2.1), robustness seems hard to achieve, make sure this is not due to unrealistic features of Γ. Of course, if robustness with respect to Γ is obtained, then one is also robust with respect to the more reasonable subclass.

It is, of course, possible to have Γ much less clearly specified than in the above examples, such as when θ is a high dimensional vector or, even worse, a nonparametric index. Only extremely crude or general features of the prior might then be obtainable, so Γ could be very large.

2.4 Updating Γ

Since we will primarily be concerned with posterior measures, the question of updating Γ by the data is obviously crucial. When making posterior conclusions, the obvious class of posteriors to consider is simply

$$\Gamma^* = \{\pi(\cdot|x): \pi \in \Gamma\}.$$

Unfortunately, more flexibility must be allowed if realism is to be achieved. The main difficulty is that, especially in multivariate problems, it would be far too time consuming (if even possible), to accurately ascertain even the most important features of the prior ahead of time. After seeing the data, however, one can determine which features of the prior will have a real impact and must carefully be considered. For example, in a complicated linear model the data may illuminate which variables are important and hence should be the focus of the prior elicitation. Or the data may indicate that some variables are accurately determined by the data, and hence prior information concerning them is likely to be less important, while other variables are very inaccurately determined from the data (due, say, to multicollinearity) and hence need accurate prior specification. Thus the data may cause further prior elicitation resulting in a reduced class Γ^* (which will then be updated by the data in the usual Bayesian fashion).

Major objections to the above approach can, of course, be raised, most troubling being the apparent dependence of the prior (not just the posterior) on the data. This offends many Bayesians, and also smacks of cheating and adhocery to non-Bayesians. To Bayesians, I can only reply that there is no choice. Typical situations have high dimensional Θ , for which it is very unrealistic to suppose that suitably accurate prior specification can be achieved; i.e., only very large Γ can be determined prior to experimentation. It will be very unlikely that robustness can be achieved with respect to such a large Γ. Hence a narrowing down of Γ will be needed, with the data indicating where further refinement is necessary. Note that the data is not to be used to shape your beliefs, but only to indicate how this narrowing down should be done, and when a point is reached which allows reasonably robust Bayesian conclusions to be drawn. As Hill (1965) says

> "...it is only the degree of care we take in approximating our prior, not the prior itself, that depends on the data."

A more troubling situation is when the data reveals that Γ was in some sense wrong, and not just too big. One could argue that Γ should have been kept flexible enough to encompass all possibilities, but realistically the data will often suggest new relationships, hypotheses, or models that were not included in, and may even contradict, the original specification of Γ. One must then go back and suitably enlarge or change Γ, as observed in deFinetti (1972, Chapter 8) and Savage (1962). As Savage, said, however,

> "It takes a lot of self-discipline not to exaggerate the probabilities you would have attached to hypotheses before they were suggested to you."

The disturbing nature of allowing the data to affect Γ directly does not seem quite so bad if a slightly different perspective is adopted. Instead of viewing the situation as that of updating prior information, think of it as an attempt to quantify (after the experiment) the relevant experimental and non-experimental information, and then combine the two. This, of course, is the view that outsiders, evaluating a robust statistical analysis, will take. A good analysis will present a suitable summarization of the data along with a description of the experimenter's Γ and his conclusions. In evaluating this, an outsider would consider the suitability of Γ, and alter Γ to reach his own conclusions if needed. How Γ was obtained is essentially irrelevant; either it seems a reasonable representation of the non-experimental evidence or it does not. The emphasis here is on the <u>effect</u> of the data <u>on</u> opinions, or on the prior to posterior <u>transformation</u>, a concept convincingly promoted by Dickey (1973). Another way of thinking of this is that one learns by passing a variety of reasonable priors over the likelihood function $\ell_x(\theta)$ and seeing what happens. The strict prior to posterior mode of reasoning is then deemphasized. (Indeed, Shafer (1981b) argues convincingly that practical

Bayesians almost never think in this strict mode, but instead view the problem as that of combining different sources of information.)

The clear difficulties of updating, by merely conditioning on the data via Bayes rule, have led to the development of other theories or methods for Bayesian or pseudo-Bayesian analysis. (See, for example, Jeffrey (1968), Shafer (1976, 1979, 1981a, 1981b, 1982), Teller (1976), and Diaconis and Zabell (1982).) These alternatives are interesting, but I remain unconvinced as to the practical necessity of developing a methodology which goes beyond post-data modification of Γ, followed by updating via Bayes rule. First of all, most of the examples against Bayesian updating can be handled by allowing post-data modification of Γ. Secondly, complex situations are understood by trying to break them into simple components for separate analysis; the prior - data, or alternatively, experimental - nonexperimental information decomposition is a very useful such breakdown, with a known method (Bayes rule) for recombination. Although contamination of information is certainly a real danger, and there may be situations where this breakdown is not necessary, in the overwhelming majority of the cases it is successful. A final argument for staying within the framework of Bayesian conditioning is that, as alluded to earlier, it is very important in statistical reports to separate the information contained in the data from that in the prior, and so this breakdown should be attempted even when not convenient.

While the above reasons argue against basing one's methodology on non-Bayesian updating, it would be foolish to rule out alternate methods completely. (See the discussion of this issue in Shafer (1979, 1981a, 1981b).) Also, certain ideas derived from these alternate viewpoints are useful in post-data modification of Γ. One such idea is the use of Jeffrey's rule, as discussed in Diaconis and Zabell (1982) and Shafer (1981a).

3. MEASURES OF ROBUSTNESS

The natural Bayesian measure of robustness is insensitivity of the final (posterior) conclusion to the choice of $\pi \in \Gamma$. This will be discussed in the next section. Though of central importance, this measure of robustness will be seen to be inadequate in some situations, necessitating measures of robustness of procedures based on overall performance, such as Bayes risk, $r(\pi,\delta)$, in decision-theoretic situations. This will be discussed in subsection 3.2. Subsection 3.3 discusses the role of each of these two methods of measuring robustness.

3.1 Posterior Robustness

Assumption I, being the cornerstone of the robust Bayesian viewpoint, must be followed. Hence, after observing all the data, any inference or decision made should be satisfactory from a posterior viewpoint.

Definition 1. An inference or decision is posterior robust with respect to Γ if it is satisfactory with respect to $\pi(\cdot|x)$ for all $\pi \in \Gamma$.

This definition is necessarily very vague, but could be tightened up in specific situations, such as in the following reasonable definition for decision-theoretic settings.

Definition 2. In a decision-theoretic setting (see subsection 1.3), an action a_0 is ε-posterior robust with respect to Γ for the observed x if

$$(3.1) \qquad \sup_{\pi \in \Gamma} |\rho(\pi,x,a_0) - \inf_{a \in \mathcal{a}} \rho(\pi,x,a)| \leq \varepsilon.$$

It is important to realize that whether or not posterior robustness exists will often depend on which x is observed. Thus Barnard (1982) says

> "We should recognise that 'robustness' of inference is a conditional property - some inferences from some samples are robust..."

Consider the following example.

Example 4. Assume that $X \sim \mathcal{N}(\theta,1)$ is observed, and that it is desired to estimate θ under loss $L(\theta,a) = (\theta-a)^2$. Here $\boldsymbol{\Theta} = \mathcal{A} = R^1$. Suppose $\Gamma = \{\pi_N,\pi_C\}$, where π_N is the $\mathcal{N}(0,2.19)$ distribution and π_C is the $\mathcal{C}(0,1)$ distribution. (This Γ is a very specialized subset of the Γ in Example 3, but behaves similarly in many respects.) If π_N were the true prior, then one would want to use the Bayes estimate

$$\delta^N(x) = \frac{2.19}{1+2.19}\, x \ ,$$

while if π_C were the true prior, then one would want to use the Bayes estimate

$$\delta^C(x) = \frac{\int\theta(1+\theta^2)^{-1}\exp\{-\frac{1}{2}(x-\theta)^2\}d\theta}{\int(1+\theta^2)^{-1}\exp\{-\frac{1}{2}(x-\theta)^2\}d\theta} \ .$$

Table 1 gives a few values of δ^N and δ^C.

Table 1. δ^N and δ^C

x	0	1	2	10
δ^N	0	.69	1.37	6.87
δ^C	0	.52	1.27	9.80

An easy calculation shows that, for squared error loss,

$$|\rho(\pi,x,a_0) - \inf_a \rho(\pi,x,a)| = (a_0-\mu_\pi(x))^2 ,$$

where $\mu_\pi(x)$ is the posterior mean for π. Since δ^N and δ^C are μ_{π_N} and μ_{π_C}, respectively, it follows that the posterior robustness of either $\delta^N(x)$ or $\delta^C(x)$ is measured by

$$[\delta^N(x)-\delta^C(x)]^2.$$

From Table 1, it is clear that either action is quite posterior robust (i.e., $\delta^N(x)$ is close to $\delta^C(x)$) for x near zero, while for $x = 10$, neither action is posterior robust. (For large x, the tail of the prior becomes very significant, and π_N and π_C have substantially different tails.)

If posterior robustness is attainable in a given situation, then the problem is essentially solved. If posterior robustness is not attainable, however, as happens in Example 4 when $x = 10$, then something else must be done. The natural thought is to attempt further elicitation of the prior distribution, and indeed it is precisely when posterior robustness does not obtain that more detailed elicitation is indicated. If this resolves the issue, fine, but if further elicitation is not possible or won't prove helpful (as in Example 4 for $x = 10$, where the prior tail will be next to impossible to accurately specify), then we must look beyond posterior robustness. (Of course, the above example is extreme, in that if encountered in practice one would seriously suspect the model for X. Extreme examples like this are useful for emphasizing the issues, however. They also provide insight which can be used in less extreme situations. Sections 5 and 6 deal with more practical issues.)

3.2 Procedure Robustness

Faced with the (X,θ) experiment, one can talk about the procedure $\delta(X)$ to be used when X is observed. Although the Bayesian tends to think conditionally on the observation $X = x$, it is certainly possible to consider the collection $\{\delta(x), x \in \mathcal{X}\}$ of inferences or decisions to be made for all possible X. (This may seem an unnecessary complication, but is logically sound.) Since, preexperimentally, the Bayesian thinks that X will be occurring according to the marginal distribution $m(\cdot)$, he would (in a decision theoretic setting, for simplicity) evaluate the overall performance of a procedure by

$$r(\pi,\delta) = E^m[\rho(\pi,X,\delta(X))].$$

A reasonable method of measuring the robustness of a procedure in such a situation is given in the following definition.

Definition 3. In a decision-theoretic setting, the procedure δ^0 is ε-procedure robust with respect to Γ if

$$\sup_{\pi\in\Gamma}\,[r(\pi,\delta^0) - \inf_{\delta} r(\pi,\delta)] < \varepsilon \ .$$

Example 4 (continued). Calculation shows that $r(\pi_C,\delta^N) = \infty$, $r(\pi_N,\delta^N) = .697$, $r(\pi_C,\delta^C) < 1$, and $r(\pi_N,\delta^C) = .736$. Hence the procedure robustness of δ^C (with respect to Γ) is measured by

$$r(\pi_N,\delta^C) - r(\pi_N,\delta^N) = .049 \ ,$$

while that of δ^N is measured by

$$r(\pi_C,\delta^N) - r(\pi_C,\delta^C) = \infty \ .$$

Clearly δ^C is much superior according to this measure of robustness. (Of course, the use of an unbounded loss function can be criticized, but even for many reasonable bounded losses δ^C would prove far superior.)

Many Bayesians object to the use of $r(\pi,\delta)$ as a measure of anything, because it involves an average over the sample space. A statistician should be responsible for the long run performance of his methodology, however. In the situation of Example 4, for instance, the Bayesian who time after time uses the conjugate prior Bayes rule δ^N will have very bad long run performance if π_C is the true prior fairly regularly, while the Bayesian who uses δ^C suffers no such danger when π_N is the true prior. In other words, if a Bayesian is to employ a methodology leading to the use of a procedure δ, he should be concerned that his methodology is sound, as reflected by $r(\pi,\delta)$. This is not to say that a procedure δ is good for all x if $r(\pi,\delta)$ is good (the fallacy in reasoning underlying frequentist statistics), but does say that δ is bad if $r(\pi,\delta)$ is bad. (Discussion of other reasons for considering $r(\pi,\delta)$ will be given in subsection 4.4.)

Many Bayesians react to the above argument by asking how $r(\pi,\delta)$ can be bad if $\delta(x)$ is chosen to be good from a posterior viewpoint for each x. Example 4 provides an illustration of how his can happen. From the viewpoint of posterior robustness, $\delta^N(x)$ and $\delta^C(x)$ were equivalent, in that the posterior robustness of each (with respect to Γ) was measured by

$$[\delta^N(x)-\delta^C(x)]^2.$$

But from the procedure robustness viewpoint, it seems clear that δ^C is considerably better than δ^N.

From the procedure robustness viewpoint, several specific criteria have been proposed for the selection of procedures. The two most common are the Γ-minimax and Γ-minimax regret criteria, which propose the use of the procedure δ^* which minimizes

$$\sup_{\pi \in \Gamma} r(\pi,\delta^*) \tag{3.2}$$

or

$$\sup_{\pi \in \Gamma} [r(\pi,\delta^*) - \inf_{\delta} r(\pi,\delta)], \tag{3.3}$$

respectively. Discussion of the literature on these criteria will be delayed until Section 5.

3.3 Discussion

It is important to realize that posterior robustness is the ideal goal. If it can be attained, the problem is solved. Also, when posterior robustness is not present, a careful Bayesian will attempt further refinement of Γ or, if possible, attempt to obtain more data. Unfortunately, situations where posterior robustness is simply unattainable are common, such as when (i) because of time or mental limitations further refinement of Γ is impossible; (ii) no more data can be obtained; or (iii) Bayesian analysis is technically too difficult to implement for a convincing variety of plausible priors (as in many nonparametric problems).

What alternatives are available when posterior robustness cannot be found? First, one could simply say that there is no clearcut answer to the problem. This is reasonable, at least in those situations where Γ is clearly defined and different priors in Γ give substantially different answers. If, however, the problem is due to technical difficulties in implementing the Bayesian approach, or if an answer simply must be obtained, then something else must be tried.

The natural Bayesian inclination would be to put some "metaprior" on Γ itself, and use the resulting Bayes rule. If technically feasible, this may well be a good adhoc solution. We stress "adhoc" because the assumption is that no further prior elicitation is possible. Thus the metaprior is simply some arbitrarily chosen distribution used as a technical device to obtain an answer. The analysis with metapriors can be very formidable, however, especially with Γ such as discussed in Section 2.3. Also, there is nothing to guarantee that the resulting answer will be good. Hence it may well be useful to consider procedure robustness and/or use of frequency measures as an aid in obtaining an answer. A more extensive discussion of the use of procedure robustness and frequency measures will be given in Sections 4.4 and 4.5.

The complaint can be raised that use of procedure robustness may violate Assumption I and the Likelihood Principle, and also that use of such measures as (3.2) and (3.3) and frequency measures will violate the rationality or coherency axioms. This is a valid complaint, yet carries no real force since a point has been reached where there is no clearcut "coherent" way to proceed. Here, coherent is being used in a broad sense, since it would formally be coherent (in the usual sense) to arbitrarily select some metaprior on Γ and do a Bayesian analysis, yet few Bayesians would say that arbitrary choice of a prior (i.e. a choice not based on any subjective opinions) is necessarily good. Thus Levi (1980) says

> "We should, therefore, recognize a distinction between principles of rationality regulating an agent's commitments

> and the suggestions which may be made when he cannot live up to them."

It should be stressed that we are not recommending any definite way of proceeding when posterior robustness is lacking. Often, putting an artificial prior on Γ may work. Often (see Sections 4.4 and 4.5) use of procedure robustness or frequency measures may prove helpful. Or entirely different statistical methodologies may provide good answers. In fact, the coherency arguments essentially suggest that no single automatic prescription concerning what to do in this situation will always prove successful.

It is crucial, finally, to recall that we are contemplating straying from the Bayesian path only to select from among answers which are plausible from a posterior Bayesian viewpoint, and hence will not be knowingly violating Assumption I or coherency by any substantial amount. Thus Good (1976) says

> "...non-Bayesian methods are acceptable provided that they are not seen to contradict your honest judgements, when combined with the axioms of rationality."

4. IMPLICATIONS OF THE ROBUST BAYESIAN VIEWPOINT

The major implications of the robust Bayesian viewpoint have already been discussed, but the flexibility of the approach allows incorporation of various sensible,yet ostensibly "non-Bayesian",techniques. Some of these are briefly discussed below.

4.1 Data Analysis

The data summarization part of data analysis is justifiable from any viewpoint, so it is the interactive modeling aspect which is of interest. This activity always involves the combining of subjective knowledge with the data to suggest or modify models for the phenomenon being studied, and is hence essentially Bayesian in nature. As discussed in subsection 2.3, it seems sensible and necessary to allow modification of Γ based on the data, and indeed, with this option, the robust Bayesian and data analyst behave in essentially the same way. The differences are, first, that the robust Bayesian believes in quantifying the subjective information (to the extent possible) in Γ, rather than incorporating it in an adhoc fashion; and, second, the robust Bayesian uses posterior measures in evaluating the evidence for any model or conclusion. This last feature eliminates, in a sensible fashion, the problems of evaluation of the strength of the evidence for a model selected by the data. (The posterior weight given to the model is based on a product of the prior weight and likelihood according to the data, automatically discounting the "significance" of the data for the model it selects.) Hence, contrary to popular opinion, the robust Bayesian is not the slave of a particular prior distribution he must pre-experimentally specify, and can engage in sensible data analysis (as opposed to non-Bayesian data analysis).

4.2 Randomization

Most statisticians are convinced of the value of randomization in statistical design (e.g. random allocation of subjects to two treatments), yet the single prior Bayesian position does not allow this. If all unknowns in the situation have been identified and their true prior distribution obtained, then the optimal Bayesian design will not require any form of randomization. When, however, uncertainty in the prior information is admitted, randomization becomes available.

The use of randomization to a robust Bayesian, however, is essentially limited to the effort of avoiding experimenter induced bias. In other words, because the robust Bayesian is worried that there are experimental factors which he has not thought of and which may be correlated with any nonrandom subject selection or allocation scheme, he will find randomization to be useful in (hopefully) preventing such bias.

The robust Bayesian does not (as an ideal) find randomization to be of use in drawing conclusions from the data. The probabilistic mechanism of randomization will usually be independent of θ, and hence by Assumption I the robust Bayesian will want to draw conclusions conditional on the given selected sample. Of course, even the non-Bayesian agrees with this to some extent, the "selection" of a new randomization design if the original design doesn't look random enough being one example. And even the most ardent anti-Bayesian would not go through with a standard classical analysis based on the randomization if significant cofactors were revealed which, by bad luck, turned out to be highly correlated with, say, the treatment groups. Yet the Bayesian conditional viewpoint argues against making any use of the randomization mechanism. Arguments for this viewpoint can be found in Basu (1971) and Basu (1980). (See also the discussion by Lindley in Basu (1980).)

It is possible to argue that robustness considerations allow the use of the randomization mechanism. For instance, Rubin (1978) argues that the prior specification is so immensely complicated in typical situations that it will often be better to "ignore" part of the data (i.e. the known outcome of the randomization) to simplify the needed prior specification. The probability mechanism of the randomization does then become part of the Bayesian analysis, and can indeed simplify matters.

The danger in this is, of course, the usual danger befalling any attempt to analyze data in violation of Assumption I; the analysis conditional on the data could differ substantially from the analysis averaging over data points that could have been obtained. Although this is something that will probably occur fairly rarely, it is unappealing to adopt as a basic method of analysis techniques which can lead to conclusions at odds with all the actual data. Note that the robustness advocated in this paper is not of this potentially dangerous type, since satisfactory conditional posterior behavior is of primary importance.

There may, of course, be very pragmatic considerations involved. For example, a randomized design will be useful if it seems important to convince others that the experiment was "unbiased" (although this is rather illusory impartiality). Also one can be very sympathetic to the argument that any Bayesian analysis here, much less a robust Bayesian analysis, is simply unmanageable.

Discussion of the randomization issue can also be found in Savage, et. al. (1962), Hill (1970), Good (1976 and earlier), Basu (1980), Lindley and Novick (1981), and Berger and Wolpert (1982b). Also, the debate in sampling theory concerning the use of superpopulation models as opposed to analysis based on the probabilistic mechanism of the sampling rule is essentially the same as the randomization debate.

Indeed Godambe and Thompson (1977), Godambe (1982), and Royall and Pfefferman (1982) specifically argue that suitable random sampling plans can lead to a form of Bayesian robustness. Other discussion and references can be found in Cassel et. al. (1977), Basu (1978), Hájek (1981), and Berger and Wolpert (1982b).

4.3 Classical Robustness

By classical robustness is meant robustness with respect to the distribution $P_\theta(\cdot)$ of the observation X. This is obviously a crucial aspect of statistical analysis, and can be included in the robust Bayesian framework by the simple expedient of allowing Θ to be a nonparametric index set (indexing the distributions for X which are of concern), and having Γ reflect the prior knowledge available about these distributions. Indeed, to many Bayesians the difference between "model" and "parameter" seems fuzzy at best. The subjective choice of the model is often a far more drastic use of prior information than is use of prior distributions on parameters of the model.

Classical robustness results tend to be in terms of measures such as "asymptotic minimaxity" (c.f. Huber (1972)), which can be related to procedure robustness. Procedure robustness is of interest here, because Bayesian analysis when P_θ is uncertain can be technically very difficult. A number of successful Bayesian analyses of model robustness problems have been carried out, however. For the most part, these studies proceed by embedding a standard family of distributions in a larger parametric family (such as embedding the normal distributions in the class of all t-distributions), and then performing a Bayesian analysis. Excellent discussions of this, along with earlier references, can be found in Box and Tiao (1973), Dempster (1975), and Box (1980).

One important point brought out in the Bayesian view is that model robustness should be viewed conditionally. If a data set gives residuals which are a gorgeous fit to normality, worrying about robustness to normality is a waste of time. Discussion and examples can be found in Dempster (1975) and Barnard (1982). Efron and Hinkley (1978) and Hinkley (1982) also discuss important situations in which model robustness should be investigated conditional on shape features of the data. All this is in line with our view that having valid conditional (posterior) measures is of primary importance.

4.4 Uses of Frequency Measures

Frequency measures can have a role to play in robust Bayesian analysis. The basic idea of frequency measures is, of course, to also consider x other than that which occurs. The simplest form of such reasoning, which can be useful to a Bayesian, is simply to imagine possible data x, compute the Bayes rule for a prior being investigated, and see if the result makes sense. In the situation of Example 4, for instance, the fact that δ^N appears inadequate for x=10 provides a warning that δ^N might also be inferior for a smaller (yet possible) observation such as x=5. Several very interesting examples of this type of reasoning are given in Diaconis and Freedman (1981). Looking at the behavior of a Bayes rule for a variety of x (often extreme x) may point out unsuspected and unacceptable features of any chosen prior. This has been called the "device of imaginary results" by I.J. Good, and has been extensively promoted by him (c.f. Good(1965, 1976, 1981)).

More formally, frequentist measures, such as operating characteristic curves and risk functions can be of interest through their relationship to procedure robustness. (This was briefly discussed in subsection 3.2, but, since the issue is quite controversial, an expanded discussion is in order.) The basic reason for this relationship is (1.2), namely that

$$(4.1) \qquad E^m\rho(\pi,X,\delta(X)) = r(\pi,\delta) = E^\pi R(\theta,\delta).$$

(Although $R(\theta,\delta)$ and $\rho(\pi,x,\delta(x))$ were defined as frequentist risk and posterior expected loss, respectively, through appropriate choice of the loss function they can be made to represent non-decision theoretic measures such as coverage probability and posterior probability of containing θ, respectively.) If, now, $R(\theta,\delta)$ is known to be "good" for all θ, then from (4.1) it follows that $E^m\rho(\pi,X,\delta(X))$ will be "good" for all π. Although this doesn't guarantee that $\rho(\pi,x,\delta(x))$ is actually good for the observed x and π of interest, there is a good chance that it will be. Conversely, if $R(\theta,\delta)$ is bad for some θ, then before using δ it is imperative to make sure that such θ are really very unlikely apriori. In Example 4, for instance,

$$R(\theta,\delta^N) = .471 + (.0983)\theta^2,$$

which is terrible for large θ. Looking at this risk would cause one to realize that, unless the large θ really are as unlikely (subjectively) as indicated by the tail of the presumed normal prior, then use of δ^N may not be wise.

Besides this aspect of using frequency measures as a check on Bayesian robustness, two closely related reasons for admitting consideration of frequency measures should be discussed. First, there are simply many problems which have a good frequency answer, and yet which do not have clearly trustworthy Bayesian answers. Because of (4.1), the frequency procedure has a good chance of also being sensible from a conditional posterior Bayesian viewpoint. Thus it can be viewed as a good "stab in the dark". Of course, as Bayesian methodology expands, there will be less and less need to depend on such frequency evaluations. (See Berger (1982d) for examples, discussion, and references.)

The final reason for consideration of frequency measures and procedure robustness is that, like it or not, the majority of users of statistics are not going to be extremely well trained, and will probably not be capable of careful Bayesian sensitivity analyses. For such users it is necessary to provide procedures, which are as Bayesian as possible, and yet are automatically robust. Since these procedures will be used repeatedly, their long run frequency performance is definitely relevant. Example 4, for instance, suggests that in estimating a normal mean it would be reasonable to ask the unsophisticated user to specify a "guess" and an estimate of the accuracy of this guess, and then fit this to a Cauchy prior and calculate the Bayes estimate (all of which could be automatically done by a computer). Fitting to a conjugate normal prior is contraindicated, however, at least for such automatic use. This section concludes with a very brief review of some useful frequency concepts.

A. Design, Prediction, and Sequential Analysis

In these problems it is absolutely imperative to average over the data likely to occur, and no Bayesian would think otherwise. Of course, these problems also have a large Bayesian component. In design, for instance, one must use subjective guesses for θ to predict what data will occur and hence what design to use. Also, a Bayesian will have the goal of obtaining good conditional performance, which may lead to a quite different design than a classical design.

B. Confidence Procedures

If C(x) is a confidence procedure for θ with confidence level $1-\alpha$, then

$$P_\theta(C(X) \text{ contains } \theta) \geq 1-\alpha.$$

As in (4.1), it follows that

$$(4.2) \qquad E^m P^{\pi(\theta|X)}(\theta \in C(X)) \geq 1-\alpha,$$

so that, for small α, C(x) has a pretty good "chance" of containing θ (according to a valid posterior measure) no matter what π is. This use of confidence procedures was discussed in Pratt (1965).

Morris (1981, 1982b) has advocated the development of procedures satisfying (4.2) for all priors π in a feasible class Γ, and has called this "empirical Bayes confidence". For the reasons discussed earlier, this may well be a valid objective, as long as it is kept in mind that the real goal is to obtain a set with good posterior probability of containing θ for the given observation x. Similar ideas are employed in Godambe and Thompson (1976) and Godambe (1982) to argue for use of frequentist concepts in obtaining robust Bayesian confidence procedures in survey sampling. Other work on the relationship between frequency and Bayesian confidence methods can be found in Welch and Peers (1963) and Stein (1981b), which also contain earlier references.

C. Minimaxity

The robust Bayesian interest in minimaxity arises from the fact that

$$\sup_{\pi} r(\pi,\delta) = \sup_{\theta} R(\theta,\delta), \tag{4.3}$$

and hence a minimax decision rule (i.e. a rule minimizing the right hand side of (4.1)) is also the "most procedure robust" Bayesian decision rule (being Γ-minimax when Γ is the class of all priors). Although realistic Γ will rarely be so large that

$$\sup_{\pi\in\Gamma} r(\pi,\delta) = \sup_{\theta} R(\theta,\delta),$$

a minimax rule can provide a basis of comparison for procedure robustness.

D. Admissibility

If δ is inadmissible, there will often exist a δ^* such that

$$R(\theta,\delta^*) < R(\theta,\delta)$$

for all θ, and hence such that $r(\pi,\delta^*) < r(\pi,\delta)$ for all priors π for which the Bayes risk exists. Because of procedure robustness and (4.1), it can be convincingly argued that this should preclude consideration of inadmissible decision rules. (See also Hill (1974).) The restriction to consideration of only admissible rules can be a very helpful reduction of the problem, particularly in areas such as sequential Bayesian analysis where even determination of a Bayes rule can be very difficult.

E. Asymptotics

Much of the frequentist work on asymptotics has relevance to a Bayesian. Some such work is discussed in Section 5. Also, asymptotics can be helpful in determining Bayesian robustness. For example, in Diaconis and Freedman (1981) it is shown that certain partially nonparametric Bayes rules can be inconsistent, giving real cause for concern as to the robustness of use of the corresponding priors.

F. Significance Testing

There are sometimes relationships between P-values in significance testing of a hypothesis and posterior probabilities of the hypothesis (c.f. Good (1950), Jeffreys (1961), Pratt (1965),and Berger and Wolpert (1982b) which has later references), and this may sometimes justify use of the often much easier to compute P-values. Also, in Section 5 the role of Bayesian significance testing in Bayesian robustness will be briefly discussed.

We have, of course, barely touched the surface of the possible uses of frequency concepts in robust Bayesian analysis. Invariance concepts, for instance, can have many uses. Also, many explicit frequentist procedures turn out to be perfectly satisfactory from a Bayesian viewpoint.

4.5 Estimating a Multivariate Mean: The Stein Effect

We conclude this section with an example interesting from several aspects. First, it is an example wherein both the frequentist decision-theorist and the robust Bayesian decision-theorist end up wanting to solve the same problem. Second, it is an example wherein the Bayesian can be amazingly robust and the frequentist can make significant use of prior information at no or little cost. Finally, it illustrates the fact that good robust Bayes procedures need not be Bayes procedures for any prior in Γ, and indeed can violate natural Bayesian intuition.

Suppose we must simultaneously deal with p independent estimation problems ($p \geq 3$), where $X_i \sim \mathcal{N}(\theta_i,1)$ is the observation in the ith problem, and the loss in estimating $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_p)$ by $\underset{\sim}{\delta} = (\delta_1,\ldots,\delta_p)$ is $\sum_{i=1}^{p} (\theta_i-\delta_i)^2$. The θ_i are apriori known to be independent and, as a quick approximation, are felt to have $\mathcal{N}(0,1)$ prior distributions, to be denoted π_i^N, $i=1,\ldots,p$. (Different prior medians could be allowed in the following analysis.) This last facet of the prior distribution is deemed uncertain, however, and hence robustness is sought with respect to the class of priors

$$(4.4) \qquad \Gamma = \{\pi = \prod_{i=1}^{p} \pi_i\colon\ \pi_i=(1-\varepsilon)\pi_i^N+\varepsilon P_i,$$
$$P_i \text{ arbitrary probability measures}\} .$$

(It is essentially certain that the θ_i are apriori independent, and ε is the assumed error in the approximations π_i^N.)

A non-Bayesian frequentist analysis of the problem must take note of the Stein phenomenon, which is that estimators δ^* exist which are better than the natural estimator $\delta^0(x) = x$, i.e.,

$$(4.5) \qquad R(\theta,\delta^*) < R(\theta,\delta^0) = p \quad \text{for all } \theta.$$

The frequentist finds himself forced somewhat into the Bayesian ballpark, however, since any such δ^* is significantly better than δ^0 only in a relatively small region of the parameter space. Intuitively, therefore, δ^* should be selected by deciding where significant improvement is most desired, and it seems manifest that significant improvement will be most desired for those θ felt likely to occur apriori. A very reasonable way of proceeding, therefore, is to elicit a rough prior distribution π_A, and then to find that δ^* which minimizes $r(\pi_A,\delta^*)$ subject to (4.5). (Such a δ^* will clearly perform best for those θ felt apriori to be most likely.) The frequentist willing to sacrifice some minimaxity (here p is the minimax risk) for more Bayesian gain would be interested in the problem

(4.6) $$\text{Minimize } r(\pi_A,\delta), \text{ subject to } R(\theta,\delta) \leq p + C.$$

A fascinating feature of this situation is that a robust Bayesian can become concerned with the same problem. Indeed, suppose he seeks procedure robustness by trying to be Γ-minimax (see (3.2)) with respect to the Γ in (4.4), and furthermore does the "obvious" thing and restricts attention to coordinatewise independent rules, i.e., rules of the form

(4.7) $$\delta(x) = (\delta_1(x_1),\delta_2(x_2),\ldots,\delta_p(x_p)).$$

(Since the θ_i are apriori independent, any Bayes rule with respect to a prior in Γ will be of this form.) A relatively simple game theoretic argument shows that this problem is then equivalent to the problem in (4.6) (with δ restricted to be of the form (4.7), of course), in that there exists a continuous increasing function ρ such that $C = \rho(\varepsilon)$ defines an equivalence of solutions. It is interesting to see what happens if the restriction to estimators of the form (4.7) is dropped, so we will consider the general problem posed in (4.6).

Exact results on problems of this form are very complicated but simple approximate solutions are given in Berger (1982b) and Berger (1982c). For the special case considered here, and when C = 0 in (4.6) for simplicity, the approximate solutions are

$$\delta^*(x) = \begin{cases} \frac{1}{2} x & \text{if} \quad |x|^2 \leq 4(p-2) \\ \left(1 - \frac{2(p-2)}{|x|^2}\right) x & \text{if} \quad |x|^2 \geq 4(p-2) \end{cases} .$$

This estimator is minimax, and hence,not only satisfactory from the frequentist viewpoint, but also procedure robust with respect to the class of <u>all</u> priors. The estimator is also quite acceptable from the posterior viewpoint, since for $|x|^2 \leq 4(p-2)$

$$\delta^*(x) = \frac{1}{2} x = \delta^N(x) ,$$

where δ^N is the Bayes procedure with respect to the approximate prior $\pi^N = \prod_{i=1}^{p} \pi_i^N$. (For the class Γ in (4.4), posterior robustness is achieved for small $|x|$ by any Bayes rule with respect to a prior in the class, while for large $|x|$ posterior robustness is not attainable.) As to Bayes risk, this estimator astonishingly has

$$\lambda = r(\pi^N,\delta^*)/r(\pi^N,\delta^N)$$

as indicated in Table 2.

Table 2. Bayes Risk Ratio of δ^* to δ^N

p	3	4	5	6	7	8	9	10	15	20
λ	1.296	1.135	1.0727	1.0427	1.0267	1.0174	1.0117	1.008	1.0016	1.0004

Thus when p = 5, for instance, δ^* is only 7% worse than δ^N if π^N is the true prior. Indeed $\sup_{\pi \in \Gamma} r(\pi,\delta^*)$ will be very satisfactory, as indicated by the crude upper bound

$$\sup_{\pi \in \Gamma} r(\pi,\delta^*) < (1-p\varepsilon)r(\pi^N,\delta^*) + p\varepsilon \ .$$

(Compare this with the fact that

$$\sup_{\pi \in \Gamma} r(\pi,\delta^N) = \infty \ .)$$

That one can have such fine Bayesian performance and be so robust (or, from a frequentist viewpoint, be minimax) is quite surprising. What is even more surprising from a Bayesian viewpoint is that we <u>know</u> apriori that the θ_i are independent, and hence we <u>know</u> that our "true Bayes rule" would be of the form (4.7). But it is shown in Efron and Morris (1971) that if only estimators of this form are considered, then about the best that can be done is to have an estimator δ^T with $\lambda = 1.4$ and $\sup_\theta R(\theta,\delta^T) = (1.3)p$. This is 40% worse than δ^N when π^N is true, and 30% worse than a minimax estimator in terms of minimax risk (indicating considerably less procedure robustness), which is significantly inferior to the performance of δ^*. Hence good robust Bayesian procedures can differ substantially from what a straightforward Bayesian viewpoint might dictate, and need not be Bayes with respect to any prior in Γ. (The "formal" Bayesian solution to this problem of putting a metaprior on Γ would probably also work, although care might be needed in choosing the formal metaprior and the resulting procedure would probably be extremely messy.)

This example was, of course, very special, particularly in that the approximate priors for each θ_i were assumed to have equal variances. (Talking in terms of "variance" is convenient for specifying π_A, here, but Γ does not assume that the prior variance is known.) Almost certainly in reality, apriori independent θ_i will have different approximate prior variances. Some partial results for the general nonsymmetric situation can be found in Berger (1982b). Similarly,it will often be unrealistic to assume that the error in the specification of each of the π_i^N is the same value ε. The almost astounding power of the Stein effect in achieving Bayesian robustness in this "ideal" situation, however, certainly argues for its value in less ideal situations.

5. HISTORY AND GUIDELINES

There has been comparatively little research in Bayesian robustness, and only a few specific guidelines are available in attempting to achieve robustness. In subsection 5.1 we briefly review the literature on Bayesian robustness, although this was not intended as a review article per se, and hence little more than a categorization of results is attempted. In subsection 5.2 the few available guidelines are presented.

5.1 History

5.1.1 Posterior Robustness

A. Asymptotics

It is intuitively plausible that,as the sample size goes to infinity, the information from the data becomes conclusive, and hence the conclusions will depend very little on the prior (automatically achieving posterior robustness). Results in this area can be divided into the categories of "stable measurement", "consistency", and "sequential analysis". Summaries of much of this work can be found in DeGroot (1970).

A(i). Stable Measurement. The principle of stable measurement is roughly that, as the sample size goes to infinity, the posterior distribution of θ becomes essentially proportional to the likelihood function (i.e., the prior distribution washes out). This concept was extensively promoted by Savage (cf. Edwards, Lindeman, and Savage (1963) and most of the other works of Savage listed in the references). Blackwell and Dubins (1962) explored a similar concept.

Since the likelihood function will generally be asymptotically normal, it is reasonable to expect the posterior distribution to be asymptotically normal. Results in this direction were obtained by LeCam (1956), Johnson (1967, 1970), Walker (1969), Dawid (1970), Brunk and Pierce (1977), Heyde and Johnstone (1979), and Ghosh et. al. (1982).

One difficulty with stable measurement is that the sample size which is large enough for the asymptotics to apply will often depend on the observations themselves. Hence, in a sense, one is forced to do a complete posterior robustness check even for large samples.

A(ii). Consistency. Results concerning the consistency of Bayes estimates (and hence a degree of asymptotic robustness with respect to the prior distribution) can be found in LeCam (1953), Freedman (1963, 1965), Fabius (1964), Schwartz (1965), Berk (1966, 1970), Strasser (1981), DeRobertis and Hartigan (1981), and Diaconis and Freedman (1982). These results tend to say that,if θ is in the support of the prior distribution, then the Bayes estimates are consistent for θ, and otherwise they are not. The results of Freedman (1963, 1965) and Diaconis and Freedman (1982) indicate, however, that Bayes estimates can be inconsistent even when θ is in the support of the prior, unless care is taken in the selection of the prior.

A(iii). Sequential Analysis. Asymptotic sequential Bayes decision theory is concerned with sequential Bayes decision problems when the cost of each observation is very small. As the cost goes to zero, the number of observations likely to be taken goes to infinity, allowing the large sample Bayesian asymptotics discussed previously to apply. Most of the results on this subject obtain limiting forms of the Bayes stopping rule or Bayes risks. See, for instance, Chernoff (1959), Schwarz (1962, 1968), Kiefer and Sacks (1963), Bickel and Yahav (1967, 1969), Gleser and Kunte (1976), Fortus (1979), Vardi (1979a, 1979b), and Woodroofe (1980). Often, this limiting form is independent of the assumed prior distribution, indicating a large sample robustness. Certain seemingly robust nonasymptotic Bayes stopping rules for estimation problems can be found in Alvo

(1977). (See also Berger (1980b) for a general discussion.)

B. Sensitivity Theory.

Sensitivity analysis is a standard name for the process of investigating changes in the conclusions caused by changes in the initial assumptions (including the prior distributions). Such analysis is present in many good Bayesian papers. Dempster (1976) gives an interesting general discussion of this with examples. Any attempt to mention all such works would be nearly hopeless, so instead only the more formal works concerned with developing bounds on the range of the posterior conclusions based on variation in the assumed prior distributions will be mentioned. (Such works will be called Sensitivity Theory.)

B(i). Bounds on the Posterior Distributions. There have been many works seeking to bound the amount of variation in the posterior distribution itself (or certain posterior probabilities) for classes Γ of prior distributions, or the closely related "upper and lower probabilities". Results for classes of priors can be found in DeGroot (1970), Huber (1973), Chamberlain and Leamer (1976), Dickey (1976b), Leamer (1978), Davis (1979), Hill (1980c), Rios and Girón (1980), and DeRobertis and Hartigan (1980). (Some of these works are closely related to stable estimation.) Results in Stein (1965) are also relevant.

The idea of "upper and lower probabilities" is essentially to try and find upper and lower bounds on the prior distributions (these bounds will typically just be finite measures, i.e., will not have mass one), and from these obtain bounds on the posterior distributions. Such ideas can be found in Boole (1854), Koopman (1940), Good (1950, 1962a, 1976), Smith (1961), Dempster (1966, 1967, 1968, 1971), Beran (1970, 1971), Fine (1973), Huber and Strassen (1973), Kyburg (1974, 1976), Kleyle (1975), Suppes (1975), Williams (1976), Suppes and Zanotti (1977), West (1979), Levi (1980), DeRobertis and Hartigan (1981), and Wolfenson and Fine (1982), although several of these works propose alternative modes of reasoning based on the upper and lower probabilities.

B(ii). Bounds on Posterior Actions and Expected Loss. Sensitivity theory is often concerned with bounding the variation in the optimal posterior action or posterior expected loss caused by variation in the prior. Results for finite parameter spaces can be found in Isaacs (1963), Fishburn (1965), Fishburn, Murphy, and Isaacs (1968), and Pierce and Folks (1969). More general theories can be found in Skibinsky and Cote (1963), Dickey (1974, 1976b), Bansal (1978), Kadane and Chuang (1978), Rios and Girón (1980), and DeRobertis and Hartigan (1981). Leamer and Polasek (c.f. Leamer (1978) and Polasek (1983), which also contain earlier references) give bounds on the posterior Bayes action for a wide variety of problems involving variation of (hierarchical) conjugate priors, an analysis they call "global sensitivity" analysis. They also discuss "local sensitivity", which is essentially the rate of change of the posterior Bayes action with respect to change in the parameters of the conjugate prior. Although not generally as useful as global sensitivity, local sensitivity can be of assistance in identifying those prior parameters which have the greatest influence on the conclusion, and hence which must be considered most carefully.

C. Partial Prior Knowledge

There are a number of results in the literature concerned with determining reasonable posterior actions when only limited facets of the prior distribution are known. For example, Stone (1963), Hartigan (1969), and Goldstein (1974, 1979, 1980) consider estimation problems where knowledge is available concerning only the first two moments of the prior distribution. The estimators that result from such an assumption are linear estimators, and much of the huge literature on linear estimation (including much of linear filtering theory in stochastic processes) can be recast in this light. A serious concern is that prior moments are almost never knowable (see subsection 2.3), and that resulting linear estimators will often not be robust (see also subsection 5.1.3.A).

Other analyses based on limited prior knowledge can be found in Godambe and Thompson (1971), Hill (1975), Leamer (1978), Levi (1980), and Lambert and Duncan (1981).

D. Detecting a Lack of Posterior Robustness.

It is particularly important to identify common statistical situations in which posterior robustness is lacking, since such situations call for very careful consideration of prior information.

When the likelihood function is flat, the prior distribution will be the main factor in determining the posterior distribution, and hence the conclusions are liable to be very sensitive to the prior. This commonly occurs in high dimensional situations, where due to such problems as multicollinearity or often simply a lack of sufficient data for all the parameters of interest, the likelihood function will be flat in certain directions. Among the many discussions of this issue are Hill (1977), Leamer (1978), Hill (1980a), Posasek (1983), and Smith and Campbell (1980). The latter article addresses this problem in a critique of ridge regression, and references a number of other ridge regression papers dealing with the same issue.

Another situation in which the likelihood function is flat is in the random model analysis of variance when the usual unbiased estimator of the between variance component is negative. This is discussed in Hill (1965), Hill (1970), and Hill (1980a).

The value of m(x) (the marginal density of X) can be of use in determining robustness, in that a particularly small value of m(x) indicates that surprising data has occurred; the data and the prior information would seem to be in conflict. In such situations the likelihood function will tend to be concentrated in the tail of the prior distribution, a very uncertain part of the prior. Of course, the initial implication of a small value of m(x) is that the situation was incorrectly modeled, and hence (prior) assumptions concerning the data model need to be reconsidered or discarded. Excellent discussions of this and other references can be found in Jeffreys (1961), Dempster (1971, 1975), Box and Tiao (1973), Geisser and Eddy (1979), Box (1980), and Good (1965, 1981).

5.1.2 Procedure Robustness

A. Asymptotic Bayes Risk.

One can work with decision problems and Bayes risk $r(\pi,\delta)$ as the sample size goes to infinity. Asymptotic approximations to $r(\pi,\delta)$ are then available. Some work in this direction can be found in Chernoff (1952, 1956, 1970), Lindley (1960), Rubin and Sethuraman (1965), Rubin (1971, 1972), Johnson and Truax (1978), Burnasev (1979), Woodroofe (1980), and Ghosh et. al. (1982). Some of the articles mentioned in subsection 5.1.1(A.) are also of this type.

Interesting robustness phenomenon can occur, when asymptotics are considered, as shown in the following example due to Rubin (1971).

Example 5. Consider the situation of testing a "fuzzy" point null hypothesis. This concerns the reasonable formulation of the point null testing problem in which the null hypothesis can be phrased as H_0: $\theta \in \Theta_0 = (\theta_0-\varepsilon,\theta_0+\varepsilon)$, where ε is quite small. (Rubin (1971) formulates the problem solely in terms of losses, in which case $\theta_0 \pm \varepsilon$ are the points at which the losses in accepting and rejecting are equal.) The prior density $\pi(\theta)$ is assumed to have a sharp peak inside Θ_0, and to be fairly flat away from the peak. (This corresponds to common sense evaluations when a point null is involved.) Relevant, also, are the loss functions $L_A(\theta)$, of accepting H_0, and $L_R(\theta)$, of rejecting H_0. The Bayesian will be making a decision based on the weight function

$$W(\theta) = \pi(\theta)[L_A(\theta)-L_R(\theta)],$$

accepting H_0 if

$$\int w(\theta)p_\theta(x)d\theta < 0, \tag{5.1}$$

and rejecting otherwise (since (5.1) implies that the posterior expected loss of accepting is smaller than that of rejecting). The sample size is assumed to be large enough so approximate normality holds, i.e., $p_\theta(x)$ is $\mathcal{N}(\theta, \frac{\sigma^2}{n})$. Three cases must be distinguished:

(i) Somewhat large n (i.e., $\varepsilon << \frac{\sigma}{\sqrt{n}}$). In this situation H_0 can essentially by treated as a point null, in that

$$(5.2) \qquad \int W(\theta)p_\theta(x)d\theta \cong p_{\theta_0}(x) \int_{\Theta_0} W(\theta)d\theta + \int_{\theta \notin \Theta_0} W(\theta)p_\theta(x)d\theta \; .$$

The mass of the weight function in Θ_0 is comparatively easy to specify. Also, outside of Θ_0, $W(\theta)$ will be a fairly smooth function and, since n is somewhat large (so that the region with high likelihood is fairly small), it should be possibly to specify the last integral in (5.2) fairly accurately. Thus we have reasonable Bayesian robustness (i.e., the subjective inputs that are needed are fairly easy to elicit.)

(ii) Extremely large n (i.e., $\frac{\sigma}{\sqrt{n}} << \varepsilon$). Here it will essentially be known whether $\theta \in \Theta_0$ or not, so the prior will not matter. (Robustness with respect to ε could be a concern, however.) This situation is the usual "stable measurement" situation.

(iii) Moderately large n (i.e., all n not covered in cases (i) and (ii)). Here, surprisingly enough, robustness is lacking, in that the shape of $W(\theta)$ in Θ_0 is very important. (See Rubin (1971).) This is disturbing, in that determining the shape of the prior in this region is almost impossible. (Of course, the overall risk will be small since n is moderately large, but Rubin (1971) has shown that even mild misspecification of the shape of $W(\theta)$ can cause an increase of Bayes risk of 40% in the most favorable cases, with much larger increases in unfavorable cases.) The phenomenon observed in this example, of robustness for somewhat large n and extremely large n, but not for n in between, is striking.

B. Γ-Minimax and Γ-Minimax Regret Procedures

The Γ-minimax and Γ-minimax regret criteria (see subsection 3.2) are natural criteria to follow if procedure robustness is sought. The basic concepts were originally developed in Robbins (1951, 1964) and Good (1952). Other general discussions can be found in Menges (1966), Blum and Rosenblatt (1967), Kudo (1967), and Berger (1980b).

The Γ-minimax regret criterion seems somewhat more reasonable than the Γ-minimax criterion, in that it is based on the loss in risk by not using the theoretically optimal Bayes rule, rather than the absolute Bayes risk. The danger in using $r(\pi,\delta)$ itself is that there could be an "unfavorable" prior $\pi_0 \in \Gamma$ with excessively large Bayes risk

$$r(\pi_0) = \inf_{\delta} r(\pi_0,\delta),$$

in which case the Γ-minimax procedure would be the Bayes rule with respect to π_0. Unless there is some reason to be especially concerned with π_0, however, it would better to eliminate its prominence by using the Γ-minimax regret criterion. The Γ-minimax regret criterion will, on the other hand, be somewhat more difficult to work with, so if Γ contains no "unfavorable" prior it might be better to consider Γ-minimaxity.

Recall from subsection 2.3 that Γ should generally be specified in terms of percentiles and relative likelihoods. This has been done in the Γ-minimax literature on testing, multiple decision theory, and nonparametrics. The literature on estimation, however, makes unfortunate use of Γ specified by prior moments.

Results on Γ-minimax estimation can be found in Jackson, O'Donovan, Zimmer, and Deeley (1970), Solomon (1972a, 1972b), DeRouen and Mitchell (1974), Watson (1974), and Morris (1982a). Testing and multiple decision theory results can be found in Rubin (1965', 1971), Randles and Hollander (1971), Gupta and Huang (1975, 1977), Berger (1979), Gupta and Kim (1980), Gupta and Hsiao (1981), Miescke (1981), and Hsiao (1982). Some nonparametric Γ-minimax studies were done in Doksum (1970), Campbell and Hollander (1979), and Lambert and Duncan (1981).

C. Controlled Frequentist Risk

As discussed in subsection 3.3, the frequentist risk $R(\theta,\delta)$ can be a good indicator of procedure robustness. In particular, if

$$R(\theta,\delta) \leq C \tag{5.3}$$

for all θ, then $r(\pi,\delta) \leq C$ for all π, giving an upper bound on the possible harm from use of the procedure δ. Theoretical work finding bounds on the frequentist risk of Bayes estimators can be found in LeCam (1982), which also contains some earlier references. Studies of particular Bayesian estimators which seem to have good frequentist risk have been done in Novick (1969), Strawderman (1971), Lindley and Smith (1972), Efron and Morris (1972, 1973), Clevenson and Zidek (1975), Leonard (1976), Rubin (1977), Faith (1978), Berger (1979, 1980a, 1982a, 1982b, 1982c), Dey (1980), Dey and Berger (1980), Albert (1981), Berliner (1981), Ghosh and Parsian (1981), Hudson and Tsui (1981), Stein (1981), Berger and Wolpert (1982), Bock (1982), Wolpert and Berger (1982), and Zheng (1982).

A more systematic approach to the robustness problem is the restricted risk Bayes approach, initiated by Hodges and Lehmann (1952), which seeks to minimize the Bayes risk $r(\pi_0,\delta)$ for a chosen prior π_0, subject to the constraint (5.3). This guarantees robustness (in a conservative sense) with respect to the class of all priors. Interestingly, as discussed in subsection 4.5, the restricted risk Bayes problem often corresponds to the true Γ-minimax problem with

$$\Gamma = \{\pi\colon \pi(\cdot)=(1-\varepsilon)\pi_0(\cdot)+\varepsilon P(\cdot),\ \text{P arbitrary}\}\ ,$$

where ε, of course, depends on the C in (5.3). Results for the restricted risk Bayes problem can be found in Efron and Morris (1971), Shapiro (1972, 1975), Masreliez and Martin (1977), Bickel (1979), Marazzi (1980), and Berger (1982c, 1982b).

5.1.3 Robust Priors

The difficulty of working with a class Γ of priors makes very appealing the idea of finding prior distributions which give Bayes rules which are naturally robust with respect to reasonable misspecification of the prior. Indeed as Huber says in the discussion of Box (1980)

> "Essentially, by now the Bayesian approach should be concerned not with the ad hoc construction of super models but with deriving reliable guide-lines on how to choose the super model (within the inherent arbitrariness) so as to guarantee robustness, and how to do so in a best possible fashion."

A. Conjugate Priors are Often Not Robust.

Conjugate priors, by definition, have tails of the same type as the tails of the likelihood function; this can cause robustness problems as indicated in subsections 3.2 and 3.3. Priors with tails flatter than the tails of the likelihood function are generally superior (at least for estimation problems). This observation has been made in Anscombe (1963), Tiao and Zellner (1964), Lindley (1968), Dawid (1973), Hill (1974), Dickey (1974), Meeden and Isaacson (1977), Rubin (1977), Umbach (1978), Ramsay and Novick (1980), and Berger (1980a, 1980b). Rubin (1977) gives an excellent numerical study showing the value of choosing flatter tailed priors.

Incidentally, conjugate priors in estimation problems in exponential families tend to result in linear estimators (see Diaconis and Ylvisaker (1979), indicating a general lack of procedure robustness of linear estimators (except for those arising from noninformative priors). This can be seen directly by examining risk functions of linear estimators.

Of course, a major advantage of conjugate priors is that they are very easy to work with. Hence if posterior robustness is present, it is often appealing to use conjugate priors. If robustness is of concern, yet simple posteriors are desired, an attractive way to proceed in estimation problems is to use a (robust) flat tailed prior, calculate (usually numerically) moments (or maybe percentiles) of the posterior, and then match these to a distribution (usually conjugate) of desired simple form. For instance, if $\bar{X} \sim \mathcal{N}(\theta,1/n)$ is observed, and it is desired to estimate θ, a Cauchy prior will tend to be robust but will result in an ugly posterior. Calculating (numerically) the first two posterior moments and pretending that the posterior is normal with these moments should be reasonably accurate and will result in a posterior which is easy to communicate and use. Uses of this idea can be found in Bakan and Oleksenko (1977), Morris (1977), and Berger (1980b).

B. Noninformative Priors.

Noninformative priors are designed to be flat and as uninfluential as possible. They tend to work well (if carefully determined), and can hence be considered to provide robust solutions to problems where very little is known apriori. The literature on this subject is vast. Much of it is summarized (and other references are given) in Jeffreys (1961), Zellner (1971), Box and Tiao (1973), Bernardo (1979), and Berger (1980b).

There are problems with the use of noninformative priors, however, principally the arbitrariness in their definition. (Bernardo (1979) seems to have the most workable definition, of what he calls reference priors.) Hence even the user of noninformative priors should be concerned with robustness with respect to the class of reasonable noninformative priors. Also, if a noninformative priors is being used as an approximation to a vague proper prior, it is wise to, at least informally, verify that the results obtained are suitable for vague proper priors.

In testing problems, standard noninformative priors cannot be used when they give infinite mass to one of the hypotheses. Such situations can be handled (in a robust fashion) by use of "reference informative priors" (c.f. Jeffreys (1961) and Zellner (1982)).

C. Priors on the Boundary of Admissibility.

While flat-tailed priors tend to be desirable, priors with tails that are too flat may give rise to inadmissible decision rules, especially in higher dimensions. The most important example is estimation of a p-variate ($p \geq 3$) normal mean under quadratic loss (although almost any sensible loss gives similar results). The usual estimator (the vector of sample means or the least squares estimator in a linear regression) is the (generalized) Bayes estimator with respect to the (noninformative) uniform generalized prior on $\mathbb{R}^p$. This estimator is inadmissible, because the prior has tails which are too flat.

Much of the recent work in admissibility has been to find the "boundary of admissibility" in various problems. Priors with tails flatter than those on the "boundary" will tend to give inadmissible decision rules, while priors with sharper tails will tend to give admissible decision rules. Since flat tails are desirable for robustness, yet inadmissible decision rules are unappealing, priors on this "boundary" are natural choices for use. Results of this nature can be found in Stein (1965, 1981), Brown (1971, 1979), Strawderman (1971), Strawderman and Cohen (1971), Berger (1976a, 1976b, 1980a, 1982c), Srinivasan (1980), Berliner (1981), Ghosh and Parsian (1981), Berger, Berliner and Zaman (1982), and Hwang (1982a, 1982b).

D. Maximum Entropy and Reference Priors

An appealing idea when faced with a class Γ of possible priors is to choose that prior which maximizes entropy or some measure of loss, or minimizes some measure of information. Such priors are likely to lead to robustness, in that they are as noninformative as possible subject to being in Γ, and have been called "minimax information" priors (Good (1968)), "maximum entropy" priors (Jaynes (1968, 1981) and Rosenkranz (1977)), and "reference" priors (Bernardo (1979, 1981)).

The most extensively developed such theory is that of maximum entropy priors, much of the development being due to E.T. Jaynes. While I would call the theory highly successful, there are certain difficulties which are cause for concern. First, when Θ is infinite and the partial prior knowledge is (sensibly) the specification of certain percentiles, the maximum entropy prior does not exist. Even when Θ is bounded, the maximum entropy prior in this situation will have unpleasant jumps. Finally, it is not really clear that a maximum entropy prior will be robust. For example, if $\Theta = R^1$ and the first two prior moments are known (an unrealistic assumption, of course), then the maximum entropy prior is normal with the given moments. Although this is not terribly unreasonable when the first two prior moments are exactly known, it still seems preferable to use a flatter tailed prior; say, a t-distribution with the given moments and a small number of degrees of freedom.

E. Multistage Bayes Priors

Multistage (or hierarchical) priors are priors composed of several stages: at stage one the prior is assumed to be of a given functional form (usually the conjugate prior form) with unknown parameters (called hyperparameters); at stage two these parameters are given a prior distribution with possibly unknown hyperparameters; with the process repeating until the final stage (seldom more than the third stage), at which point a completely specified prior distribution (often noninformative) is given to the hyperparameters of the preceeding stage. Such priors are particularly useful in multivariate situations where relationships among the parameters are thought to exist and can be modeled in stages. They are also a useful enrichment of the class of conjugate priors when either robustness or more flexibility is sought, in that Bayesian calculations can be done in stages with these priors and will often be relatively easy if the first stage is of a conjugate form.

A multistage prior can, of course, be thought of as a single stage prior; merely integrate out the mulitstage prior over all hyperparameters.

The robustness of the multistage prior follows from the fact that, virtually always, the single stage version has flat tails. If, for example, $\Theta = R^1$ and the first stage prior is $\mathcal{N}(\mu,\tau^2)$, putting a prior on τ^2 and integrating will usually result in a flat tailed prior.

The literature on multistage priors is too large to be mentioned here. Good (1952) was the first to extensively discuss the technique, and has a very substantial body of work on the subject and its relationship to Bayesian robustness (c.f. Good (1980, 1981)). Lindley and Smith (1972) is also an important landmark.

F. Empirical Bayes Priors.

If $X_1,\ldots,X_n$ are observed and the X_i have distributions depending on θ_i, where the θ_i can be assumed to be generated from a particular prior distribution π_0, then π_0 can itself often be estimated from the data. This is the empirical Bayes idea, first formalized by Robbins (c.f. Robbins (1955, 1964)). The approach is particularly easy if π_0 is chosen to be of a known functional form (say the conjugate form) with unknown hyperparameters, and these hyperparameters are estimated from the data. (This is then actually very closely related to the multistage Bayes approach, with similar answers being obtained under either method.) Providing <u>all</u> the data is used to estimate the hyperparameters (as opposed to, say, using just "past data" to estimate the hyperparameters) the resulting prior seems to be quite robust. This is because "extreme" data (the bane of nonrobust priors) will tend to give hyperparameter estimates leading to flat priors. For more thorough discussion of this see Berger (1980b).

The empirical Bayes literature is also too large to mention. Good discussions and references can be found in Maritz (1970), Berger (1980b), and Morris (1982b).

5.2 Guidelines

The (woefully) few guidelines that have been discussed for achieving Bayesian robustness are summarized here, with a few additional observations.

5.2.1 General Considerations.

As stressed in subsection 2.3 and elsewhere, it is very important to consider robustness with respect to reasonable classes of priors. Unfortunately, easy to work with classes, such as classes of conjugate priors and classes based on prior moments, are usually unsuitable.

It cannot be overemphasized that if posterior robustness obtains, for the data at hand, then the search is ended. This can often best be discovered by simply varying the prior (over Γ) and seeing how the conclusion changes. Increasingly easy to use interactive computer systems should eventually make this relatively easy to do. It will often suffice to merely check posterior robustness for several, fairly different, priors in Γ. For instance, in Example 3 (subsection 2.3), if posterior robustness with respect to the normal and Cauchy priors is present, then posterior robustness with respect to all of Γ probably also obtains. Two useful indicators of a lack of posterior robustness are a flat likelihood function (or more commonly a likelihood function which is flat in certain directions of Θ) and a surprisingly small value of $m(x)$.

When posterior robustness is lacking, the situation must be reconsidered. First of all, one naturally looks for experimental causes or modeling failures accounting for this unpleasant situation. If nothing is turned up, further refinement of Γ is called for. If the limit of the elicitation process has been reached, however, then now, and only now, does procedure robustness and the possible use of frequency concepts (see Section 4.4) come into play. (Of course, if one is developing procedures for automatic use by nonsophisticated users, then posterior robustness is relevant from the start. To many, this may be deemed to be a major purpose of the theoretical statistician.) One could formally attempt some type of Γ-minimax or Γ-minimax regret analysis, but this will tend to prove enormously difficult. Indications of a lack of posterior robustness can be obtained from frequentist measures of the performance of a procedure; if the frequentist measure looks bad for certain Γ which are not completely implausible, concern is indicated.

A natural Bayesian attempt to obtain procedure robustness would be to put a metaprior on Γ itself. Since we are assuming that the elicitation process has ended, this would be merely a technical device to hopefully achieve robustness. Experience indicates that this probably works reasonably well, although it is difficult to do. (For one example, see Dickey and Freeman (1975).) It will usually be nearly impossible to construct a reasonable meta prior with support equal to all of Γ, so careful selection of a representative subset of Γ on which to place the meta prior would be needed. Note that this "two stage" prior could be written as a one stage prior, and hence the technique can be interpreted as simply a way of constructing hopefully robust priors.

Due to the difficulties of formally working with Γ for procedure robustness, it may simply be best to investigate the robustness of a procedure with respect to a few carefully chosen disparate priors in Γ.

The material on "robust priors" in the preceding subsection will not be repeated here, although it is certainly relevant to general guidelines.

5.2.2 Guidelines for Particular Types of Problems.

The following essentially obvious comments are not too much better than nothing, but may sometimes prove helpful.

A. Estimation.

Posterior robustness will typically be obtained when the likelihood function is concentrated in the "central" portion of the prior. (This "center" will usually be similar for all π in Γ.) When this is not the case, flat tailed priors will at least give procedure robustness. Note that,in multivariate estimation problems,it will often be the case that the robustness situation is very different for different coordinates of θ.

B. Testing.

In testing problems the tail of the prior will usually be unimportant (in contrast to the estimation situation), in that if the likelihood function is concentrated in the tail of the prior there is usually very strong evidence for a particular hypothesis. This robustness with respect to the tail of the prior is very pleasant. Note, however, that the posterior odds of the hypotheses can be drastically affected by the tail of the prior (as pointed out by Savage et. al. (1962)), so Bayesian measures of the strength of the conclusion are not necessarily robust.

Conclusions in testing problems will, naturally, be frequently sensitive to the prior mass given each hypothesis. This is unavoidable and, to a Bayesian, completely sensible.

C. Design of Experiments and Sequential Analysis.

Optimal Bayesian designs are usually robust with respect to small changes in the prior, such as changes in the tail. At least this is true when overall average measures of performance (say Bayes risk) are deemed relevant, since such averages are dominated by the contributions from the "likely" θ, or alternatively the "likely" x. (Of course, after taking the data and being faced with the need to draw some conclusion, robustness may have to be completely reevaluated.)

Note that, in design, there may be real technical advantages in working with frequentist measures averaged over the prior, rather than posterior measures averaged over the marginal distribution of X (which is more instinctively appealing to a Bayesian). This is because the (decision) procedure to be used may be fairly accurately known (say, when the sample size will be moderately large), so that involving the (uncertain) prior only at the last stage can lead to a technically easier robustness analysis.

Sequential analysis is, in a sense, just a design problem, in that the real difficulty is deciding, at a given stage, whether to cease sampling or to continue taking observations. This problem should again be relatively immune to such things as the tail of the posterior distribution (upon which the decision to stop or not is based). Of course, if the likelihood becomes concentrated in the tail of the original prior, then this tail can become relevant through its effect on the posterior.

In a very practical sense there may be little problem with robustness in sequential Bayesian analysis, since it will often be the case that one simply continues sampling until enough information has been accumulated so that posterior robustness obtains.

5.2.3 Actual Practice.

In many realistic statistical situations involving complicated Θ, any type of Bayesian approach becomes very difficult. Also, the uncertainties in specifying a prior for such Θ are very acute, meaning that Bayesian robustness becomes a very real concern. Unfortunately, one quickly encounters an instance of "Type II rationality" (c.f. Good (1973)) in that, if straightforward Bayesian analysis is difficult, then a robust Bayesian analysis might be next to impossible. Type II rationality simply says, in this situation, that if you cannot trust a single prior Bayes analysis and Bayesian robustness results are unavailable, then it is permissible to use some type of non-Bayesian analysis, providing it is deemed to be the lesser evil. In other words, if the dangers of a Bayesian analysis with an ill-specified prior seem large (and cannot be eliminated by robustness considerations), and if an easier non-Bayesian or partially non-Bayesian analysis seems sensible (see subsection 4.4), then go ahead and abandon ship.

The need to compromise the "purist" robust Bayesian position was already encountered in subsection 2.4, where post-data modification of Γ was discussed. (Of course, allowing such modification somewhat alleviates the current problem, since all prior knowledge concerning a complicated Θ need not then be exactingly quantified prior to experimentation.) This compromise was still within the general Bayesian framework, however.

A more significant departure from the usual Bayesian framework occurs when it is necessary to ignore data. Such situations surely abound in statistics. Survey sampling provides one such situation, in that all sorts of data may be available about the sample, most of which may seem irrelevant to the attribute of interest. Constructing a general Bayesian (superpopulation) model for all the data would be very difficult and, perhaps, would not be trustworthy. The same issue arises in the use of randomization, as discussed in subsection 4.2. Hildreth (1963), Pratt (1965, with his discussion on "insufficient statistics"), Dempster (1968), Hill (1975, 1980a, 1980c), and Good (1976 and earlier with his "Statistician's Stooge") also contain useful discussions on this issue.

Ignoring data causes no real problem to a Bayesian if the data seems unlikely to have an effect on the posterior distribution of the parameters of interest. Often, of course, this can only be ascertained through, at least informal,Bayesian reasoning. Consider the following examples.

Example 6. (Fraser and Mackay (1976)). Suppose independent observations $X_1,\dots,X_n$ from an $\mathcal{N}(\mu,\sigma^2)$ distribution are observed, where it is desired to estimate μ but σ^2 is also unknown. Independent observations $Y_1,\dots,Y_m$ are also available, where Y_i is $\mathcal{N}(\mu_i,\sigma^2)$, μ_i unknown, $i=1,\dots,p$. If virtually nothing is known apriori about the μ_i (and they are in no way related to μ), it is certainly reasonable to ignore the Y_i when estimating μ. (A formal Bayesian analysis would certainly show that the Y_i had almost negligible influence on the posterior distribution of μ.)

Example 7. In a medical trial comparing two surgical techniques, a significant relationship was found between the time of the day in which the surgery was performed and the success of the surgery. Suppose the relationship was one of the following: (i) the later in the day the surgery occurred, the less successful it was; (ii) when surgery began on even hours, it was more successful than when it began on odd hours; (iii) when surgery ended on an even minute, it was more successful than when it ended on an odd minute. The question before us is - can we ignore the data "time of day"? The answer in case (i) is almost certainly no, and we better hope that the two treatment groups were not unbalanced concerning this covariate. The answer in case (iii) is almost certainly yes; it is hard to believe that this relationship is anything more than a coincidence. In case (ii) the answer is not so certain, and indeed some investigation is called for. (Did certain surgeons work at certain times, etc.?)

The decision about ignoring data in Example 7 clearly involves prior opinions. The point, however, is that it may be possible to informally reason that certain data can be ignored, without having to go through a full blown Bayesian analysis. This is not really a violation of Bayesian principles either, since the posteriors obtained by ignoring part of the data are felt to be the same as what would have been obtained by a sound Bayesian anlysis with all of the data.

The real difficulty arises when it is necessary to throw away potentially relevant data. The reason for doing this would be an inability to carry out a (robust) Bayesian analysis involving everything. Hill (1975) considers a nonparametric problem of this nature, in which a trustworthy complete Bayesian analysis seems almost impossible. Hill says

> "When such a formal analysis simply cannot be made, or even when it is merely very difficult and of dubious validity, then there is little choice but to condition on that part of the data that can be effectively dealt with, and rely upon some form of stable estimation argument."

The last part of the comment can be interpreted to mean that, if you must ignore data, at least convince yourself that there is no reason to expect it to have a large effect on the posterior (or, more properly, the final conclusion). This point is extensively discussed in Pratt (1965) and also Dempster (1975) which has interesting examples and other references.

The above discussion is not to be interpreted as advocating the frequently encountered viewpoint of "using whatever approach works well for a given problem". Indeed the major point in this article is that the only way to ensure that a conclusion being reached is sensible it to verify that it is sensible from a posterior robust Bayesian viewpoint. But if a robust Bayesian analysis is not implementable, then compromises must be made. The robust Bayesian makes this compromise only when he has to, however, and only to the extent necessitated by technical limitations.

Perhaps the most important practical advice the robust Bayesian has to offer is "think like a robust Bayesian". (In the same way,it has been argued that the most important thing to learn from decision theory is simply the ability to think decision-theoretically.) Merely thinking of problems from this perspective, without even doing a formal analysis, will frequently illuminate the truth. Once the truth (or the direction in which it lies) has been discerned, a method of analysis can undoubtedly be found which is acceptable to the relevant audience and leads to this truth.

6. FINAL COMMENTS

In an (obviously unsuccessful) effort to keep from meandering from the central argument, a number of side issues have been deferred to this final section. These include a discussion of various criticisms that can be raised against the robust Bayesian viewpoint (by non-Bayesians, Bayesians, and Foundationalists), and a very brief discussion of needed theoretical developments. Many of the criticisms are founded on very deep issues, so all that can be done here is to give a superficial view of the arguments and counter arguments.

6.1 Non-Bayesian Criticisms

The primary non-Bayesian objection to the robust Bayesian viewpoint is, of course, that Assumption I is wrong. Since an extensive justification of Assumption I was not attempted here, this objection will not be pursued, except for one brief comment. Much of the philosophical difference in attitude between Bayesians and non-Bayesians seems to be due to Bayesians being optimistic about the existence of truth and pessimistic about the use of intuition, while non-Bayesians are just the opposite. The Bayesian feels there is (at least theoretically) a single correct way of doing things , not many correct ways. Also, the Bayesian (and the decision theorist) do not trust intuition to properly combine and relate all relevant factors of a problem to arrive at a conclusion.

Perhaps the most common non-Bayesian objection to anything Bayesian is the Bayesian's lack of objectivity. Bayesian rebuttals range from the subjectivist opinion that "objectivity" is a myth, to the objective Bayesian assertions that objectivity can only be attained if consciously sought from a Bayesian perspective. The former viewpoint is reflected by the following quotation from Good (1973).

> "The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science."

The objective Bayesian viewpoint is that the only way to avoid "biasing" the analysis is to do a Bayesian analysis with a noninformative prior distribution (see subsection 5.1.3(B)for references). Strong support for this view can be obtained from "Reason (ii)" in subsection 2.1. If a supposedly objective non-Bayesian procedure actually corresponds to a Bayesian procedure for a very biased prior distribution, the claim of objectivity seems somewhat silly. The vehement condemnation of the use of noninformative priors by some non-Bayesians is indeed somewhat mystifying, since subjective prior beliefs are not being incorporated. Of course, there are problems in finding and using noninformative priors, but I have seen no better, easier to use, and less error prone technique for deriving reasonable objective procedures. (Although when being a purist I would argue against the possibility of objectivity, for a variety of robust Bayesian and Type II Rationality reasons the noninformative prior approach seems extremely valuable.)

It should be mentioned that there are several different non-Bayesian theories that reject Assumption I. Besides the various classical theories, these include the fiducial inference of R.A. Fisher (see Wilkinson (1977) for an up-to-date version), the structural inference of D.A.S. Fraser (see Fraser and Mackay (1976) and Fraser (1979) for references), pivotal inference (c.f. Barnard (1981)) and the theory of belief functions (c.f. Shafer (1982)). Since we are foregoing a serious effort to justify Assumption I, these alternative theories will not be discussed.

6.2 Bayesian Criticisms

Many natural Bayesian objections to the viewpoint expressed in this paper, such as the violation of the Likelihood Principle (and to an extent Assumption I) by procedure robustness, have been discussed elsewhere. Several other criticisms can be raised, however. Three are discussed in this subsection.

A. "Just Report What the Data Says."

A very admirable Bayesian desire is to provide a mechanism by which the data can be easily assimilated to allow others to reach a conclusion. The likelihood function of θ, $\ell_x(\theta)$, is the most basic such mechanism, since anyone can determine his own posterior for θ by simply multiplying $\ell_x(\theta)$ by his prior (or priors) and normalizing. Thus reporting the likelihood function is definitely reasonable (c.f. Box and Tiao (1973)).

A similar idea (c.f. Bernardo (1979)) would be to just report a noninformative or reference posterior, since this will be more meaningful intuitively than $\ell_x(\theta)$, and anyone can easily still determine his own posterior. Considerable effort has also been spent on finding easier to digest data communication vehicles such as Bayes factors between hypotheses (c.f. Dickey (1973, 1974)). A criticism of the robust Bayesian position is that, if the above pursuit is the true job of the statistician, then he need not be concerned with robustness, which afterall only becomes of concern when data-prior interactions are being studied.

I must have stated the criticism unfairly, since it seems clearly unworthy. We cannot, after all, abandon the user at the critical point of combining the data with his prior information, particularly when some action or conclusion must be taken. Also, it is frequently impossible to even separate data from prior information in a useful way. For example, the usual likelihood function is very model dependent, but the model is often unknown and should be considered part of the parameter.

B. "Why Single Out the Prior? Model Robustness Is Just As Serious a Problem."

First of all, since the data model was allowed to be part of Θ, we did not really ignore model robustness. On the other hand, there are sometimes reasons to be more concerned about the parameters than the model. For example, the model may have some theoretical basis, while prior opinions about parameters of the model might be much more subjective. Of course, there are many problems in which the reverse is true, where the choice of a model is somewhat arbitrary and will have a much more profound effect on the answer than the choice of a prior on the parameters of the model.

Nevertheless, the prevalent statistical attitude is to trust models more than priors, and in dealing with this attitude the robust Bayesian viewpoint can be very helpful. Also, even when considerable uncertainty about the model exists, it may cause less of a problem than uncertainty about the prior information, as the following example indicates.

Example 8. Suppose $X_1,\ldots,X_n$ is an independent sample from a location density $f(x-\theta)$ on R^1, where $f(z)$ is symmetric and unimodal. It is deemed reasonable to model f as a t-distribution with quartiles $(\theta\pm1)$ but specification of the degrees of freedom, α, is judged to be impossible. Prior elicitation reveals that θ is thought to have median zero and quartiles ±1, with the prior having a symmetric unimodal density. It is desired to estimate θ. Although the model and prior uncertainties seem similar here, the likelihood function will be

$$\ell_x(\theta) = \prod_{i=1}^{p} f(x_i-\theta) \ ,$$

which, for even moderate n, will most likely have sharper tails than the prior. (The tail of $\ell_x(\theta)$ will be like the nth power of the tail of f.) This indicates that the robustness problem with respect to f will be less serious than that with respect to the prior.

C. "Robustness Is a Rare Problem and Can Be Dealt With Entirely Within the Bayesian Framework."

These issues have been discussed throughout the paper. I have argued that posterior robustness will be lacking in a significant portion of our problems (at least at the present stage of Bayesian development), and that techniques of proceeding, which at least partly lie outside the pure

Bayesian framework, can prove useful. Some people may argue that the non-Bayesian components of the robust Bayesian viewpoint will seldom be needed, while others will argue that it is these non-Bayesian components which will be of most use. This is exactly what we should be arguing about: what is the best method to achieve the robust Bayesian goal. (See also Berger (1982d).)

From a very pragmatic viewpoint, also, the pure Bayesian position strikes me as unwise. The only truly overwhelming problem facing Bayesians is that of convincing non-Bayesians that the Bayesian viewpoint is correct. The major stumbling block in the entire controversy is that Bayesians (as a whole, not individually) have not openly admitted the validity of Assumption II, and been willing to accept its consequences. This allows the non-Bayesian to refuse to think about Assumption I, because he feels certain that Assumption II is correct and hence that the Bayesians must be wrong. Again, I am talking about the overall image of Bayesian, and not necessarily about the viewpoints of particular Bayesians. Even accepting Assumption II, but staying within the purely Bayesian framework of posterior robustness, will not provide a general enough structure to satisfy many of the criticisms of non-Bayesians. The practicing Bayesian might find that he seldom needs to leave the pure Bayesian structure, and hence that procedure robustness, etc., are concepts only rarely needed, but having them available can never hurt and can, I believe, help substantially in promoting the cause.

6.3 Foundational Criticisms

While non-Bayesians attack Assumption I from above (loftily disdaining from grubbing around in subjective probabilities) certain foundationalists attack it from below (urging even deeper submersion in subjectivism). The issue is whether reasoning in terms of a class Γ of (countably additive) prior probability distributions with updating by conditioning (and post data modification of Γ) suffices, or whether more general or more basic concepts are needed. I will basically argue that the above concepts not only suffice, but are what we should train ourselves to think in terms of. The robust Bayesian viewpoint is not an attempt to model how intuition works, but rather an attempt to create a structure of components which are simple enough to be accessible to intuition, and which when combined give the truth. As Good (1976) says

> "The main merit that I claim for the Doogian philosophy is that it codifies and exemplifies an adequately complete and simple theory of rationality, complete in the sense that it is I believe not subject to the criticisms that are usually directed at other forms of Bayesianism, and simple in the sense that it attains realism with a minimum of machinery."

My rebuttal to the foundational criticisms will thus tend to be that the alternative structures proposed either have components which are not reasonably accessible to intuition, or have unnecessarily complicated structures. I, of course, admit that it may be personal taste, rather than sound reasoning, which leads me to reject these alternative theories. Also, through lack of careful enough study or just general thick-headedness I may misrepresent certain arguments or be foolish in my response to them, in which case I apologize and look forward to being set straight. In any case, besides being somewhat fun, these foundational issues serve well to illuminate the edges of the robust Bayesian theory.

A. Measurability Criticisms.

DeFinetti (1974, 1975) and Good (1962a) argue that sometimes it is inappropriate to stay within the confines of measurable events. The real concern here is that the data may cause one to desire an enlargement of the σ-field (of measurable events in Θ) that was originally chosen as adequate. This could be subsumed under post-data modification of Γ, of course. In any case, technical measurability concerns are certainly not particularly relevant to the validity of the robust Bayesian viewpoint.

At the other extreme, Manski (1981) and Lambert and Duncan (1981) argue that, in specifying a prior, the σ-field of measureable events should be restricted to those events about which prior information is to be elicited. This is somewhat appealing intuitively, since a single measurable prior with respect to this σ-field would correspond to a class Γ of priors in the usual setup with, say, the Borel σ-field. The difficulties with this approach are that (i) it is very hard to update σ-fields based on the data, surely an essential ingredient of the approach; and (ii) it will generally be much more revealing to investigate robustness by varying π over Γ, than to simply make conclusions within the restricted σ-field formulation. The point is that, by using the robust Bayesian framework, one is often alerted as to what features of the prior need special consideration. Restricted σ-fields may hide these facets of the prior. To some extent, my view may be based on simply feeling comfortable with usual probability distributions, and hence development of this alternate approach is of interest, but I would be surprised if it led to a more useful framework.

B. "Finitely Additive Priors Should Be Allowed."

Among Bayesians there is a considerable faction that believes that insisting on countably additive priors is too restrictive, in that such conceivably desirable priors as proper "uniform" priors on R^1 or on the integers are prohibited. DeFinetti has long argued this (c.f. deFinetti (1972, 1974, 1975)). Other persuasive cases have been made by Dubins (1975), Heath and Sudderth (1978), Kadane, Schervish, and Seidenfeld (1981), and Hill (1980b). The case for allowing finitely additive priors also rests on the fact that the "rationality" or "coherency" justifications of Bayesian analysis lead only to finitely additive priors, although under slightly stronger axioms countable additivity emerges (c.f. Savage (1954), DeGroot (1970), Spielman (1977), and Kadane, Schervish, and Seidenfeld (1981)).

The arguments for staying within the countably additive framework are that (i) the examples espousing a need for finite additivity are not really convincing; (ii) even if "uniform" type priors on unbounded spaces are needed, countably additive improper priors can be used; and (iii) finitely additive priors require extremely careful handling.

The first point is that, while convincing "thought" examples have been constructed of the need for finite additivity, I have not yet seen a real example, involving an actual real world action that must be taken, in which my prior opinions would be uniform on an unbounded set. I certainly admit, however, that situations will exist in which I might want to use a "noninformative" prior, either as a robust approximation to my true prior beliefs or in a situation in which the appearance of objectivity is deemed necessary. Such situations can be dealt with either by using improper countably additive "noninformative" priors or by using proper finitely additive noninformative priors. Each approach has certain theoretical drawbacks, which will not be discussed here. They both also have the practical drawback of there being no clearcut definition of noninformative priors, and a careless choice can give unsatisfactory results. (The situation for finitely additive priors is particularly bizarre: for instance, there are $2^{2^{\aleph_0}}$ different "diffuse" finitely additive priors on the positive integers, as opposed to the single (constant) countably additive diffuse (improper) prior.) The main difference, to me, lies in the ease with which they can be used. Finitely additive priors frequently fail to have the Radon-Nikodym property (that conditional probabilities on sets of measure zero can be uniquely defined as limits of conditional probabilities of sets of nonzero measure), and hence frequently do not have well defined conditional (posterior) distributions. (Heath and Sudderth (1978) discuss some common statistical situations involving amenable group structures where meaningful posterior distributions can be defined. See also Dubins (1969).) This makes the typical Bayesian conditional analysis very difficult or impossible in general. Also, unconditional Bayesian analysis can seem very silly if finitely additive priors are used, as the following example shows.

Example 9. It is desired to estimate, under squared error loss, a normal mean θ based on $\bar{X} \sim \mathcal{N}(\theta,1/n)$. If one were to be repeatedly faced with this problem (with different θ_i) then it might seem reasonable to ask for a procedure with good average Bayes risk $r(\pi,\delta)$. If nothing was felt known about the θ_i, one might be tempted to use a noninformative prior. Use of the improper countably additive uniform prior will give infinite risks, so one is alerted to approach the problem differently. The usual finitely additive uniform prior, however, has finite Bayes risk equal to 1/n, but there are many estimators which achieve this, among them $\delta^0(\bar{x}) = \bar{x}$ and $\delta^*(\bar{x}) = \bar{x} - 10^{1000}/\bar{x}$. A posterior analysis of the problem (which can be done here using Heath and Sudderth (1978)) shows that δ^0, not δ^*, is correct, but the need for careful unconditional handling of finitely additive priors is, at least, indicated.

It should be noted that there is no foundational reason not to allow finitely additive priors into the robust Bayesian framework, so that those who feel comfortable with them are invited to do so.

C. "The Use of Probability Distributions is Too Restrictive."

The first point, made initially by Kraft, Pratt, and Seidenberg (1959) (see also Fine (1973)), is that there may exist "likelihood orderings" of events that are internally consistent and yet which are not consistent with any probability distribution. Although unsettled by this fact, I would argue that it is irrelevant, in that I would myself heavily distrust any likelihood ordering not consistent with some probability distribution. The consistent modes of behavior are those induced by probability distributions, so I would rather take them as my "primitives" than I would a concept such as "likelihood orderings". This is another situation in which I am not concerned with modeling how the mind could work, but rather with developing a framework within which the mind can successfully work.

Many foundational theories have been proposed which are based on generalization of probability distributions. Various such attempts can be found in Koopman (1940), Good (1950, 1962a, 1976), Smith (1961), Dempster (1966, 1967, 1968, 1971), Jeffrey (1968), Beran (1972), Huber and Strassen (1973), Fine (1973), Kyburg (1974), Suppes (1975), Suppes and Zanotti (1977), Levi (1980), DeRobertis and Hartigan (1981), Wolfenson and Fine (1982), and Rios and Girón (1980). (Some of these deviate only slightly from the robust Bayesian approach, and hence are not really susceptible to the following criticisms.)

A starting point for several of these theories is a rather ill-considered criticism of prior probabilities. They often begin with a "counterexample" such as the following.

Example 10. Suppose you pull a coin from your pocket and, without looking at it, are interested in the event A that it will come up heads when flipped. Suppose you (reasonably) judge the subjective probability of this event to be close to $\frac{1}{2}$. Next, you contemplate an experiment in which two drugs, about which you know nothing, will be tested, and are interested in the event B, that Drug 1 is better than Drug 2. You (reasonably) judge your subjective probability of event B to also be $\frac{1}{2}$. The argument now proceeds:

"Even though both probabilities were $\frac{1}{2}$, you have a stronger 'belief' in the probability specified for event A, in that if you were told that five flips of the coin were all heads your opinion about the fairness of the coin would probably change very little, while if you were told that in tests on five patients Drug 1 worked better than Drug 2 you would probably change your opinion substantially about the worth of Drug 1." Thus, the argument goes, it is necessary to go beyond probability distributions and have measures of the "strength of belief" in probabilities.

It is easy to see the flaw in this reasoning. Before getting any data, I would be equally secure in probabilities of $\frac{1}{2}$ for each A and B, in that I would be indifferent between placing a single bet on either event. My knowledge about the events A and B is well described by a probability of $\frac{1}{2}$. However, my knowledge about the overall phenomena being investigated in each case is quite different. A description of my overall knowledge about the situations is more fully described by defining the unknown (and fictitious to a true subjective Bayesian) quantities p_C and p_M, reflecting the "true" proportion of heads and "true" proportion of patients for which Drug 1 would work better than Drug 2, respectively, and then quantifying prior distributions (or classes thereof) for p_C and p_M. The prior distributions for p_C will undoubtedly be much more tightly concentrated about $\frac{1}{2}$, than will the prior distributions for p_M. Note that the subjective probabilities of events A and B are just the means of the respective prior distributions. (I first saw an analysis of this common misconception done by D. Lindley, though I cannot recall the reference.)

Thus prior distributions prove to be rich enough to reflect whatever is reasonably desired. Even more interesting is the observation that, in taking account of experimental evidence, one is almost forced to think in the correct fashion. Thus, in Example 10, if at the beginning it was only felt necessary to quantify the probabilities of A and B, reflection on the experiment to be performed reveals that the data information can be combined with prior information via Bayes theorem only if prior information is specified in terms of quantities such as p_C and p_M.

A second, more substantial, reason that alternative theories to Bayesian analysis have been developed is the recognition of the validity of Assumption II, and the perception that Bayesian analysis could not incorporate this assumption (although there were numerous works, about 50 by I. J. Good alone I believe, indicating that Assumption II could be incorporated). Some of the approaches do suggest alternate methods of dealing with probabilistic uncertainty, such as using lower and upper probabilities. The robust Bayesian approach seems much more straightforward, however, and does not demand the introduction of all sorts of new and supposedly "intuitive" criteria. Indeed, I have seen no new criterion that is obviously trustworthy, and the very same reasoning that forced me to accept the Bayesian viewpoint, as opposed to the "intuitive" classical viewpoint, argues against the existence of any such other criterion. This is a mild echo of E. T. Jaynes (1976), who said

> "It doesn't matter how many new words you drag into the discussion to avoid having to utter the word 'probability' in a sense different from frequency: likelihood, confidence, significance, propensity, support, credibility, acceptability, indiffidence, consonance, tenability, - and so on, until the resources of the good Dr. Roget are exhausted....It doesn't matter what approach you happen to like philosophically - by the time you have made your methods fully consistent, you will be forced, kicking and screaming, back to the ones given by

Laplace." (Author's note: Laplace argued for noninformative prior Bayesian analysis. We have, of course, allowed ourself proper subjective priors also, but the following of Assumption I is the most important part of Laplace's methods.)

D. Updating Γ.

The fourth reason sometimes proposed for broadening Bayesian analysis is the clear need to sometimes update the prior information by means other than Bayes theorem. This problem was discussed in subsection 2.4.

E. Conclusions.

A reading of the above suggests that the espoused robust Bayesian viewpoint was constructed by starting with pure Bayesian analysis and modifying it to handle every meaningful objection raised. This is exactly right. Assumption I is the cornerstone, and provides the starting point for the theory. At every stage where additional flexibility was needed, it was allowed into the theory, but in a way which minimized the resulting deviation from Assumption I. Any attempt to modify the theory, not satisfying this "minimum distance from Assumption I" constraint,is unlikely to prove successful.

6.4 Future Development

I agree with Dempster (1976) that

> "The ultimate goal of research on Bayesian robustness should be to classify applied situations so that a plausible prepackaged robustness analysis within each class will be available. I believe that only the faintest beginnings have been made on this task."

This would enable users to investigate robustness themselves, surely the most desirable goal.

Because it will probably always be the case that many (most?) users of statistics will not have the skill or the inclination to do such analyses, however, it behooves researchers to find specific robust Bayesian procedures or families of robust prior distributions (to use in place of conjugate families where warranted) for important situations. Again, relatively little has been done in this area.

Alerting users to situations *lacking* robustness is also very important. They can then know when, and on what, it is necessary to concentrate their prior elicitation.

As a concluding comment, note that a common criticism of Bayesian analysis is that it is too automatic. Thus Kiefer (1977) states that

> "...statistics is too complex to be codified in terms of a simple prescription that is a panacea for all settings..."

As we have seen, robust Bayesian analysis offers no single prescription, and instead urges flexibility in thought and methods. It demands only that the proper goal be kept in mind.

ACKNOWLEDGEMENTS

The list of people who have contributed to my views on this subject would be too large to include. During the writing of the paper I received useful ideas and had helpful discussions with Leon Gleser, Prem Goel, Bruce Hill, Ker-Chau Li, Prem Puri, Herman Rubin, Arif Zaman, and participants in a seminar on Foundations of Statistics held at Purdue University. My thanks to all these people. Herman Rubin deserves special mention as my source for clear thinking whenever I get confused.

A preliminary version of the paper was read by José Bernardo, Lawrence Brown, Morris DeGroot, Arthur Dempster, Persi Diaconis, Joseph Kadane, Isaac Levi, Dennis Lindley, and Melvin Novick, all of whom had many useful comments which greatly improved the paper. I am particularly grateful to Morris DeGroot for a very extensive set of insightful comments.

REFERENCES

Albert, J.H. (1981). Simultaneous estimation of Poisson means. J. Multivariate Anal. 11, 400-417.

Alvo, M. (1977). Bayesian sequential estimates. Ann. Statist. 5, 955-968.

Anscombe, F.J. (1963). Bayesian inference concerning many parameters with reference to supersaturated designs. Bull. Int. Statist. Inst. 40, 721-733.

Bakan, G.J. and Oleksenko, O.M. (1977). Nonlinear estimation by approximating the a posteriori density by a normal distribution. Soviet Automat. Control 10, 6-10.

Bansal, A.K. (1978). Robustness of a Bayes estimator for the mean of a normal population with nonnormal prior. Commun. Statist. - Theor. Meth. A7, 453-460.

Barnard, G. (1982). A coherent view of statistical inference. To be published in the proceedings of the statistical symposium held at Waterloo in 1981.

Basu, D. (1971). An essay on the logical foundations of survey sampling. In Foundations of Statistical Inference, V.P. Godombe and D.A. Sprott (eds.) Holt, Rinehart, and Winston, Toronto.

Basu, D. (1975). Statistical information and likelihood (with discussion). Sankhya A37, 1-71.

Basu, D. (1978). On the relevance of randomization in data analysis (with discussion). In Survey Sampling and Measurement, N.K. Namboodiri (ed.). Academic Press, New York.

Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test. J. Amer. Statist. Assoc. 75, 575-595.

Beran, R.J. (1970). Upper and lower risks and minimax procedures. Sixth Berkeley Symp. Math. Statist, Prob. 1, 1-16, University of California Press, Berkeley.

Beran, R.J. (1971). On distribution-free statistical inference with upper and lower probabilities. Ann. Math. Statist. 42, 157-168.

Berger, J. (1976a). Inadmissibility results for generalized Bayes estimators of coordinates of a location vector. Ann. Statist. 4, 302-333.

Berger, J. (1976b). Admissibility results for generalized Bayes estimators of coordinates of a location vector. Ann. Statist. 4, 334-356.

Berger, J. (1979). Multivariate estimation with nonsymmetric loss functions. In Optimizing Methods in Statistics, J.S. Rustagi (Ed.). Academic Press, New York.

Berger, J. (1980a). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. Ann. Statist. 8, 716-761.

Berger, J. (1980b). Statistical Decision Theory: Foundations, Concepts, and Methods. Springer-Verlag, New York.

Berger, J. (1982a). Selecting a minimax estimator of a multivariate normal mean. Ann. Statist. 10, 81-92.

Berger, J. (1982b). Bayesian robustness and the Stein effect. J. Amer. Statist. Assoc. 77, 358-368.

Berger, J. (1982c). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In Statistical Decision Theory and Related Topics III, S.S. Gupta and J. Berger (eds.). Academic Press, New York.

Berger, J. (1982d). Bayesian salesmanship. To appear in Bayesian Inference and Decision Techniques with Applications, P.K. Goel and A. Zellner (eds.). North-Holland, Amsterdam.

Berger, J., Berliner, L.M., and Zaman, A. (1982). General admissibility and inadmissibility results for estimation in a control problem. Ann. Statist. 10, 838-856.

Berger. J. and Wolpert, R. (1982a). Estimating the mean function of a Gaussian process and the Stein effect. To appear in J. Multivariate Analysis.

Berger, J. and Wolpert, R. (1982b). The Likelihood Principle: a review and generalizations. Technical Report #82-33, Department of Statistics, Purdue University, West Lafayette, Indiana.

Berger, R. (1979). Gamma minimax robustness of Bayes rules. Comm. Statist. 8, 543-560.

Berk, R. (1966). Limiting behavior of posterior distributions when the model is incorrect. Ann. Math. Statist. 37, 51-58.

Berk, R. (1970). Consistency a posteriori. Ann. Math. Statist. 41, 894-906.

Berliner, L.M. (1981). Improving on inadmissible estimators in the control problem. To appear in Ann. Statist..

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with Discussion). J. Roy. Statist. Soc. B 41, 113-147.

Barnardo, J.M. (1981). Reference decisions. Symposia Mathematica XXV, 85-94.

Bickel, P.J. (1979). Minimax estimation of the mean of a normal distribution subject to doing well at a point. Technical Report, Dept. of Statistics, Univ. of California at Berkeley.

Bickel, P. J. and Yahav, J. A. (1967). Asymptotically pointwise optimal procedures in sequential analysis. In Proc. 5th Berkeley Symp. Math. Statist. Prob. 1, 401-413, Univ. of California Press, Berkeley.

Bickel, P. J. and Yahav, J. A. (1969). Some contributions to the asymptotic theory of Bayes solutions. Z. Warsch. verw. Gebiete 11, 257-276.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). J. Amer. Statist. Assoc. 57, 269-326.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. H. (1975). Discrete Multivariate Analysis: Theory and Practice. M.I.T. Press, Cambridge.

Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing information. Ann. Math. Statist. 33, 882-886.

Blum, J. R. and Rosenblatt, J. (1967). On partial a priori information in statistical inference. Ann. Math. Statist. 38, 1671-1678.

Bock, M. E. (1982). Employing vague inequality prior information in estimation of a normal mean vector. In Statistical Decision Theory and Related Topics III, S. S. Gupta and J. Berger (eds.). Academic Press, New York.

Boole, G. (1854). An Investigation of the Laws of Thought: Reprinted by Dover (1958). (Chapters XVII and XVIII.)

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with Discussion). J. Roy. Statist. Soc. A 143, 383-430.

Box, G. E. P. and Tiao G. C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. Ann. Math. Statist. 42, 855-904.

Brown, L. D. (1979). An heuristic method for determining admissibility of estimators - with applications. Ann. Statist. 7, 960-994.

Brunk, H. D. and Pierce, D. A. (1977). Large sample posterior normality of the population mean. Commun. Statist. - Theor. Meth. A6, 1-14.

Burnasev, M. V. (1979). Asymptotic expansions of the integral risk of statistical estimators of location parameter in a scheme of independent observations. Soviet Math. Dokl. 20, 788-791.

Campbell, G. and Hollander, M. (1979). Nonparametric Bayes estimation with incomplete Dirichlet prior information. In Optimizing Methods in Statistics, J. S. Rustagi (ed.). Academic Press, New York.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1977). Foundations of Inference in Survey Sampling. Wiley, New York.

Chamberlain, G. and Leamer, E. E. (1976). Matrix weighted averages and posterior bounds. J. Roy. Statist. Soc. Ser. B. 38, 73-84.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Statist. 23, 493-507.

Chernoff, H. (1956). Large-sample theory: Parametric case. Ann. Math. Statist. 27, 1-22.

Chernoff, H. (1959). Sequential design of experiments. Ann. Math. Statist. 30, 755-770.

Chernoff, H. (1970). Sequential Analysis and Optimal Design. S.I.A.M.

Clevenson, M. and Zidek, J. (1975). Simultaneous estimation of the mean of independent Poisson laws. J. Amer. Statist. Assoc. 70, 698-705.

Davis, W.A. (1979). Approximate Bayesian predictive distributions and model selection. J. Amer. Statist. Assoc. 74, 312-317.

Dawid, A.P. (1979). On the limiting normality of posterior distributions. Proc. Camb. Phil. Soc. 67, 625-633.

Dawid, A. P. (1973). Posterior expectations for large observations. Biometrika 60, 664-666.

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. In Studies in Subjective Probability (1964), H. E. Kyburg and H. E. Smokler (eds.). Wiley, New York.

de Finetti, B. (1972). Probability, Induction, and Statistics. Wiley, New York.

de Finetti, B. (1974, 1975). Theory of Probability, Volumes 1 and 2, Wiley, New York.

De Groot, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill, New York.

Dempster, A. P. (1966). New methods of reasoning toward posterior distributions based on sample data. Ann. Math. Statist. 37, 355-374.

Dempster, A. P. (1967). Upper and lower probabilities induced by multi-valued maps. Ann. Math. Statist. 38, 325-339.

Dempster, A. P. (1968). A generalization of Bayesian inference. J. Roy. Statist. Soc. Ser. B. 30, 205-248.

Dempster, A. P. (1971). Model searching and estimation in the logic of inference. In Foundations of Statistical Inference, V. P. Godambe and D. A. Sprott (eds.). Holt, Rinehart and Winston, Toronto.

Dempster, A. P. (1975). A subjectivist look at robustness. Bull. of the International Statistical Institute 46, 349-374.

Dempster, A. P. (1976). Examples relevant to the robustness of applied inferences. In Statistical Decision Theory and Related Topics II, S. S. Gupta and D. S. Moore (eds.). Academic Press, New York.

De Robertis, L. and Hartigan, J. A. (1981). Bayesian inference using intervals of measures. Ann. Statist. 9, 235-244.

De Rouen, T. A. and Mitchell, T. J. (1974). A G_1-minimax estimator for a linear combination of binomial probabilities. J. Amer. Statist. Assoc. 69, 231-233.

Dey, D. (1980). On the choice of coordinates in simultaneous estimation of normal means. Technical Report 80-32, Dept. of Statistics, Purdue University.

Dey, D. and Berger, J. (1980). Combining coordinates in simultaneous estimation of normal means. To appear in J. Statist. Planning and Inference.

Diaconis, P. and Freedman, D. (1981). Frequency properties of Bayes rules. To appear in the Proceedings of the Conference on Scienctific Inference, Data Analysis, and Robustness, University of Wisconsin, Madison.

Diaconis, P. and Freedman, D. (1982). Bayes rules for location problems. In Statistical Decision Theory and Related Topics III, S.S. Gupta and J. Berger (eds.). Academic Press, New York.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. Ann. Statist. 7, 269-281.

Diaconis, P. and Zabell, S. (1982). Updating subjective probability. J. Amer. Statist. Assoc. 77, 822-830.

Dickey, J.M. (1973). Scientific reporting. J. Roy. Statist. Soc. B 35, 285-305.

Dickey, J.M. (1974). Bayesian alternatives to the F-test and least-squares estimator in the normal linear model. In Studies in Bayesian Econometrics and Statistics, S.E. Fienberg and A. Zellner (eds.). North Holland, Amsterdam.

Dickey, J.M. (1976a). Discussion of "Strong inconsistency from uniform priors" by M. Stone, J. Amer. Statist. Assoc. 71, 119-125.

Dickey, J.M. (1976b). Approximate posterior distributions. J. Amer. Statist. Assoc. 71, 680-689.

Dickey, J.M. and Freeman, P. (1975). Population - distributed personal probabilities. J. Amer. Statist. Assoc. 70, 362-364.

Doksum, K. (1970). Decision theory for some nonparametric models. In Proc. Sixth Berkeley Symp. Math. Statist. Prob. 1, 331-343. University of California Press, Berkeley.

Dubins, L.E. (1969). An elementary proof of Bochner's finitely additive Radon-Nikodym theorem. Amer. Math. Monthly 76, 520-523.

Dubins, L.E. (1975). Finitely additive conditional probabilities, conglomerability, and disintegrations. Ann. Probability 3, 89-99.

Edwards, W., Lindman, H., and Savage, L.J. (1963). Bayesian statistical inference for psychological research. Psychological Review 70, 193-242.

Efron, B. and Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika 65, 457-482.

Efron, B. and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators - Part I: the Bayes case. J. Amer. Statist. Assoc. 66, 807-815.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators - Part 2: the empirical Bayes case. J. Amer. Statist. Assoc. 66, 807-815.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. J. Amer. Statist. Assoc. 68, 117-130.

Fabius, J. (1964). Asymptotic behavior of Bayes estimates. Ann. Math. Statist. 35, 846-856.

Faith, R. (1978). Minimax Bayes estimators of a multivariate normal mean. J. Multivariate Anal. 8, 372-379.

Fine, T. (1973). Theories of Probability. Academic Press, New York.

Fishburn, P. C. (1965). Analysis of decisions with incomplete knowledge of probabilities. Operations Research 13, 217-237.

Fishburn, P. C., Murphy, A. H., and Isaacs, H. H. (1968). Sensitivity of decisions to probability estimation errors: a re-examination. Operations Research 16, 253-268.

Fortus, R. (1979). Approximations to Bayesian sequential tests of composite hypotheses. Ann. Statist. 7, 579-591.

Fraser, D.A.S. (1979). Inference and Linear Models. McGraw-Hill, New York.

Fraser, D.A.S. and Mackay, J. (1976). On the equivalence of standard inference procedures. In Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II , W.L. Harper and C.A. Hooker (eds.). D. Reidel, Boston.

Freedman, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. Ann. Math. Statist. 34, 1386-1403.

Freedman, D. (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. Ann. Math. Statist. 35, 454-456.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. J. Amer. Statist. Assoc. 74, 153-160.

Ghosh, J.K., Sinha, B.K., and Joshi, S.N. (1982). Expansions for posterior probability and integrated Bayes risk. In Statistical Decision Theory and Related Topics III, S.S. Gupta and J. Berger (eds.). Academic Press, New York.

Ghosh, M. and Parsian, A. (1981). Bayes minimax estimation of multiple Poisson parameters. J. Multivariate Anal. 11, 280-288.

Gleser, L.J. and Kunte, S. (1976). On asymptotically optimal sequential Bayes interval estimation procedures. Ann. Statist. 4, 685-711.

Godambe, V.P. (1982). Estimation in survey sampling: robustness and optimality. J. Amer. Statist. Assoc. 77, 393-406.

Godambe, V.P. and Thompson, M.E. (1971). The specification of prior knowledge by classes of prior distributions in survey sampling estimation. In Foundations of Statistical Inference, V.P. Godambe and D.A. Sprott (eds.). Holt, Rinehart, and Winston, Toronto.

Godambe, V.P. and Thompson, M.E. (1977). Robust near optimal estimation in survey practice. Bulletin of the International Statist. Inst. XLVII, 127-170.

Goldstein, M. (1974). Approximate Bayesian inference with incompletely specified prior distributions. Biometrika 61, 629-631.

Goldstein, M. (1979). The variance modified Bayes estimator. J. Roy. Statist. Soc. B 41, 96-100.

Goldstein, M. (1980). The linear Bayes regression estimator under weak prior assumptions. Biometrika 67, 621-628.

Good, I. J. (1950). Probability and the Weighing of Evidence. Griffin, London.

Good, I.J. (1952). Rational decisions. J. Roy. Statist. Soc. B 14, 107-114.

Good, I. J. (1962a). Subjective probability as the measure of a nonmeasurable set. In Logic, Methodology, and Philosophy of Science. Stanford.

Good, I.J. (1962b). How rational should a manager be? Management Science 8, 383-393.

Good, I.J. (1965). The Estimation of Probabilities: An Essay on Modern Bayesian Methods. M.I.T. Press, Cambridge.

Good, I.J. (1968). The utility of a distribution. Nature 219, 1392.

Good, I.J. (1973). The probabilistic explication of evidence, surprise, causality, explanation and utility. In Foundations of Statistical Inference, V.P. Godambe and D.A. Sprott (eds.). Holt, Rinehart, and Winston, Toronto.

Good, I.J. (1976). The Bayesian influence, or how to sweep subjectivism under the carpet. In Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II, W.L. Harper and C.A. Hooker (eds.). D. Reidel, Boston.

Good, I.J. (1980). Some history of the hierarchical Bayesian methodology. In Bayesian Statistics, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.). University Press, Valencia.

Good, I.J. (1981). The robustness of a hierarchical model for multinomials and contingency tables. To appear in the Proceedings of the Conference on Scientific Inference, Data Analysis, and Robustness, University of Wisconsin, Madison.

Good, I. J. and Crook, J. F. (1974). The Bayes/non Bayes compromise and the multinomial distribution. J. Amer. Statist. Assoc. 69, 711-720.

Gupta, S. S. and Hsiao, P. (1981). On Γ-minimax, minimax, and Bayes procedures for selecting populations close to a control. Sankhyā Ser. B.

Gupta, S. S. and Huang, D. Y. (1975). On Γ-minimax classification procedures. Proceedings of the 40th Session of the International Statistical Institute 46, Book 3, 330-335.

Gupta, S. S. and Huang, D. Y. (1977). On some Γ-minimax selection and multiple comparison procedures. In Statistical Decision Theory and Related Topics II, S. S. Gupta and D. S. Moore (eds.). Academic Press, New York.

Gupta, S. S. and Kim, W. C. (1980). Γ-minimax and minimax rules for comparison of treatments with a control. Recent Developments in Statistical Inference and Data Analysis, K. Matusita (ed.). North Holland, Amsterdam.

Hájek, J. (1981). Sampling from a Finite Population. Marcel Dekker, New York.

Hartigan, J. A. (1969). Linear Bayesian methods. J. Roy. Statist. Soc. B 31, 446-454.

Heath, D.C. and Sudderth, W.D. (1978). On finitively additive priors, coherence, and extended admissibility. Ann. Statist. 6, 333-345.

Heyde, C.C. and Johnstone, I.M. (1979). On asymptotic posterior normality for stochastic processes. J. Roy. Statist. Soc. B 41, 184-189.

Hildreth, C. (1963). Bayesian statisticians and remote clients. Econometrica 31, 422-438.

Hill, B. (1965). Inference about variance components in the one-way model. J. Amer. Statist. Assoc. 60, 806-825.

Hill, B. (1969). Foundations for the theory of least squares. J. Roy. Statist. Soc. B 31, 89-97.

Hill, B. (1970). Some contrasts between Bayesian and classical inference in the analysis of variance and in the testing of models. In Bayesian Statistics, D. L. Meyer and R. O. Collier, Jr. (eds.). Peacock Publ., Itasca, Ill.

Hill, B. (1974). On coherence, inadmissibility, and inference about many parameters in the theory of least squares. In Studies in Bayesian Econometrics and Statistics, S. Fienberg and A. Zellner (eds.). North Holland, Amsterdam.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. Ann. Statist. 3, 1163-1174.

Hill, B. (1977). Exact and approximate Bayesian solutions for inference about variance components and multivariate inadmissibility. In New Developments in the Applications of Bayesian Methods, A. Aykac and C. Brumat (eds.), 129-152. North Holland, Amsterdam.

Hill, B. (1980a). Robust analysis of the random model and weighted least squares regression. In Evaluation of Econometric Models. Academic Press, New York.

Hill, B. (1980b). On some statistical paradoxes and non-conglomerability. Bayesian Statistics, J. M. Bernardo, M. H. De Groot, D. V. Lindley, and A. F. M. Smith (eds.). University Press, Valencia.

Hill, B. (1980c). Invariance and robustness of the posterior distribution of characteristics of a finite population, with reference to contingency tables and the sampling of species. In Bayesian Analysis in Econometrics and Statistics, Essays in Honor of Harold Jeffreys, A. Zellner (ed.). North-Holland, Amsterdam.

Hinkley, D.V. (1982). Can frequentist inference be very wrong? A conditional 'Yes'. Technical Report No. 397, Department of Theoretical Statistics, University of Minnesota.

Hodges, J.L. and Lehmann, E.L. (1952). The use of previous experience in reaching statistical decisions. Ann. Math. Statist. 23, 396-407.

Hsiao, P. (1982). Γ-minimax procedures for selecting good location parameters in some multivariate distributions. In Statistical Decision Theory and Related Topics III, S.S. Gupta and J. Berger (eds.). Academic Press, New York.

Huber, P.J. (1972). Robust statistics: a review. Ann. Math. Statist. 43, 1041-1067.

Huber, P.J. (1973). The use of Choquet capacities in statistics. Bulletin of the International Statist. Inst. 45, 181-191.

Huber, P.J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. Ann. Statist. 1, 251-263.

Hudson, H.M. and Tsui, K. (1981). Simultaneous Poisson estimators for apriori hypotheses about means. J. Amer. Statist. Assoc. 76, 182-187.

Hwang, J.T. (1982a). Semi tail upper bounds on the class of admissible estimators in discrete exponential families with applications to Poisson and negative binomial distributions. Ann. Statist. 10, 1137-1147.

Hwang, J.T. (1982b). Certain bounds on the class of admissible estimators in continuous exponential families. Statistical Decision Theory and Related Topics III, S.S. Gupta and J. Berger (eds.). Academic Press, New York.

Isaacs, H.H. (1963). Sensitivity of decisions to probability estimation errors. Operations Research 11, 536-552.

Jackson, D.A., Donovan, T.M., Zimmer, W.J., and Deely, J.J. (1970). Γ_2-minimax estimators in the exponential family. Biometrika 57, 439-443.

Jackson, P., Novick, M., and DeKeyrel, D. (1980). Adversary preposterior analysis for simple parametric models. In Bayesian Analysis in Economics and Statistics, A. Zellner (eds.). North-Holland, Amsterdam.

Jaynes, E.T. (1968). Prior probabilities. IEEE Transactions on Systems Science and Cybernetics SSC-4, 227-241.

Jaynes, E.T. (1976). Confidence intervals versus Bayesian intervals. In Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II, W.L. Harper and C.A. Hooker (eds.). D. Reidel, Boston.

Jaynes, E.T. (1981). The intuitive inadequacy of classical statistics. Presented at the International Convention on Fundamentals of Probability and Statistics, Luino, Italy.

Jeffrey, R. (1968). Probable knowledge. In The Problem of Inductive Logic, I. Lakatos (ed.). North Holland, Amsterdam

Jeffreys, H. (1961). Theory of Probability, 3rd Edition. Oxford University Press, Oxford.

Johnson, B. R. and Truax, D. R. (1978). Asymptotic behavior of Bayes procedures for testing simple hypotheses in multiparameter exponential families. Ann. Statist. 6, 346-361.

Johnson, R. A. (1967). An asymptotic expansion for posterior distributions. Ann. Math. Statist. 38, 1899-1906.

Johnson, R. A. (1970). Asymptotic expansions associated with posterior distributions. Ann. Math. Statist. 41, 851-864.

Kadane, J. B. and Chuang, D. T. (1978). Stable decision problems. Ann. Statist. 6, 1095-1110.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C. (1980). Interactive elicitation of opinion for a normal linear model. J. Amer. Statist. Assoc. 75, 845-854.

Kadane, J.B., Schervish, M., and Seidenfeld, T. (1981). Statistical implications of finitely additive probability. Technical Report #206, Carnegie-Mellon University, Pittsburgh.

Kiefer, J. (1977). The foundations of statistics - are there any? Synthese 36, 161-176.

Kiefer, J. and Sacks, J. (1963). Asymptotically optimum sequential inference and design. Ann. Math. Statist. 34, 705-750.

Kleyle, R. (1975). Upper and lower probabilities for discrete distributions. Ann. Statist. 3, 504-511.

Koopman, B. O. (1940). The axioms and algebra of intuitive probability. Ann. Math. 41, 269-278.

Kraft, C. H., Pratt, J.W., and Seidenberg, A.(1959). Intuitive probability on finite sets. Ann. Math. Statist. 30, 408-419.

Kudō, H. (1967). On partial prior information and the property of parametric sufficiency. Proc. Fifth Berkeley Sym. Prob. Statist. 1. University of California Press, Berkeley.

Kyburg, H. (1974). The Logical Foundations of Statistical Inference. D. Reidel, Dordrecht.

Kyburg, H. E. (1976). Statistical knowledge and statistical inference. In Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II, W. L. Harper and C. A. Hooker (eds.). D. Reidel, Boston.

Lambert, D. and Duncan, G. (1981). Bayesian learning based on partial prior information. Technical Report No. 209, Department of Statistics, Carnegie-Mellon University.

Leamer, E. E. (1978). Specification Searches. Wiley, New York.

Le Cam, L. (1953). On some asymptotic properties of the maximum likelihood estimates and related Bayes estimates. University of California Pub. Statist. 1, 277-330.

Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. Proc. 3rd Berkeley Symp. Math. Statist. Probability 1, 129-156. University of California Press, Berkeley.

Le Cam, L. (1982). On the risk of Bayes estimates. In Statistical Decision Theory and Related Topics III, S. S. Gupta and J. Berger (eds.). Academic Press, New York.

Leonard, T. (1976). Some alternative approaches to multiparameter estimation. Biometrika 63, 69-76.

Levi, I. (1974). On indeterminate probabilities. J. of Philosophy LXXI.

Levi, I. (1980). The Enterprise of Knowledge. MIT Press, Cambridge.

Lindley, D.V. (1960). The use of prior probability distributions in statistical inference and decisions. Proc. Berkeley Symp. Math. Statist. Probability 1, 453-468. University of California Press, Berkeley.

Lindley, D.V. (1968). The choice of variables in multiple regression (with discussion). J. Roy. Statist. Soc. B 30, 31-66.

Lindley, D.V. (1982). Scoring rules and the inevitability of probability. International Statistical Review 50, 1-26.

Lindley, D.V. and Novick, M. (1981). The role of exchangeability in inference. Ann. Statist. 9, 45-58.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. J. Roy. Statist. Soc. B 34, 1-41.

Manski, C.F. (1981). Learning and decision making when subjective probabilities have subjective domains. Ann. Statist. 9, 59-65.

Marazzi, A. (1980). Robust Bayesian estimation for the linear model. Research Report No. 27, Fachgruppe fuer Statistik, Eidgenoessische Technische Hochschule, Zurich.

Maritz, J.S. (1970). Empirical Bayes methods. Methuen, London.

Masreliez, C.J. and Martin, R.D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. IEEE Transactions on Automatic Control AC-22, 361-371.

Meeden, G. and Isaacson, D. (1977). Approximate behavior of the posterior distribution for a large observation. Ann. Statist. 5, 899-908.

Menges, G. (1966). On the Bayesification of the minimax principle. Unternehmensforschung 10, 81-91.

Miescke, K.J. (1981). Γ-minimax selection procedures in simultaneous testing problems. Ann. Statist. 9, 215-220.

Morris, C. (1977). Interval estimation for empirical Bayes generalizations of Stein's estimator. The Rand Paper Series, Rand Corp., Santa Monica.

Morris, C. (1981). Parametric empirical Bayes confidence intervals. To appear in the Proceedings of the Conference on Scientific Inference, Data Analysis, and Robustness, University of Wisconsin, Madison.

Morris, C. (1982a). Natural exponential families with quadratic variance functions. Ann. Statist. 10, 65-80.

Morris, C. (1982b). Parametric empirical Bayes inference: theory and applications. To appear in J. Amer. Statist. Assoc.

Novick, M.R. (1969). Multiparameter Bayesian indifference procedures (with discussion). J. Roy. Statist. Soc. B 31, 29-64.

Pierce, D.A. and Folks, J.L. (1969). Sensitivity of Bayes procedures to the prior distribution. Operations Research 17, 344-350.

Polasek, W. (1983). Multivariate Regression Systems: Estimation and Sensitivity Analysis of Two-Dimensional Data. (This Volume).

Potter, J.M. and Anderson, B.D.O. (1980). Prior information and decision making. IEEE Trans. Syst., Man., Cybern. 10, 125-133.

Pratt, J.W. (1965). Bayesian interpretation of standard inference statements (with discussion). J. Roy. Statist. Soc. B 27, 169-203.

Pratt, J.W. Raiffa, H. and Schlaifer, R. (1965). Introduction to Statistical Decision Theory. McGraw-Hill, New York.

Raiffa, H. and Schlaifer, R. (1961). Applied Statistical Decision Theory. Harvard University, Boston.

Ramsay, J.O. and Novick, M.R. (1980). PLU robust Bayesian decision theory: point estimation. J. Amer. Statist. Assoc. 75, 901-907.

Randles, H.R. and Hollander, M. (1971). Γ-minimax selection procedures in treatment versus control problems. Ann. Math. Statist. 42, 330-341.

Rios, S. and Girón, F.J. (1980). ıQuasi-Bayesian behavior: a more realistic approach to decision making? Bayesian Statistics, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (eds.). University Press, Valencia.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. Proc. Second Berkeley Symp. Math. Statist. Prob. 1, 131-148. Univ. of California Press, Berkeley.

Robbins, H. (1955). An empirical Bayes approach to statistics. Proc. Third Berkeley Symposium Math. Statist. Prob. 1, 157-164. University of California Press, Berkeley.

Robbins, H.E. (1964). The empirical Bayes approach to statistical decision problems. Ann. Math. Statist. 35, 1-20.

Rosenkrantz, R.D. (1977). Inference, Method, and Decision: Towards a Bayesian Philosophy of Science. Reidel, Boston.

Royall, R.M. and Pfeffermann, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. Biometrika 69, 401-409.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. Ann. Statist. 6, 34-58.

Rubin, H. (1971). A decision-theoretic approach to the problem of testing a null hypothesis. In Statistical Decision Theory and Related Topics, S. S. Gupta and J. Yackel (eds.). Academic Press, New York.

Rubin, H. (1972). On large sample properties of certain nonparametric procedures. Proc. Sixth Berkeley Symp. Math. Statistics and Prob. 429-435. University of California Press, Berkeley.

Rubin, H. (1977). Robust Bayesian estimation. In Statistical Decision Theory and Related Topics II, S. S. Gupta and D. Moore (eds.). Academic Press, New York.

Rubin, H. and Sethuraman, J. (1965). Bayes risk efficiency. Sankhyā A 27, 347-356.

Savage, L. J. (1954). The Foundations of Statistics. Wiley, New York.

Savage, L. J. (1961). The foundations of statistics reconsidered. Proc. Fourth Berkeley Symp. Math. Statistics and Prob. 575-586. University of California Press, Berkeley.

Savage, L. J. (1962). Bayesian statistics. In Recent Developments in Information and Decision Processes, R. E. Machol and P. Gray (eds.). Macmillan, New York.

Savage, L. J. (et. al.) (1962). The Foundations of Statistical Inference. Methuen, London.

Schneeweiss, H. (1964). Eine Entscheidungsregel fur den Fall partiell bekannter Wahrscheinlichkeiten. Unternehmensforschung. 8 No. 2, 86-95.

Schwartz, L. (1965). On Bayes procedures. Z. Wahrsch. Verw. Gebiete 4, 10-26.

Schwarz, G. (1962). Asymptotic shapes of Bayes sequential testing regions. Ann. Math. Statist. 33, 224-236.

Schwarz, G. (1968). Asymptotic shapes for sequential testing of truncation parameters. Ann. Math. Statist. 39, 2038-2043.

Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton University Press, Princeton.

Shafer, G. (1979). Two theories of probability. In PSA 1978, Vol. 2, Philosphy of Science Association, East Lansing, Michigan.

Shafer, G. (1981a). Jeffrey's rule of conditioning. Philosophy of Science, 48, 337-362.

Shafer, G. (1981b). Constructive probability. Synthese 48, 1-60.

Shafer, G. (1982). Belief functions and parametric models (with Discussion). J. Roy. Statist. Soc. B 44.

Shapiro, S. H. (1972). A compromise between the Bayes and minimax approaches to estimation. Technical Report No. 31, Department of Statistics, Stanford University.

Shapiro, S. H. (1975). Estimation of location and scale parameters - a compromise. Commun. Statist. 4(12), 1093-1108.

Skibinski, M. and Cote, L. (1963). On the inadmissibility of some standard estimates in the presence of prior information. Ann. Statist. 34, 539-548.

Smith, C. A. B. (1961). Consistency in statistical inference and decision. J. Roy. Statist. Soc. B 23, 1-25.

Smith, G. and Campbell, F. (1980). A critique of some ridge regression methods (with discussion). J. Amer. Statist. Assoc. 75, 74-103.

Solomon, D. L. (1972a). Λ-minimax estimation of a multivariate location parameter. J. Amer. Statist. Assoc. 67, 641-646.

Solomon, D. L. (1972b). Λ-minimax estimation of a scale parameter. J. Amer. Statist. Assoc. 67, 647-649.

Spielman, S. (1977). Physical probability and Bayesian statistics. Synthese 36, 235-269.

Srinivasan, C. (1980). Admissible generalized Bayes estimators and exterior boundary value problem. Sankhyā.

Stein, C. (1965). Approximation of improper prior measures by prior probability measures. In Bernouilli-Bayes-Laplace Festchr., 217-240. Springer-Verlag, New York.

Stein, C. (1981a). Estimation of the mean of a multivariate normal distribution. Ann. Statist. 9, 1135-1151.

Stein, C. (1981b). On the coverage probability of confidence sets based on a prior distribution. Technical Report No. 180, Dept. of Statistics, Stanford University.

Stone, M. (1963). Robustness of nonideal decision procedures. J. Amer. Statist. Assoc. 58, 480-486.

Stasser, H. (1981). Consistency of maximum likelihood and Bayes estimates. Ann. Statist. 9, 1107-1113.

Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. Ann. Math. Statist. 42, 385-388.

Strawderman, W. and Cohen, A. (1971). Admissibility of estimators of the mean vector of a multivariate normal distribution with quadratic loss. Ann. Math. Statist. 42, 270-296.

Suppes, P. (1975). Approximate probability and expectation of gambles. Erkenntnis 9, 153-161.

Suppes, P. and Zanotti, M. (1977). On using random relations to generate upper and lower probabilities. Synthese 36, 427-440.

Teller, P. (1976). Conditionalization, observation, and change of preference. In Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, W. Harper and C.A. Hooker (eds.). Reidel, Dordrecht.

Theil, H. (1963). On the use of incomplete prior information in regression analysis. J. Amer. Statist. Assoc. 58, 401-414.

Tiao, G.C. and Zellner, A. (1964). On the Bayesian estimation of multivariate regression. J. Roy. Statist. Soc. B 26, 277-285.

Umbach, D. (1978). On the approximate behavior of the posterior distribution for an extreme multivariate observation. J. Multivariate Anal. 8, 518-531.

Vardi, Y. (1979a). Asymptotic optimality of certain sequential estimators. Ann. Statist. 7, 1034-1039.

Vardi, Y. (1979b). Asymptotic optimal sequential estimation. Ann. Statist. 7, 1040-1051.

Walker, A.M. (1969). On the asymptotic behavior of posterior distributions. J. Roy. Statist. Soc. Ser. B. 31, 80-88.

Watson, S.R. (1974). On Bayesian inference with incompletely specified prior distributions. Biometrika, 61, 193-196.

Weerhandi, S. and Zidek, J.V. (1981). Multi-Bayesian statistical decision theory. J. Roy. Statist. Soc. A 144, 85-93.

Welch, B.L. and Peers, H.W. (1963). On formulas for confidence points based on integrals of weighted likelihoods. J. Roy. Statist. Soc. B 25, 318-329.

West, M. (1981). Robust sequential approximate Bayesian estimation. J. Royal Statist. Soc. B 43, 157-166.

West, S. (1979). Upper and lower probability inferences for the logistic function. Ann. Statist. 7, 400-413.

Wilkinson, G.N. (1977). On resolving the controversy in statistical inference (with Discussion). J. Roy. Statist. Soc. B 39, 119-171.

Williams, P.M. (1976). Indeterminate probabilities. In Formal Methods in the Methodology of Empirical Sciences, M. Przelecki, K. Szaniawski, and R. Wójciki (eds.). D. Reidel, Dordrecht.

Wolfenson, M. and Fine, T. (1982). Bayes-like decision making with upper and lower probabilities. J. Amer. Statist. Assoc. 77, 80-88.

Wolpert, R. and Berger, J. (1982). Incorporating prior information in minimax estimation of the mean of a Gaussian process. Statistical Decision Theory and Related Topics III, S.S. Gupta and J. Berger (eds.). Academic Press, New York.

Woodroofe, M. (1980). On the Bayes risk incurred by using asymptotic shapes. Commun. Statist.-Theor. Meth. A9, 1727-1748.

Zaman, Asad (1982). Quasitransitive preferences over lotteries. Technical Report, University of Pennsylvania.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. Wiley, New York.

Zellner, A. (1976). Bayesian analysis of the regression model with multivariate Student-t error terms. J. Amer. Statist. Assoc. 71, 400-405.

Zellner, A. (1982). Applications of Bayesian analysis in econometrics. Technical Report, H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago, Chicago.

Zellner, A. and Geisel, M. (1968). Sensitivity to control of uncertainty and form of the criterion function. In the Future of Statistics, D.G. Watts (ed.). Academic Press, New York.

Zheng, Z. (1982). A class of generalized Bayes minimax estimators. In Statistical Decision Theory and Related Topics III, S.S. Gupta and J. Berger (eds.). Academic Press, New York.

Zidek, J.V. (1982). Aspects of multi-Bayesian theory. Technical Report No. 82-11, Dept. of Mathematics, University of British Columbia, Vancouver.

82-9

PART II

THE ROBUST BAYESIAN VIEWPOINT*

James O. Berger
Purdue University

*Research supported by the National Science Foundation under Grant No. MCS-8101670A1.

Robustness of Bayesian Analyses, edited by Joseph B. Kadane
© Elsevier Science Publishers B.V. (1984) 63–124

1. Introduction

1.1. Introduction

Statistics needs a 'foundation', by which I mean a framework of analysis within which *any* statistical investigation can *theoretically* be planned, performed, and meaningfully evaluated. The words 'any' and 'theoretically' are key, in that the framework should apply to any situation but may only theoretically be implementable. Practical difficulties or time limitations may prevent complete (or even partial) utilization of such a framework, but the direction in which 'truth' could be found would at least be known.

To a large number of statisticians the above goal is deemed unattainable, with the attendant attitude being that one must 'keep an open mind' and use 'whatever works well for a given problem'. Besides seeming unnecessarily pessimistic and somewhat unscientific, such a position seems almost meaningless in that without the desired foundational framework there would be no way of determining what works well in a given problem.

The main contender for the crown is Bayesian analysis. ('Classical' statisticians tend to be of the 'there is no foundation' ilk.) The main justification for Bayesian analysis is a belief (for a variety of reasons) in

Assumption I. In any statistical investigation, one will ultimately be faced with making reports, inferences, or decisions which involve uncertainties. Of interest is the information available about these uncertainties after seeing the data, and the only trustworthy and sensible measures of this information are Bayesian posterior measures.

Belief in Assumption I leads many Bayesians to argue that the desired foundation is simply the usual Bayesian analysis in which one specifies a prior distribution for the unknowns and processes the data via Bayes rule.

This attitude is vigorously opposed by non-Bayesians, partly because of objections to Assumption I but, more often, because of a belief in

Assumption II. Prior distributions can never be quantified or elicited exactly (i.e. without error), especially in a finite amount of time.

Because of this distrust of prior distributions, many statisticians reject the Bayesian viewpoint out of hand.

A Bayesian viewpoint has long existed, however, which is based on belief in both Assumptions I and II. While Assumption I calls for a basically Bayesian outlook, Assumption II precludes the obvious Bayesian solution of writing down a single prior distribution and doing a Bayesian analysis. Instead, the viewpoint is essentially that one should strive for Bayesian behavior which is satisfactory for all prior distributions which remain plausible after the prior elicitation process has been terminated. I will call this the *robust Bayesian* viewpoint, and argue that it provides the desired foundational framework.

The robust Bayesian viewpoint is by no means new, of course, and virtually all Bayesians will ascribe to it to some degree. For example, de Finetti (as quoted by Dempster (1975)) stated

> "Subjectivists should feel obligated to recognize that any opinion (so much more the initial one) is only vaguely acceptable... So it is important not only to know the exact answer for an exactly specified initial problem, but what happens changing in a reasonable neighborhood the assumed initial opinion."

Most of the arguments and examples presented herein have undoubtedly been presented elsewhere. For instance, a very large proportion of the ideas can be found in the works of I.J. Good, even as early as Good (1950). (Indeed, I would have very few qualms about calling myself a Doogian.) Herman Rubin and Bruce Hill (among others) have also always espoused similar views. In some sense, therefore, this should be thought of as basically a review paper, with the goal of tying together the various elements of the robust Bayesian viewpoint in an attempt to present a convincing case. (To keep the account readable, I will defer most historical references to Section 5.)

This article is written more for the 'non-robust' Bayesian, than for the non-Bayesian. In other words, little attempt will be made to justify Assumption I. Besides the sheer impossibility of adequately discussing Assumption I in a single paper, the rationale is that the Bayesian should have clean hands before he accuses someone else's hands of being dirty. Presenting the (enormously convincing) arguments for Assumption I seems to have little effect on non-Bayesians if they are able to come back with the complaint that Assumption II seems totally obvious to them and they refuse to operate in violation of it. Fully admitting (and even expounding on) the truth of Assumption II, while showing how Assumption I can still basically be followed, should greatly enhance the Bayesian argument. (See Berger (1982d).)

In reading the paper, keep in mind that the robust Bayesian viewpoint is being advocated as the framework for ultimately verifying the sensibility of an analysis, and is not necessarily being advocated as an applied methodology to do all of statistics. Comparatively little work has been done on robust Bayesian methods, so while it is a very illuminating viewpoint from which to understand things, it is not to be expected to provide easy answers to all our problems.

As a final caveat, although I will talk about various 'classes' of Bayesians and non-Bayesians (such as 'objective' Bayesians, 'pure' subjective Bayesians, frequentists, etc.), these classes are to a large extent imaginary; most statisticians are a composite of a number of such classes. These distinctions will be made only for convenience in representing certain basic viewpoints.

In Section 2, justifications for Assumptions I and II will briefly be outlined. Since II implies that one must consider classes of plausible prior distributions, reasonable such classes will also be discussed, along with the problems of updating prior information. Section 3 is concerned with methods of measuring Bayesian robustness, and the somewhat surprising conclusion is reached that frequentist measures can be useful in measuring robustness. (This seems to conflict with Assumption I, and indeed behavior violating Assumption I can occur under this viewpoint, but *only* to the extent necessary to achieve robustness.) Section 4 deals with certain consequences of adopting this viewpoint, showing how certain features of many non-Bayesian techniques can be partially justified from the robust Bayesian viewpoint. Section 4 also presents an example, involving the Stein effect, which demonstrates that naive Bayesian intuition is not always trustworthy in the face of robustness considerations. Section 5 gives a brief survey of existing work related to Bayesian robustness, and contains some useful guidelines for achieving robustness. Section 6 consists of some conclusions and philosophical meanderings concerning the robust Bayesian viewpoint and objections to it.

1.2. Notation

In this paper it will be assumed that the data x is a realization of a random variable X with distribution $P_\theta(\cdot)$ on the sample space $\mathscr{X}$ for some unknown $\theta \in \Theta$. Although Θ will be referred to as the parameter space and $\{P_\theta \colon \theta \in \Theta\}$ will usually be a parametric family in the examples, the basic arguments hold for any index set Θ; thus the nonparametric situation would be included by letting Θ index any desired class of probability distributions. A prior distribution on Θ will be denoted π, $\pi(\cdot \mid x)$ will denote the posterior distribution of θ given the observation x, and $m(\cdot) = \mathrm{E}^\pi[P_\theta(\cdot)]$ will denote the marginal (or unconditional or predictive) distribution of X. (E will stand for expectation, with superscripts indicating what the expectation is being taken over, and subscripts indicating fixed parameter values.)

1.3. Decision theory

Many of the examples discussed will be presented from a decision theoretical viewpoint. The reason is mainly that, if a point is to be made, it can most clearly be done in a precisely quantifiable situation. It can, of course, be argued that, just as the robust Bayesian viewpoint seems necessary for understanding, so the robust decision theoretic viewpoint is also essential. ('Inference' problems would simply be problems where very little knowledge concerning the loss function was obtainable, and hence where robustness over a wide class of loss functions would be sought.) I certainly support this view, feeling that there are great dangers in refusing to at least think in decision theoretic terms.

(Incidentally, it has always struck me as curious that there are violently antidecision-theoretic Bayesians and violently anti-Bayesian decision theorists. Is there really such a big difference between the two types of subjective inputs?) To keep the paper contained, however, the decision theoretic issue will not be explicitly considered; issues get clouded when too much is attempted.

When employing a decision theoretic viewpoint, the action space will be denoted $\mathscr{A}$, the loss in taking action $a \in \mathscr{A}$ when $\theta \in \Theta$ obtains will be denoted $L(\theta, a)$, and the posterior expected loss of action a with respect to the prior π and observation x will be denoted

$$\rho(\pi, x, a) = \mathrm{E}^{\pi(\cdot|x)} L(\theta, a) = \int_{\Theta} L(\theta, a)\pi(\mathrm{d}\theta(\mathrm{d}\theta \mid x)\,. \tag{1.1}$$

A decision rule (for simplicity assumed to be a nonrandomized function from $\mathscr{X}$ into $\mathscr{A}$) will be denoted $\delta(x)$. We will have cause to consider the (frequentist) risk function

$$R(\theta, \delta) = \mathrm{E}_{\theta} L(\theta, \delta(X)) = \int_{\mathscr{X}} L(\theta, \delta(x)) P_{\theta}(\mathrm{d}x)$$

and the Bayes risk

$$\begin{aligned} r(\pi, \delta) &= \mathrm{E}^{\pi} R(\theta, \delta) = \int_{\Theta} R(\theta, \delta)\pi(\mathrm{d}\theta) \\ &= \mathrm{E}^{m} \rho(\pi, X, \delta(X)) = \int_{\mathscr{X}} \rho(\pi, x, \delta(x)) m(\mathrm{d}x) \end{aligned} \tag{1.2}$$

2. The robust Bayesian viewpoint

As the robust Bayesian viewpoint is founded on a belief in Assumptions I and II, these assumptions will be discussed in the first two subsections. Subsection 2.3 discusses reasonable classes of prior distributions which could be considered in light of Assumption II. Subsection 2.4 discusses the issue of updating uncertain prior information.

2.1. Justification for Assumption I

There are at least seven basic reasons that have been advanced for being a Bayesian, these being:

(i) Prior information is too important to ignore or deal with in an ad hoc fashion.

(ii) According to most 'classical' criteria, the class of 'optimal' procedures corresponds to the class of Bayes procedures, so one should select from among this class according to prior information.

(iii) The Bayesian viewpoint works better than any other in revealing the common sense features of a situation and producing reasonable procedures.

(iv) The goal of statistics is to communicate evidence about uncertainties, and the correct language of uncertainty is probability. Only subjective probability provides a broad enough framework to encompass the types of uncertainties encountered, and Bayes theorem tells how to process information in the language of subjective probability.

(v) Axioms of rational behavior imply that any 'coherent' mode of behavior corresponds to Bayesian behavior with respect to some prior distribution.

(vi) The Likelihood Principle seems irrefutable, yet the only general way of implementing it seems to be through Bayesian analysis.

(vii) Bayesian posterior measures of accuracy seem to be the only meaningful measures of accuracy.

Many papers and books have been written about these reasons, and no attempt will be made to review or explain these reasons in detail. A few comments seem in order concerning the importance and effectiveness of each of these reasons, however.

Reasons (i), (ii), and (iii) do not bear directly on Assumption I, but do lend considerable support to the Bayesian position. Reason (i) is important, especially when it is realized that choice of such things as a model is really just a (perhaps rather extreme) use of prior information. Nevertheless, reason (i) is not very effective for 'conversion' since it can always be argued (incorrectly or not) that in many problems no (or very little) prior information is available.

Reason (ii) is very suggestive, pointing out a frequently occurring one-to-one correspondence between 'good' classical procedures or methods and Bayesian procedures. In testing between two simple hypotheses, for example (the 'dichotomy' discussed from this perspective by Lindley and Savage in, for instance, Savage et al. (1962)), the classical Neyman–Pearson tests are the Bayes tests. In selecting a test, therefore, one can either make a grand intellectual leap to α and β, or can carefully consider the available prior information (and information about the loss or consequences of accepting and rejecting and cost of experimentation) and select the test on Bayesian (decision-theoretic) grounds. To me the essence of reasoning is to reduce a complicated problem to simple components, analyze the components separately, and recombine to get an answer. I distrust grand intellectual leaps.

Another example, involving current research, is the work on finding alternatives to the least squares estimator, due either to pursuance of the Stein phenomenon (that in three or more dimensions the usual estimator is inadmissible) or ridge regression ideas. Again there is a one-to-one correspondence between 'good' procedures (say, as measured by mean squared error) and Bayesian procedures (as shown in the normal case by Brown (1971)). One can thus select an alternative to the usual estimator either by a (mystical to me) intuitive method, or by considering which θ are a priori most likely to occur and selecting a Bayes estimator designed to do well for these θ (while preserving mean squared error dominance if desired). Further discussion of this example is given in subsection 4.5.

Reason (iii) is certainly not very good for conversion, but is the reason Bayesians tend to become more and more Bayesian as time progresses. Application of Bayesian reasoning will time and again clear up mystifying situations, and easily arrived at Bayesian procedures (say, with respect to noninformative prior distributions) often perform much better than complicated and difficult to determine classical procedures. (It is a shame that the very simplicity of much of Bayesian analysis is considered an indictment of it; I may find very stimulating a difficult mathematical derivation of, say, a minimax rule, and not be so intellectually excited at the routine calculation of the corresponding noninformative prior Bayes rule, yet (if done sensibly) the latter rule will virtually always be better.)

The last four reasons all pertain to the validity of Assumption I, and indeed are very related. They correspond to essentially four different modes of argument for Assumption I, however, and hence are listed separately.

Reason (iv) has been eloquently argued by many scientists, philosophers, probabilists, and statisticians (cf. Jeffreys (1961), de Finetti (1972, 1974, 1975), Good (1950), Jaynes (1981), and Lindley (1982)). We often work very hard in elementary statistics courses to suppress in students their natural instincts to talk about the 'chance that θ is in the interval' or the 'probability that the hypothesis is true', telling them that (although these are what they really want to know) we must be 'objective' and create an artificial language of confidence statements and error probabilities. Such artificial languages do not seem able to withstand deep scrutiny.

Reason (v) is compelling to many, but is perhaps a touch overemphasized.

The axioms of rationality are, for the most part, very believable, and it is interesting to know that any coherent method of behavior corresponds to Bayesian behavior with respect to some prior distribution, but this does not say that the right way to behave is to write down a prior distribution and perform a Bayesian analysis. Indeed I would term this latter behavior incoherent (in a broad sense), in that the prior distribution used can only be an approximation to true prior beliefs (see the next subsection). The value of rationality and coherence is that they indicate that my 'optimal' analysis will correspond to a Bayesian analysis with respect to my 'true' prior distribution (admittedly circularly defined here), and hence indicate the direction in which I should look to determine my optimal analysis. See Section 3.3 and Berger (1982d) for further discussion and references.

Reasons (vi) and (vii) are often the most convincing to non-Bayesians. They bring out the key point that many Bayesians became Bayesians, not because they were infatuated with prior information, but because they could see no other meaningful solution to the conditional inference problems besetting classical statistics.

The Likelihood Principle is wonderful, in that so much follows from so little. The Likelihood Principle essentially says that if the family of distributions $\{P_\theta\}$ has densities $\{p_\theta\}$ with respect to some dominating measure and the observation from the experiment is x, then the evidence about θ obtainable from the experiment is contained in the likelihood function $l_x(\theta) = p_\theta(x)$ (considered as a function of θ). The appeal of the principle is partly the fact that (as shown by Birnbaum (1962)) it follows from the Principles of Sufficiency and Conditionality; indeed all that is needed of the Conditionality Principle is that if one chooses between two experiments based on the flip of an (independent) fair coin, then the evidence about θ obtained is precisely the evidence obtained from the experiment actually performed. These latter principles seem so self-evident that it is hard to disagree with the Likelihood Principle, yet belief in the Likelihood Principle forces a complete revolution in thought; one must then think conditionally on the actual observation x. Further investigation (cf. Basu (1975) and Berger and Wolpert (1982)) leads to the conclusion that $l_x(\theta)$ can be meaningfully used only by considering it as a probability density with respect to a measure π, which should reflect prior beliefs about θ. Hence the result of this line of reasoning is that one must view things in a Bayesian fashion.

Reason (vii) is related to the Likelihood Principle, in that it argues that only conditional measures (based on the posterior distribution) given x are sensible for evaluating the evidence about θ, but it is less foundational and more of a 'proof by counterexample'. For instance, consider

Example 1. Suppose $X = \theta + 1$ or $\theta - 1$ with probability $\frac{1}{2}$ each ($\theta \in R^1$), and that a 75% confidence interval of smallest size for θ, based on independent observations X_1 and X_2, is desired. This is obviously given by

$$C(x_1, x_2) = \begin{cases} \text{the point } \tfrac{1}{2}(x_1 + x_2) & \text{if } |x_1 - x_2| = 2\,, \\ \text{the point } (x_1 + 1) & \text{if } |x_1 - x_2| = 0 \end{cases}$$

(or we could have chosen $(x_1 - 1)$ if $|x_1 - x_2| = 0$). But if $|x_1 - x_2| = 2$, we are *absolutely certain* that $\theta = \frac{1}{2}(x_1 + x_2)$, while if $|x_1 - x_2| = 0$ we are *equally uncertain* whether $\theta = X_1 + 1$ or $\theta = X_1 - 1$ (barring specific prior information). In either case, it seems absurd to report $C(X_1, X_2)$ as being a 75% confidence interval. The point, of course, is that frequentist measures such as 'confidence' can be totally misleading for given data x. The frequentist can protest that such measures as 'confidence' are not to be interpreted conditionally, but what is the sense in proposing a measure of accuracy which clearly presents a false image of the information about θ contained in the data. (The Bayesian posterior credible regions for this situation are, of course, very sensible.)

Examples are available for essentially any non-Bayesian measure of accuracy (or at least any frequency measure of accuracy), showing that the measure can very inaccurately portray the information about θ contained in the observation x. After seeing enough of these examples, posterior measures start to look very attractive.

All sorts of classical defences and objections to reasons (vi) and (vii) can, of course, be raised, such as bringing in questions of design, stopping rules in sequential analysis, analysis in nonparametric situations (where a likelihood function may not exist), allowing 'conditional' frequentist statements, etc., but they all seem to be answerable. Further discussion here would be inappropriate and can be found in Basu (1975) and Berger and Wolpert (1982), which also contain earlier references.

2.2. *Justification for Assumption* II

Assumption II seems almost transparently obvious, yet there is considerable resistance to it among many Bayesians. Hence a brief discussion seems in order.

In the first place, there are situations in which it seems simply unreasonable to expect that beliefs can even be modeled by a single prior distribution. Consider the following simple (though admittedly artificial) example, essentially given in Zaman (1982).

Example 2. Consider 3 boxes labelled A, B, and $2B$, one of which contains a ball. The *only* information you have is that box $2B$ is twice as likely to contain the ball as box B. You are to determine your subjective probabilities p_A, p_B, and p_{2B} of the ball being in the indicated box. Clearly you should have $p_{2B} = 2p_B$, but it is not clear what else can be said. Since nothing is known comparatively about A and B, it seems that one should have $p_A = p_B$, but by the same reasoning one would say $p_A = p_{2B}$, and both cannot hold. It does seem reasonable to suppose that one's prior probabilities should satisfy the constraints $p_B \leq p_A \leq p_{2B}$ and $p_{2B} = 2p_B$, but it is unreasonable to expect anything more precise to be concluded.

Even if in a situation where it is reasonable to expect beliefs to be expressible in terms of a single prior distribution π_T, can this actually be done? Consider, for instance, any of the axiomatic systems which guarantee the existence of π_T. (In the situation of Example 2, at least one of the axioms in any system will be violated.) The prior π_T is obtained by various betting or comparison schemes, but is *exactly* nailed down only after an infinite process of elicitation. This is clearly the case when Θ is infinite (or when the associated σ-field of events is infinite) since there are then simply an infinite set of probabilities to determine. Even when Θ is finite, the axiomatic systems formally call for considering an infinite number of bets or comparisons. In the betting schemes, one must compare all possible wagers, and indeed should really base the bets on a utility function which itself takes an infinite amount of time to perfect; and in the comparison schemes one must compare events with the infinite set of measurable events from some auxiliary—say, uniform—distribution. And all this assumes that Θ is known, whereas in many situations the possible states of nature are only vaguely comprehended (cf. Shafer (1979, 1981a, 1981b) and Barnard (1982)).

From a strictly intuitive viewpoint it is also clear that the single prior axiom systems are, in a sense, inapplicable, since there is obviously a lower limit to the accuracy of prior elicitation. I cannot believe that anyone could ever distinguish between $P(A) = 0.25$ and $P(A) = 0.250001$ (or $P(A) = 0.25 + 10^{-100}$ if an extreme case is needed) in terms of subjective elicitation. Thus Savage (1961) says

> "No matter how neat modern operational definitions of personal probability may look, it is usually possible to determine the personal probabilities of important events only very crudely."

Similar views can be found in Koopman (1940), Good (1950, 1962a, 1973 (priggish principle 3)), Savage (1954), Smith (1961), Dempster (1967, 1968), Fine (1973), Kyburg (1974, 1976), Suppes (1975), Levi (1980), Rios and Girón (1980), De Robertis and Hartigan (1981), and Zaman (1982).

Some Bayesians argue that the concept of a 'true' prior π_T is meaningless, in that the approximate prior π_A that one arrives at after a finite amount of time is your true prior at the moment, and should hence be used as such. In the face of infinite Θ this is clearly not very reasonable, since in a finite amount of time an infinite set of probabilities cannot be specified without introducing a large degree of arbitrariness. Even if only a finite Θ is involved, it seems unreasonable to look upon π_A as any form of truth, since further thought would likely cause further refinement and there is always a considerable fuzziness in subjective elicitation. The distinction between π_T and π_A is made very succinctly by Dickey (1976a, 1976b), who calls them the 'actual prior distribution' and the 'operational prior distribution', respectively, and points out situations in which π_A can be known to be a good approximation to π_T. (It is possible to argue philosophically that π_T is essentially an imaginary quantity itself—cf. parts of Levi (1980)—but it is often a useful imaginary quantity to consider.)

As an aside, it is interesting to observe that, in the above light, the subjectivist Bayesian objections to the objective Bayesian use of 'noninformative' priors seem less forceful. In a situation where there is very little prior information about θ, a noninformative prior may be a better approximation to π_T than any hastily derived proper subjective approximation π_A.

Another situation, in which working with a class of priors is clearly unavoidable, is when group conclusions or decisions must be made and the priors of all members of the group must be considered. (See Weerhandi and Zidek (1981) and Zidek (1982) for discussion and earlier references.) The issue of scientific communication is related to this, the (often unattainable) ideal being that of presenting a conclusion which would be the conclusion for any reasonable prior that a user of the information might have. (Among the works bearing on this issue are Hildreth (1963), Dickey (1973), and Jackson, Novick, and DeKeyrel (1980).) Although ideas in these areas must bear a strong relationship to those discussed in this paper, we will not be formally considering such group situations.

The above arguments do not, of course, establish that a serious problem exists with standard (i.e. single prior) Bayesian analysis. Indeed I am very sympathetic to the claim that single prior Bayesian analysis is the ideal goal and that the major problem remaining is that of developing good prior elicitation techniques. There is a very substantial and growing literature on the subject of prior elicitation (cf. Kadane, Dickey, Winkler, Smith and Peters (1980) for discussion and other references), and as better elicitation methods become available it is natural to expect the need for consideration of Bayesian robustness to decline. The validity of Assumption II from a philosophical viewpoint seems clear, however.

2.3. Reasonable classes of prior distributions

In subsection 2.2 it was argued that quantification of prior beliefs can never be done without error, and hence that one is left, at the end of the elicitation process, with a set Γ of prior distributions which reflect true prior beliefs; i.e., π_T is an unknown element of Γ. Some comments are in order concerning the specification of Γ.

The first and most crucial realization is that, in quantification of prior beliefs, only prior probabilities and relative likelihoods can accurately be elicited. In other words, such features of the prior distribution as percentiles and shape features (unimodality, monotonicity, symmetry, smoothness, etc.) can be elicited with some confidence, while features such as moments and exact functional form are much harder to accurately determine. The reason for this is simply that assessment of probabilities of events (and hence of percentiles of the prior distribution) is certainly feasible, as likewise is intuitive comparison of the 'likelihood' of the various θ (at least to the extent of leading to reasonably certain structural information about the prior density). To specify a prior moment, on the other hand, demands very accurate specification of the 'tail' of the prior distribution, which will almost never be feasible. Consider the following example.

Example 3. Suppose θ is an unknown normal mean, and that a necessarily brief period of prior assessment results in the conclusions that

$$p^{\pi}(\theta \leqslant -1) = p^{\pi}(-1 < \theta \leqslant 0) = p^{\pi}(0 < \theta \leqslant 1) = p^{\pi}(1 < \theta) \cong \tfrac{1}{4},$$

and that π has a symmetric unimodal density. Thus Γ could reasonably be chosen to consist of all priors with symmetric unimodal densities having median 0 and quartiles ± 1. (To be certain of robustness, it would probably be better to choose Γ to consist of all priors with medians within ε_1 of zero and quartiles within ε_2 of ± 1, with similar leeway for error allowed in the specification of symmetry. This will not make much difference in this situation, however.)

In this example, Γ contains both the conjugate prior $N(0, 2.19)$ (normal with mean zero and variance 2.19) and the $\mathscr{C}(0, 1)$ prior (Cauchy with median zero and scale parameter 1). The normal prior has all moments, while the Cauchy prior has no moments whatsoever. It seems very unlikely in this situation to expect detailed knowledge of the tail of the distribution (i.e., detailed knowledge of a set of very small prior probability), so any attempt to specify Γ by prior moments seems fraught with peril.

An alternative reasonable approach to specifying Γ is to approximate π_T by a specific assessed approximation π_A, and then let Γ consist of all priors 'close' to π_A. Again, 'close' should be measured in terms of close probabilities, such as in the class

$$\begin{aligned}\Gamma = \{\pi\colon \pi(\cdot) = (1-\varepsilon)\pi_A(\cdot) + \varepsilon P(\cdot),\\ P \text{ an arbitrary probability distribution}\},\end{aligned} \tag{2.1}$$

where ε reflects the believed accuracy of the prior assessment. This class Γ was first considered in Schneeweiss (1964), Blun and Rosenblatt (1967), and Huber (1973). Other reasonable classes can be found in Berger (1980b).

Much of the literature involving classes of priors chooses Γ to be either the set of priors with certain moments in specified ranges or a set of priors of a particular functional form with parameters in specified ranges. While these tend to be much easier to work with than are Γ such as in Example 3 or (2.1), they are unsuitable, as discussed earlier.

Easily specified classes, such as (2.1), are often somewhat too large, in that they do not incorporate probably available smoothness information about π. In (2.1), for example, it may well be felt that π definitely has a unimodal continuous density, which would place severe restrictions on the contaminations P allowed. Thus if, in doing a robustness analysis with respect to a Γ as in (2.1), robustness seems hard to achieve, make sure this is not due to unrealistic features of Γ. Of course, if robustness with respect to Γ is obtained, then one is also robust with respect to the more reasonable subclass.

It is, of course, possible to have Γ much less clearly specified than in the above examples, such as when θ is a high dimensional vector or, even worse, a nonparametric index. Only extremely crude or general features of the prior might then be obtainable, so Γ could be very large.

2.4. *Updating Γ*

Since we will primarily be concerned with posterior measures, the question of updating Γ by the data is obviously crucial. When making posterior conclusions, the obvious class of posteriors to consider is simply

$$\Gamma^* = \{\pi(\cdot \mid x) \colon \pi \in \Gamma\}.$$

Unfortunately, more flexibility must be allowed if realism is to be achieved. The main difficulty is that, especially in multivariate problems, it would be far too time consuming (if even possible), to accurately ascertain even the most important features of the prior ahead of time. After seeing the data, however, one can determine which features of the prior will have a real impact and must carefully be considered. For example, in a complicated linear model the data may illuminate which variables are important and hence should be the focus of the prior elicitation. Or the data may indicate that some variables are accurately determined by the data, and hence prior information concerning them is likely to be less important, while other variables are very inaccurately determined from the data (due, say, to multicollinearity) and hence need accurate prior specification. Thus the data may cause further prior elicitation resulting in a reduced class Γ^* (which will then be updated by the data in the usual Bayesian fashion).

Major objections to the above approach can, of course, be raised, most troubling being the apparent dependence of the prior (not just the posterior) on the data. This offends many Bayesians, and also smacks of cheating and adhocery to non-Bayesians. To Bayesians, I can only reply that there is no choice. Typical situations have high dimensional Θ, for which it is very unrealistic to suppose that suitably accurate prior specification can be achieved; i.e., only very large Γ can be determined prior to experimentation. It will be very unlikely that robustness can be achieved with respect to such a large Γ. Hence a narrowing down of Γ will be needed, with the data indicating where further refinement is necessary. Note that the data is not to be used to shape your beliefs, but only to indicate how this narrowing down should be done, and when a point is reached which allows reasonably robust Bayesian conclusions to be drawn. As Hill (1965) says,

> "...it is only the degree of care we take in approximating our prior, not the prior itself, that depends on the data."

A more troubling situation is when the data reveals that Γ was in some sense wrong, and not just too big. One could argue that Γ should have been kept flexible enough to encompass all possibilities, but realistically the data will often suggest new relationships, hypotheses, or models that were not included in, and may even contradict, the original specification of Γ. One must then go back and suitably enlarge or change Γ, as observed in de Finetti (1972, Ch. 8) and Savage (1962). As Savage said, however,

> "It takes a lot of self-discipline not to exaggerate the probabilities you would have attached to hypotheses before they were suggested to you."

The disturbing nature of allowing the data to affect Γ directly does not seem quite so bad if a slightly different perspective is adopted. Instead of viewing the situation as that of updating prior information, think of it as an attempt to quantify (after the experiment) the relevant experimental and nonexperimental information, and then combine the two. This, of course, is the view that outsiders, evaluating a robust statistical analysis, will take. A good analysis will present a suitable summarization of the data along with a description of the experimenter's Γ and his conclusions. In evaluating this, an outsider would consider the suitability of Γ, and alter Γ to reach his own conclusions if needed. How Γ was obtained is essentially irrelevant; either it seems a reasonable representation of the nonexperimental evidence or it does not. The emphasis here is on the *effect* of the data *on* opinions, or on the prior to posterior *transformation*, a concept convincingly promoted by Dickey (1973). Another way of thinking of this is that one learns by passing a variety of reasonable priors over the likelihood function $l_x(\theta)$ and seeing what happens. The strict prior to posterior mode of reasoning is then deemphasized. (Indeed, Schafer (1981b) argues convincingly that practical Bayesians almost never think in this strict mode, but instead view the problem as that of combining different sources of information.)

The clear difficulties of updating, by merely conditioning on the data via Bayes rule, have led to the development of other theories or methods for Bayesian or pseudo-Bayesian analysis. (See, for example, Jeffrey (1968), Shafer (1976, 1979, 1981a, 1981b, 1982), Teller (1976), and Diaconis and Zabell (1982).) These alternatives are interesting, but I remain unconvinced as to the practical necessity of developing a methodology which goes beyond post-data modification of Γ, followed by updating via Bayes rule. First of all, most of the examples against Bayesian updating can be handled by allowing post-data modification of Γ. Secondly, complex situations are understood by trying to break them into simple components for separate analysis; the prior–data, or alternatively, experimental–nonexperimental information decomposition is a very useful such breakdown, with a known method (Bayes rule) for recombination. Although contamination of information is certainly a real danger, and there may be situations where this breakdown is not necessary, in the overwhelming majority of the cases it is successful. A final argument for staying within the framework of Bayesian conditioning is that, as alluded to earlier, it is very important in statistical reports to separate the information contained in the data from that in the prior, and so this breakdown should be attempted even when not convenient.

While the above reasons argue against basing one's methodology on non-Bayesian updating, it would be foolish to rule out alternate methods completely. (See the discussion of this issue in Shafer (1979, 1981a, 1981b).) Also, certain ideas derived from these alternate viewpoints are useful in post-data modification of Γ. One such idea is the use of Jeffrey's rule, as discussed in Diaconis and Zabell (1982) and Shafer (1981a).

3. Measures of robustness

The natural Bayesian measure of robustness is insensitivity of the final (posterior) conclusion to the choice of $\pi \in \Gamma$. This will be discussed in the next section. Though of central importance, this measure of robustness will be seen to be inadequate in some situations, necessitating measures of robustness of procedures based on overall performance, such as Bayes risk, $r(\pi, \delta)$, in decision theoretic situations. This will be discussed in subsection 3.2. Subsection 3.3 discusses the role of each of these two methods of measuring robustness.

3.1. Posterior robustness

Assumption I, being the cornerstone of the robust Bayesian viewpoint, must be followed. Hence, after observing all the data, any inference or decision made should be satisfactory from a posterior viewpoint.

Definition 1. An inference or decision is *posterior robust* with respect to Γ if it is satisfactory with respect to $\pi(\cdot \mid x)$ for all $\pi \in \Gamma$.

This definition is necessarily very vague, but could be tightened up in specific situations, such as in the following reasonable definition for decision theoretic settings.

Definition 2. In a decision theoretic setting (see subsection 1.3), an action a_0 is *ε-posterior robust* with respect to Γ for the observed x if

$$\sup_{\pi \in \Gamma} |\rho(\pi, x, a_0) - \inf_{a \in \mathscr{A}} \rho(\pi, x, a)| \leq \varepsilon . \tag{3.1}$$

It is important to realize that whether or not posterior robustness exists will often depend on which x is observed. Thus Barnard (1982) says

> "We should recognize that 'robustness' of inference is a conditional property—some inferences from some samples are robust . . ."

Consider the following example.

Example 4. Assume that $X \sim \mathrm{N}(\theta, 1)$ is observed, and that it is desired to estimate θ under loss $L(\theta, a) = (\theta - a)^2$. Here $\Theta = \mathscr{A} = R^1$. Suppose $\Gamma = \{\pi_N, \pi_C\}$, where π_N is the $\mathrm{N}(0, 2.19)$ distribution and π_C is the $\mathscr{C}(0, 1)$ distribution. (This Γ is a very specialized subset of the Γ in Example 3, but behaves similarly in many respects.) If π_N were the true prior, then one would want to use the Bayes estimate

$$\delta^N(x) = \frac{2.19}{1 + 2.19} x ,$$

while if π_C were the true prior, then one would want to use the Bayes estimate

$$\delta^C(x)=\frac{\int \theta(1+\theta^2)^{-1}\exp\{-\frac{1}{2}(x-\theta)^2\}\,d\theta}{\int (1+\theta^2)^{-1}\exp\{-\frac{1}{2}(x-\theta)^2\}\,d\theta}.$$

Table 1 gives a few values of δ^N and δ^C.

Table 1
δ^N and δ^C

x	0	1	2	10
δ^N	0	0.69	1.37	6.87
δ^C	0	0.52	1.27	9.80

An easy calculation shows that, for squared error loss,

$$|\rho(\pi, x, a_0)-\inf_a \rho(\pi, x, a)| = (a_0-\mu_\pi(x))^2,$$

where $\mu_\pi(x)$ is the posterior mean for π. Since δ^N and δ^C are μ_{π_N} and μ_{π_C}, respectively, it follows that the posterior robustness of either $\delta^N(x)$ or $\delta^C(x)$ is measured by

$$[\delta^N(x)-\delta^C(x)]^2.$$

From Table 1 it is clear that either action is quite posterior robust (i.e., $\delta^N(x)$ is close to $\delta^C(x)$) for x near zero, while for $x = 10$, neither action is posterior robust. (For large x, the tail of the prior becomes very significant, and π_N and π_C have substantially different tails.)

If posterior robustness is attainable in a given situation, then the problem is essentially solved. If posterior robustness is not attainable, however, as happens in Example 4 when $x = 10$, then something else most be done. The natural thought is to attempt further elicitation of the prior distribution, and indeed it is precisely when posterior robustness does not obtain that more detailed elicitation is indicated. If this resolves the issue, fine, but if further elicitation is not possible or won't prove helpful (as in Example 4 for $x = 10$, where the prior tail will be next to impossible to accurately specify), then we must look beyond posterior robustness. (Of course, the above example is extreme, in that if encountered in practice one would seriously suspect the model for X. Extreme examples like this are useful for emphasizing the issues, however. They also provide insight which can be used in less extreme situations. Sections 5 and 6 deal with more practical issues.)

3.2. Procedure robustness

Faced with the (X, θ) experiment, one can talk about the procedure $\delta(X)$ to be used when X is observed. Although the Bayesian tends to think conditionally on the observation $X = x$, it is certainly possible to consider the collection $\{\delta(x): x \in \mathscr{X}\}$ of inferences or decisions to be made for all possible X. (This may seem an unnecessary complication, but is logically sound.) Since, pre-experimentally, the Bayesian thinks that X will be occurring according to the marginal distribution $m(\cdot)$, he would (in a decision theoretic setting, for simplicity) evaluate the overall performance of a procedure by

$$r(\pi, \delta) = \mathrm{E}^m[\rho(\pi, X, \delta(X))] .$$

A reasonable method of measuring the robustness of a procedure in such a situation is given in the following definition.

Definition 3. In a decision theoretic setting, the procedure δ^0 is ε-*procedure robust* with respect to Γ if

$$\sup_{\pi \in \Gamma} [r(\pi, \delta^0) - \inf_{\delta} r(\pi, \delta)] < \varepsilon .$$

Example 4 (*continued*). Calculation shows that $r(\pi_C, \delta^N) = \infty$, $r(\pi_N, \delta^N) = 0.697$, $r(\pi_C, \delta^C) < 1$, and $r(\pi_N, \delta^C) = 0.736$. Hence the procedure robustness of δ^C (with respect to Γ) is measured by

$$r(\pi_N, \delta^C) - r(\pi_N, \delta^N) = 0.049 ,$$

while that of δ^N is measured by

$$r(\pi_C, \delta^N) - r(\pi_C, \delta^C) = \infty .$$

Clearly δ^C is much superior according to this measure of robustness. (Of course, the use of an unbounded loss function can be criticized, but even for many reasonable bounded losses δ^C would prove far superior.)

Many Bayesians object to the use of $r(\pi, \delta)$ as a measure of anything, because it involves an average over the sample space. A statistician should be responsible for the long run performance of his methodology, however. In the situation of Example 4, for instance, the Bayesian who time after time uses the conjugate prior Bayes rule δ^N will have very bad long run performance if π_C is the true prior fairly regularly, while the Bayesian who uses δ^C suffers no such danger when π_N is the true prior. In other words, if a Bayesian is to employ a methodology leading to the use of a procedure δ, he should be concerned that his methodology is sound, as reflected by $r(\pi, \delta)$. This is not to say that a procedure δ is good for all x if $r(\pi, \delta)$ is good (the fallacy in reasoning underlying frequentist statistics), but does say that δ is bad if $r(\pi, \delta)$ is bad. (Discussion of other reasons for considering $r(\pi, \delta)$ will be given in subsection 4.4.)

Many Bayesians react to the above argument by asking how $r(\pi, \delta)$ can be bad if $\delta(x)$ is chosen to be good from a posterior viewpoint for each x. Example 4 provides an illustration of how this can happen. From the viewpoint of posterior robustness, $\delta^N(x)$ and $\delta^C(x)$ were equivalent, in that the posterior robustness of each (with respect to Γ) was measured by

$$[\delta^N(x) - \delta^C(x)]^2 .$$

But from the procedure robustness viewpoint, it seems clear that δ^C is considerably better than δ^N.

From the procedure robustness viewpoint, several specific criteria have been proposed for the selection of procedures. The two most common are the *Γ-minimax* and *Γ-minimax regret* criteria, which propose the use of the procedure δ^* which minimizes

$$\sup_{\pi \in \Gamma} r(\pi, \delta^*) \tag{3.2}$$

or

$$\sup_{\pi \in \Gamma} [r(\pi, \delta^*) - \inf_{\delta} r(\pi, \delta)] , \tag{3.3}$$

respectively. Discussion of the literature on these criteria will be delayed until Section 5.

3.3. Discussion

It is important to realize that posterior robustness is the ideal goal. If it can be attained, the problem is solved. Also, when posterior robustness is not present, a careful Bayesian will attempt further refinement of Γ or, if possible, attempt to obtain more data. Unfortunately, situations where posterior robustness is simply unattainable are common, such as when (i) because of time or mental limitations further refinement of Γ is impossible; (ii) no more data can be obtained; or (iii) Bayesian analysis is technically too difficult to implement for a convincing variety of plausible priors (as in many nonparametric problems).

What alternatives are available when posterior robustness cannot be found? First, one could simply say that there is no clearcut answer to the problem. This is reasonable, at least in those situations where Γ is clearly defined and different priors in Γ give substantially different answers. If, however, the problem is due to technical difficulties in implementing the Bayesian approach, or if an answer simply must be obtained, then something else must be tried.

The natural Bayesian inclination would be to put some 'metaprior' on Γ itself and use the resulting Bayes rule. If technically feasible, this may well be a good ad hoc solution. We stress 'ad hoc' because the assumption is that no further prior elicitation is possible. Thus the metaprior is simply some arbitrarily chosen distribution used as a technical device to obtain an answer. The analysis with metapriors can be very formidable, however, especially with Γ such as discussed in Section 2.3. Also, there is nothing to guarantee that the resulting answer will be good. Hence it may well be useful to consider procedure robustness and/or use of frequency measures as an aid in obtaining an answer. A more extensive discussion of the use of procedure robustness and frequency measures will be given in Sections 4.4 and 4.5.

The complaint can be raised that use of procedure robustness may violate Assumption I and the Likelihood Principle, and also that use of such measures as (3.2) and (3.3) and frequency measures will violate the rationality or coherency axioms. This is a valid complaint, yet carries no real force since a point has been reached where there is no clearcut 'coherent' way to proceed. Here, coherent is being used in a broad sense, since it would formally be coherent (in the usual sense) to arbitrarily select some metaprior on Γ and do a Bayesian analysis, yet few Bayesians would say that arbitrary choice of a prior (i.e. a choice not based on any subjective opinions) is necessarily good. Thus Levi (1980) says

> "We should, therefore, recognize a distinction between principles of rationality regulating an agent's commitments and the suggestions which may be made when he cannot live up to them."

It should be stressed that we are not recommending any definite way of proceeding when posterior robustness is lacking. Often, putting an artificial prior on Γ may work. Often (see Sections 4.4 and 4.5) use of procedure robustness or frequency measures may prove helpful. Or entirely different statistical methodologies may provide good answers. In fact, the coherency arguments essentially suggest that no *single* automatic prescription concerning what to do in this situation will always prove successful.

It is crucial, finally, to recall that we are contemplating straying from the Bayesian path only to select from among answers which are plausible from a posterior Bayesian viewpoint, and hence will not be knowingly violating Assumption I or coherency by any substantial amount. Thus Good (1976) says

> "... non-Bayesian methods are acceptable provided that they are not seen to contradict your honest judgements, when combined with the axioms of rationality."

4. Implications of the robust Bayesian viewpoint

The major implications of the robust Bayesian viewpoint have already been discussed, but the flexibility of the approach allows incorporation of various sensible, yet ostensibly 'non-Bayesian', techniques. Some of these are briefly discussed below.

4.1. Data analysis

The data summarization part of data analysis is justifiable from any viewpoint, so it is the interactive modeling aspect which is of interest. This activity always involves the combining of subjective knowledge with the data to suggest or modify models for the phenomenon being studied, and is hence essentially Bayesian in nature. As discussed in subsection 2.3, it seems sensible and necessary to allow modification of Γ based on the data, and indeed, with this option, the robust Bayesian and data analyst behave in essentially the same way. The differences are, first, that the robust Bayesian believes in quantifying the subjective information (to the extent possible) in Γ, rather than incorporating it in an ad hoc fashion; and, second, the robust Bayesian uses posterior measures in evaluating the evidence for any model or conclusion. This last feature eliminates, in a sensible fashion, the problems of evaluation of the strength of the evidence for a model selected by the data. (The posterior weight given to the model is based on a product of the prior weight and likelihood according to the data, automatically discounting the 'significance' of the data for the model it selects.) Hence, contrary to popular opinion, the robust Bayesian is not the slave of a particular prior distribution he must pre-experimentally specify, and can engage in sensible data analysis (as opposed to non-Bayesian data analysis).

4.2. Randomization

Most statisticians are convinced of the value of randomization in statistical design (e.g. random allocation of subjects to two treatments), yet the single prior Bayesian position does not allow this. If *all* unknowns in the situation have been identified and their true prior distribution obtained, then the optimal Bayesian design will not require any form of randomization. When, however, uncertainty in the prior information is admitted, randomization becomes available.

The use of randomization to a robust Bayesian, however, is essentially limited to the effort of avoiding experimenter induced bias. In other words, because the robust Bayesian is worried that there are experimental factors which he has not thought of and which may be correlated with any nonrandom subject selection or allocation scheme, he will find randomization to be useful in (hopefully) preventing such bias.

The robust Bayesian does not (as an ideal) find randomization to be of use in drawing conclusions from the data. The probabilistic mechanism of randomization will usually be independent of θ, and hence by Assumption I the robust Bayesian will want to draw conclusions conditional on the given selected sample. Of course, even the non-Bayesian agrees with this to some extent, the 'selection' of a new randomization design if the original design doesn't look random enough being one example. And even the most ardent anti-Bayesian would not go through with a standard classical analysis based on the randomization if significant cofactors were revealed which, by bad luck, turned out to be highly correlated with, say, the treatment groups. Yet the Bayesian conditional viewpoint argues against making *any* use of the randomization mechanism. Arguments for this viewpoint can be found in Basu (1971) and Basu (1980). (See also the discussion by Lindley in Basu (1980).)

It is possible to argue that robustness considerations allow the use of the randomization mechanism. For instance, Rubin (1978) argues that the prior specification is so immensely complicated in typical situations that it will often be better to 'ignore' part of the data (i.e. the known outcome of the randomization) to simplify the needed prior specification. The probability mechanism of the randomization does then become part of the Bayesian analysis and can indeed simplify matters.

The danger in this is, of course, the usual danger befalling any attempt to analyze data in violation of Assumption I; the analysis conditional on the data could differ substantially from the analysis averaging over data points that could have been obtained. Although this is something that will probably occur fairly rarely, it is unappealing to adopt as a basic method of analysis techniques which can lead to conclusions at odds with all the actual data. Note that the robustness advocated in this paper is not of this potentially dangerous type, since satisfactory conditional posterior behavior behavior is of primary importance.

There may, of course, be very pragmatic considerations involved. For example, a randomized design will be useful if it seems important to convince others that the experiment was 'unbiased' (although this is rather illusory impartiality). Also one can be very sympathetic to the argument that any Bayesian analysis here, much less a robust Bayesian analysis, is simply unmanageable.

Discussion of the randomization issue can also be found in Savage et al. (1962), Hill (1970), Good (1976 and earlier), Basu (1980), Lindley and Novick (1981) and Berger and Wolpert (1982). Also, the debate in sampling theory concerning the use of superpopulation models as opposed to analysis based on the probabilistic mechanism of the sampling rule is essentially the same as the randomization debate. Indeed Godambe and Thompson (1977), Godambe (1982) and Royall and Pfefferman (1982) specifically argue that suitable random sampling plans can lead to a form of Bayesian robustness. Other discussion and references can be found in Cassel et al. (1977), Basu (1978), Hájek (1981) and Berger and Wolpert (1982).

4.3. Classical robustness

By classical robustness is meant robustness with respect to the distribution $P_\theta(\cdot)$ of the observation X. This is obviously a crucial aspect of statistical analysis and can be included in the robust Bayesian framework by the simple expedient of allowing Θ to be a nonparametric index set (indexing the distributions for X which are of concern) and having Γ reflect the prior knowledge available about these distributions. Indeed, to many Bayesians the difference between 'model' and 'parameter' seems fuzzy at best. The *subjective* choice of the model is often a far more drastic use of prior information than is use of prior distributions on parameters of the model.

Classical robustness results tend to be in terms of measures such as 'asymptotic minimaxity' (cf. Huber (1972)), which can be related to procedure robustness. Procedure robustness is of interest here because Bayesian analysis when P_θ is uncertain can be technically very difficult. A number of successful Bayesian analyses of model robustness problems have been carried out, however. For the most part these studies proceed by embedding a standard family of distributions in a larger parametric family (such as embedding the normal distributions in the class of all t-distributions), and then performing a Bayesian analysis. Excellent discussions of this, along with earlier references, can be found in Box and Tiao (1973), Dempster (1975) and Box (1980).

One important point brought out in the Bayesian view is that model robustness should be viewed conditionally. If a data set gives residuals which are a gorgeous fit to normality, worrying about robustness to normality is a waste of time. Discussion and examples can be found in Dempster (1975) and Barnard (1982). Efron and Hinkley (1978) and Hinkley (1983) also discuss important situations in which model robustness should be investigated conditional on shape features of the data. All this is in line with our view that having valid conditional (posterior) measures is of primary importance.

4.4. Uses of frequency measures

Frequency measures can have a role to play in robust Bayesian analysis. The basic idea of frequency measures is, of course, to also consider x other than that which occurs. The simplest form of such reasoning, which can be useful to a Bayesian, is simply to imagine possible data x, compute the Bayes rule for a prior being investigated and see if the result makes sense. In the situation of Example 4, for instance, the fact that δ^N appears inadequate for $x = 10$ provides a warning that δ^N might also be inferior for a smaller (yet possible) observation such as $x = 5$. Several very interesting examples of this type of reasoning are given in Diaconis and Freedman (1983). Looking at the behavior of a Bayes rule for a variety of x (often extreme x) may point out unsuspected and unacceptable features of any chosen prior. This has been called the 'device of imaginary results' by I.J. Good and has been extensively promoted by him (cf. Good (1965, 1976, 1981)).

More formally, frequentist measures, such as operating characteristic curves and risk functions can be of interest through their relationship to procedure robustness. (This was briefly discussed in subsection 3.2, but, since the issue is quite controversial, an expanded discussion is in order.) The basic reason for this relationship is (1.2), namely that

$$\mathrm{E}^m \rho(\pi, X, \delta(X)) = r(\pi, \delta) = \mathrm{E}^\pi R(\theta, \delta) . \tag{4.1}$$

(Although $R(\theta, \delta)$ and $\rho(\pi, x, \delta(x))$ were defined as frequentist risk and posterior expected loss, respectively, through appropriate choice of the loss function they can be made to represent nondecision theoretic measures such as coverage probability and posterior probability of containing θ, respectively.) If, now, $R(\theta, \delta)$ is known to be 'good' for all θ, then from (4.1) it follows that $\mathrm{E}^m \rho(\pi, X, \delta(X))$ will be 'good' for all π. Although this doesn't guarantee that $\rho(\pi, x, \delta(x))$ is actually good for the observed x and π of interest, there is a good chance that it will be. Conversely, if $R(\theta, \delta)$ is bad for some θ, then before using δ it is imperative to make sure that such θ are really very unlikely a priori. In Example 4, for instance,

$$R(\theta, \delta^N) = 0.471 + (0.0983)\theta^2 ,$$

which is terrible for large θ. Looking at this risk would cause one to realize that, unless the large θ really are as unlikely (subjectively) as indicated by the tail of the presumed normal prior, then use of δ^N may not be wise.

Besides this aspect of using frequency measures as a check on Bayesian robustness, two closely related reasons for admitting consideration of frequency measures should be discussed. First, there are simply many problems which have a good frequency answer, and yet which do not have clearly trustworthy Bayesian answers. Because of (4.1), the frequency procedure has a good chance of also being sensible from a conditional posterior Bayesian viewpoint. Thus it can be viewed as a good 'stab in the dark'. Of course, as Bayesian methodology expands, there will be less and less need to depend on such frequency evaluations. (See Berger (1982d) for examples, discussion and references.)

The final reason for consideration of frequency measures and procedure robustness is that, like it or not, the majority of users of statistics are not going to be extremely well trained, and will probably not be capable of careful Bayesian sensitivity analyses. For such users it is necessary to provide procedures, which are as Bayesian as possible and yet are *automatically* robust. Since these procedures will be used repeatedly, their long run frequency performance is definitely relevant. Example 4, for instance, suggests that in estimating a normal mean it would be reasonable to ask the unsophisticated user to specify a 'guess' and an estimate of the accuracy of this guess, and then fit this to a Cauchy prior and calculate the Bayes estimate (all of which could be automatically done by a computer). Fitting to a conjugate normal prior is contraindicated, however, at least for such automatic use. This section concludes with a very brief review of some useful frequency concepts.

A. *Design, prediction and sequential analysis*

In these problems it is absolutely imperative to average over the data likely to occur and no Bayesian would think otherwise. Of course, these problems also have a large Bayesian component. In design, for instance, one must use subjective gusses for θ to predict what data will occur and hence what design to use. Also, a Bayesian will have the goal of obtaining good conditional performance, which may lead to a quite different design than a classical design.

B. *Confidence procedures*

If $C(x)$ is a confidence procedure for θ with confidence level $1-\alpha$, then

$$P_\theta(C(X) \text{ contains } \theta) \geqslant 1-\alpha .$$

As in (4.1), it follows that

$$\mathrm{E}^m P^{\pi(\theta|X)}(\theta \in C(X)) \geqslant 1-\alpha , \tag{4.2}$$

so that, for small α, $C(x)$ has a pretty good 'chance' of containing θ (according to a valid posterior measure) no matter what π is. This use of confidence procedures was discussed in Pratt (1965).

Morris (1983a, 1983b) has advocated the development of procedures satisfying (4.2) for all priors π in a feasible class Γ, and has called this 'empirical Bayes confidence'. For the reasons discussed earlier, this may well be a valid objective, as long as it is kept in mind that the real goal is to obtain a set with good posterior probability of containing θ for the given observation x. Similar ideas are employed in Godambe and Thompson (1976) and Godambe (1982) to argue for use of frequentist concepts in obtaining robust Bayesian confidence procedures in survey sampling. Other work on the relationship between frequency and Bayesian confidence methods can be found in Welch and Peers (1963) and Stein (1981b), which also contain earlier references.

C. *Minimaxity*

The robust Bayesian interest in minimaxity arises from the fact that

$$\sup_\pi r(\pi, \delta) = \sup_\theta R(\theta, \delta) , \tag{4.3}$$

and hence a minimax decision rule (i.e. a rule minimizing the right hand side of (4.1)) is also the 'most procedure robust' Bayesian decision rule (being Γ-minimax when Γ is the class of all priors). Although realistic Γ will rarely be so large that

$$\sup_{\pi \in \Gamma} r(\pi, \delta) = \sup_{\theta} R(\theta, \delta),$$

a minimax rule can provide a basis of comparison for procedure robustness.

D. Admissibility

If δ is inadmissible, there will often exist a δ^* such that

$$R(\theta, \delta^*) < R(\theta, \delta)$$

for all θ, and hence such that $r(\pi, \delta^*) < r(\pi, \delta)$ for all priors π for which the Bayes risk exists. Because of procedure robustness and (4.1), it can be convincingly argued that this should preclude consideration of inadmissible decision rules. (See also Hill (1974).) The restriction to consideration of only admissible rules can be a very helpful reduction of the problem, particularly in areas such as sequential Bayesian analysis where even determination of a Bayes rule can be very difficult.

E. Asymptotics

Much of the frequentist work on asymptotics has relevance to a Bayesian. Some such work is discussed in Section 5. Also, asymptotics can be helpful in determining Bayesian robustness. For example, in Diaconis and Freedman (1983) it is shown that certain partially nonparametric Bayes rules can be inconsistent, giving real cause for concern as to the robustness of use of the corresponding priors.

F. Significance testing

There are sometimes relationships between P-values in significance testing of a hypothesis and posterior probabilities of the hypothesis (cf. Good (1950), Jeffreys (1961), Pratt (1965), and Berger and Wolpert (1982) which has later references), and this may sometimes justify use of the often much easier to compute P-values. Also, in Section 5 the role of Bayesian significance testing in Bayesian robustness will be briefly discussed.

We have, of course, barely touched the surface of the possible uses of frequency concepts in robust Bayesian analysis. Invariance concepts, for instance, can have many uses. Also, many explicit frequentist procedures turn out to be perfectly satisfactory from a Bayesian viewpoint.

4.5. Estimating a multivariate mean: The Stein effect

We conclude this section with an example interesting from several aspects. First, it is an example wherein both the frequentist decision theorist and the robust Bayesian decision theorist end up wanting to solve the same problem. Second, it is an example wherein the Bayesian can be amazingly robust and the frequentist can make significant use of prior information at no or little cost.

Finally, it illustrates the fact that good robust Bayes procedures need not be Bayes procedures for any prior in Γ, and indeed can violate natural Bayesian intuition.

Suppose we must simultaneously deal with p independent estimation problems ($p \geqslant 3$), where $X_i \sim \mathrm{N}(\theta_i, 1)$ is the observation in the ith problem, and the loss in estimating $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ by $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ is $\Sigma_{i=1}^p (\theta_i - \delta_i)^2$. The θ_i are a priori *known* to be independent and, as a quick approximation, are felt to have N(0, 1) prior distributions, to be denoted π_i^N, $i = 1, \ldots, p$. (Different prior medians could be allowed in the following analysis.) This last facet of the prior distribution is deemed uncertain, however, and hence robustness is sought with respect to the class of priors

$$\Gamma = \left\{\pi = \prod_{i=1}^{p} \pi_i \colon \pi_i = (1-\varepsilon)\pi_i^N + \varepsilon P_i,\ P_i \text{ arbitrary probability measures}\right\}. \tag{4.4}$$

(It is essentially certain that the θ_i are a priori independent, and ε is the assumed error in the approximations π_i^N.)

A non-Bayesian frequentist analysis of the problem must take note of the Stein phenomenon, which is that estimators δ^* exist which are better than the natural estimator $\delta^0(x) = x$, i.e.

$$R(\theta, \delta^*) < R(\theta, \delta^0) = p \quad \text{for all } \theta\,. \tag{4.5}$$

The frequentist finds himself forced somewhat into the Bayesian ballpark, however, since any such δ^* is significantly better than δ^0 only in a relatively small region of the parameter space. Intuitively, therefore, δ^* should be selected by deciding where significant improvement is most desired, and it seems manifest that significant improvement will be most desired for those θ felt likely to occur a priori. A very reasonable way of proceeding, therefore, is to elicit a rough prior distribution π_A, and then to find that δ^* which minimizes $r(\pi_A, \delta^*)$ subject to (4.5). (Such a δ^* will clearly perform best for those θ felt a priori to be most likely.) The frequentist willing to sacrifice some minimaxity (here p is the minimax risk) for more Bayesian gain would be interested in the problem

$$\text{Minimize } r(\pi_A, \delta), \quad \text{subject to } R(\theta, \delta) \leqslant p + C\,. \tag{4.6}$$

A fascinating feature of this situation is that a robust Bayesian can become concerned with the same problem. Indeed, suppose he seeks procedure robustness by trying to be Γ-minimax (see (3.2)) with respect to the Γ in (4.4), and furthermore does the 'obvious' thing and restricts attention to coordinatewise independent rules, i.e. rules of the form

$$\delta(x) = (\delta_1(x_1), \delta_2(x_2), \ldots, \delta_p(x_p))\,. \tag{4.7}$$

(Since the θ_i are a priori independent, any Bayes rule with respect to a prior in Γ will be of this form.) A relatively simple game theoretic argument shows that this problem is then equivalent to the problem in (4.6) (with δ restricted to be of the form (4.7) of course), in that there exists a continuous increasing function ρ such that $C = \rho(\varepsilon)$ defines an equivalence of solutions. It is interesting to see what happens if the restriction to estimators of the form (4.7) is dropped, so we will consider the general problem posed in (4.6).

Exact results on problems of this form are very complicated but simple approximate solutions are given in Berger (1982b) and Berger (1982c). For the special case considered here, and when $C = 0$ in (4.6) for simplicity, the approximate solutions are

$$\delta^*(x) = \begin{cases} \frac{1}{2}x & \text{if } |x|^2 \leqslant 4(p-2)\,, \\ (1 - 2(p-2)/|x|^2)x & \text{if } |x|^2 \geqslant 4(p-2)\,. \end{cases}$$

This estimator is minimax and hence not only satisfactory from the frequentist viewpoint, but also procedure robust with respect to the class of *all* priors. The estimator is also quite acceptable from the posterior viewpoint, since for $|x|^2 \leqslant 4(p-2)$,

$$\delta^*(x) = \tfrac{1}{2}x = \delta^N(x)\,,$$

where δ^N is the Bayes procedure with respect to the approximate prior $\pi^N = \Pi_{i=1}^p \pi_i^N$. (For the class Γ in (4.4) posterior robustness is achieved for small $|x|$ by any Bayes rule with respect to a prior in the class, while for large $|x|$ posterior robustness is not attainable.) As to Bayes risk, this estimator astonishingly has

$$\lambda = \frac{r(\pi^N, \delta^*)}{r(\pi^N, \delta^N)}$$

as indicated in Table 2. Thus when $p = 5$, for instance, δ^* is only 7% worse than δ^N if π^N is the true prior. Indeed $\sup_{\pi\in\Gamma} r(\pi, \delta^*)$ will be very satisfactory, as indicated by the crude upper bound

$$\sup_{\pi\in\Gamma} r(\pi, \delta^*) < (1 - p\varepsilon) r(\pi^N, \delta^*) + p\varepsilon\,.$$

(Compare this with the fact that $\sup_{\pi\in\Gamma} r(\pi, \delta^N) = \infty$.)

That one can have such fine Bayesian performance and be so robust (or, from a frequentist viewpoint, be minimax) is quite surprising. What is even more surprising from a Bayesian viewpoint is that we *know* a priori that the θ_i are independent and hence we *know* that our 'true Bayes rule' would be of the

Table 2
Bayes risk ratio of δ^* to δ^N

p	3	4	5	6	7	8	9	10	15	20
λ	1.296	1.135	1.0727	1.0427	1.0267	1.0174	1.0117	1.008	1.0016	1.0004

form (4.7). But it is shown in Efron and Morris (1971) that if only estimators of this form are considered, then about the best that can be done is to have an estimator δ^T with $\lambda = 1.4$ and $\sup_\theta R(\theta, \delta^T) = (1.3)p$. This is 40% worse than δ^N when π^N is true and 30% worse than a minimax estimator in terms of minimax risk (indicating considerably less procedure robustness), which is significantly inferior to the performance of δ^*. Hence good robust Bayesian procedures can differ substantially from what a straightforward Bayesian viewpoint might dictate and need not be Bayes with respect to any prior in Γ. (The 'formal' Bayesian solution to this problem of putting a metaprior on Γ would probably also work, although care might be needed in choosing the formal metaprior and the resulting procedure would probably be extremely messy.)

This example was, of course, very special, particularly in that the approximate priors for each θ_i were assumed to have equal variances. (Talking in terms of 'variance' is convenient for specifying π_A, here, but Γ does not assume that the prior variance is known.) Almost certainly in reality, a priori independent θ_i will have different approximate prior variances. Some partial results for the general nonsymmetric situation can be found in Berger (1982b). Similarly, it will often be unrealistic to assume that the error in the specification of each of the π_i^N is the same value ε. The almost astounding power of the Stein effect in achieving Bayesian robustness in this 'ideal' situation, however, certainly argues for its value in less ideal situations.

5. History and guidelines

There has been comparatively little research in Bayesian robustness and only a few specific guidelines are available in attempting to achieve robustness. In subsection 5.1 we briefly review the literature on Bayesian robustness, although this was not intended as a review article per se, and hence little more than a categorization of results is attempted. In subsection 5.2 the few available guidlines are presented.

5.1. History

5.1.1. Posterior robustness

A. Asymptotics

It is intuitively plausible that, as the sample size goes to infinity, the information from the data becomes conclusive, and hence the conclusions will depend very little on the prior (automatically achieving posterior robustness). Results in this area can be divided into the categories of 'stable measurement', 'consistency', and 'sequential analysis'. Summaries of much of this work can be found in De Groot (1970).

A(*i*). *Stable measurement.* The principle of stable measurement is roughly that, as the sample size goes to infinity, the posterior distribution of θ becomes essentially proportional to the likelihood function (i.e. the prior distribution washes out). This concept was extensively promoted by Savage (cf. Edwards, Lindeman and Savage (1963) and most of the other works of Savage listed in the references). Blackwell and Dubins (1962) explored a similar concept.

Since the likelihood function will generally be asymptotically normal, it is reasonable to expect the posterior distribution to be asymptotically normal. Results in this direction were obtained by Le Cam (1956), Johnson (1967, 1970), Walker (1969), Dawid (1970), Brunk and Pierce (1977), Heyde and Johnstone (1979) and Ghosh et al. (1982).

One difficulty with stable measurement is that the sample size which is large enough for the asymptotics to apply will often depend on the observations themselves. Hence, in a sense, one is forced to do a complete posterior robustness check even for large samples.

A(*ii*). *Consistency.* Results concerning the consistency of Bayes estimates (and hence a degree of asymptotic robustness with respect to the prior distribution) can be found in Le Cam (1953), Freedman (1963, 1965), Fabius (1964), Schwartz (1965), Berk (1966, 1970), Strasser (1981), De Robertis and Hartigan (1981) and Diaconis and Freedman (1982). These results tend to say that, if θ is in the support of the prior distribution, then the Bayes estimates are consistent for θ and otherwise they are not. The results of Freedman (1963, 1965) and Diaconis and Freedman (1982) indicate, however, that Bayes estimates can be inconsistent even when θ is in the support of the prior, unless care is taken in the selection of the prior.

A(iii). Sequential analysis. Asymptotic sequential Bayes decision theory is concerned with sequential Bayes decision problems when the cost of each observation is very small. As the cost goes to zero, the number of observations likely to be taken goes to infinity, allowing the large sample Bayesian asymptotics discussed previously to apply. Most of the results on this subject obtain limiting forms of the Bayes stopping rule or Bayes risks. See, for instance, Chernoff (1959), Schwarz (1962, 1968), Kiefer and Sacks (1963), Bickel and Yahav (1967, 1969), Gleser and Kunte (1976), Fortus (1979), Vardi (1979a, 1979b) and Woodroofe (1980). Often, this limiting form is independent of the assumed prior distribution, indicating a large sample robustness. Certain seemingly robust nonasymptotic Bayes stopping rules for estimation problems can be found in Alvo (1977). (See also Berger (1980b) for a general discussion.)

B. Sensitivity theory

Sensitivity analysis is a standard name for the process of investigating changes in the conclusions caused by changes in the initial assumptions (including the prior distributions). Such analysis is present in many good Bayesian papers. Dempster (1976) gives an interesting general discussion of this with examples. Any attempt to mention all such works would be nearly hopeless, so instead only the more formal works concerned with developing bounds on the range of the posterior conclusions based on variation in the assumed prior distributions will be mentioned. (Such works will be called Sensitivity Theory.)

B(i). Bounds on the posterior distributions. There have been many works seeking to bound the amount of variation in the posterior distribution itself (or certain posterior probabilities) for classes Γ of prior distributions, or the closely related 'upper and lower probabilities'. Results for classes of priors can be found in De Groot (1970), Huber (1973), Chamberlain and Leamer (1976), Dickey (1976b), Leamer (1978), Davis (1979), Hill (1980c), Rios and Girón (1980) and De Robertis and Hartigan (1980). (Some of these works are closely related to stable estimation.) Results in Stein (1965) are also relevant.

The idea of 'upper and lower probabilities' is essentially to try and find upper and lower bounds on the prior distributions (these bounds will typically just be finite measures, i.e., will not have mass one) and from these obtain bounds on the posterior distributions. Such ideas can be found in Boole (1854), Koopman (1940), Good (1950, 1962a, 1976), Smith (1961), Dempster (1966, 1967, 1968, 1971), Beran (1970, 1971), Fine (1973), Huber and Strassen (1973), Kyburg (1974, 1976), Kleyle (1975), Suppes (1975), Williams (1976), Suppes and Zanotti (1977), West (1979), Levi (1980), De Robertis and Hartigan (1981) and Wolfenson and Fine (1982), although several of these works propose alternative modes of reasoning based on the upper and lower probabilities.

B(ii). Bounds on posterior actions and expected loss. Sensitivity theory is often concerned with bounding the variation in the optimal posterior action or posterior expected loss caused by variation in the prior. Results for finite parameter spaces can be found in Isaacs (1963), Fishburn (1965), Fishburn, Murphy and Isaacs (1968) and Pierce and Folks (1969). More general theories can be found in Skibinsky and Cote (1963), Dickey (1974, 1976b), Bansal (1978), Kadane and Chuang (1978), Rios and Girón (1980) and De Robertis and Hartigan (1981). Leamer and Polasek (cf. Leamer (1978) and Polasek (1983), which also contain earlier references) give bounds on the posterior Bayes action for a wide variety of problems involving variation of (hierarchical) conjugate priors, an analysis they call 'global sensitivity' analysis. They also discuss 'local sensitivity', which is essentially the rate of change of the posterior Bayes action with respect to change in the parameters of the conjugate prior. Although not generally as useful as global sensitivity, local sensitivity can be of assistance in identifying those prior parameters which have the greatest influence on the conclusion, and hence which must be considered most carefully.

C. Partial prior knowledge

There are a number of results in the literature concerned with determining reasonable posterior actions when only limited facets of the prior distribution are known. For example, Stone (1963), Hartigan (1969) and Goldstein (1974, 1979, 1980) consider estimation problems were knowledge is available concerning only the first two moments of the prior distribution. The estimators that result from such an assumption are linear estimators, and much of the huge literature on linear estimation (including much of linear filtering theory in stochastic processes) can be recast in this light. A serious concern is that prior moments are almost never knowable (see subsection 2.3), and that resulting linear estimators will often not be robust (see also subsection 5.1.3.A).

Other analyses based on limited prior knowledge can be found in Godambe and Thompson (1971), Hill (1975), Leamer (1978), Levi (1980) and Lambert and Duncan (1981).

D. Detecting a lack of posterior robustness

It is particularly important to identify common statistical situations in which posterior robustness is lacking, since such situations call for very careful consideration of prior information.

When the likelihood function is flat the prior distribution will be the main factor in determining the posterior distribution and hence the conclusions are liable to be very sensitive to the prior. This commonly occurs in high dimensional situations where, due to such problems as multicollinearity or often simply a lack of sufficient data for all the parameters of interest, the likelihood function will be flat in certain directions. Among the many discussions of this issue are Hill (1977), Leamer (1978), Hill (1980a), Posasek (1983) and Smith and Campbell (1980). The latter article addresses this problem in a critique of ridge regression, and references a number of other ridge regression papers dealing with the same issue.

Another situation in which the likelihood function is flat is in the random model analysis of variance when the usual unbiased estimator of the between variance component is negative. This is discussed in Hill (1965), Hill (1970) and Hill (1980a).

The value of $m(x)$ (the marginal density of X) can be of use in determining robustness, in that a particularly small value of $m(x)$ indicates that surprising data has occurred; the data and the prior information would seem to be in conflict. In such situations the likelihood function will tend to be concentrated in the tail of the prior distribution, a very uncertain part of the prior. Of course the initial implication of a small value of $m(x)$ is that the situation was incorrectly modeled and hence (prior) assumptions concerning the data model need to be reconsidered or discarded. Excellent discussions of this and other references can be found in Jeffreys (1961), Dempster (1971, 1975), Box and Tiao (1973), Geisser and Eddy (1979), Box (1980) and Good (1965, 1983).

5.1.2. Procedure robustness

A. Asymptotic Bayes risk

One can work with decision problems and Bayes risk $r(\pi, \delta)$ as the sample size goes to infinity. Asymptotic approximations to $r(\pi, \delta)$ are then available. Some work in this direction can be found in Chernoff (1952, 1956, 1970), Lindley (1960), Rubin and Sethuraman (1965), Rubin (1971, 1972), Johnson and Truax (1978), Burnasev (1979), Woodroofe (1980), and Ghosh et al. (1982). Some of the articles mentioned in subsection 5.1.1.A are also of this type.

Interesting robustness phenomenon can occur, when asymptotics are considered, as shown in the following example due to Rubin (1971).

Example 5. Consider the situation of testing a 'fuzzy' point null hypothesis. This concerns the reasonable formulation of the point null testing problem in which the null hypothesis can be phrased as $H_0: \theta \in \Theta_0 = (\theta_0 - \varepsilon, \theta_0 - \varepsilon)$, where ε is quite small. (Rubin (1971) formulates the problem solely in terms of losses, in which case $\theta_0 \pm \varepsilon$ are the points at which the losses in accepting and rejecting are equal.) The prior density $\pi(\theta)$ is assumed to have a sharp peak inside Θ_0 and to be fairly flat away from the peak. (This corresponds to common sense evaluations when a point null is involved.) Relevant also, are the loss functions $L_A(\theta)$, of accepting H_0, and $L_R(\theta)$, of rejecting H_0. The Bayesian will be making a decision based on the weight function

$$W(\theta) = \pi(\theta)[L_A(\theta) - L_R(\theta)],$$

accepting H_0 if

$$\int w(\theta) p_\theta(x)\, \mathrm{d}\theta < 0, \tag{5.1}$$

and rejecting otherwise (since (5.1) implies that the posterior expected loss of accepting is smaller than that of rejecting). The sample size is assumed to be large enough so approximate normality holds, i.e., $p_\theta(x)$ is $\mathrm{N}(\theta, \sigma^2/n)$. Three cases must be distinguished:

(i) Somewhat large n (i.e., $\varepsilon \ll \sigma/\sqrt{n}$). In this situation H_0 can essentially be treated as a point null, in that

$$\int W(\theta)p_\theta(x)\,\mathrm{d}\theta \cong p_{\theta_0}(x)\int_{\Theta_0} W(\theta)\,\mathrm{d}\theta + \int_{\theta\notin\Theta_0} W(\theta)p_\theta(x)\,\mathrm{d}\theta\,. \tag{5.2}$$

The mass of the weight function in Θ_0 is comparatively easy to specify. Also, outside of Θ_0, $W(\theta)$ will be a fairly smooth function and, since n is somewhat large (so that the region with high likelihood is fairly small), it should be possibly to specify the last integral in (5.2) fairly accurately. Thus we have reasonable Bayesian robustness (i.e. the subjective inputs that are needed are fairly easy to elicit.)

(ii) Extremely large n (i.e., $\sigma/\sqrt{n} \ll \varepsilon$). Here it will essentially be known whether $\theta \in \Theta_0$ or not, so the prior will not matter. (Robustness with respect to ε could be a concern, however.) This situation is the usual 'stable measurement' situation.

(iii) Moderately large n (i.e. all n not covered in cases (i) and (ii)). Here, surprisingly enough, robustness is lacking, in that the shape of $W(\theta)$ in Θ_0 is very important. (See Rubin (1971).) This is disturbing, in that determining the shape of the prior in this region is almost impossible. (Of course, the overall risk will be small since n is moderately large, but Rubin (1971) has shown that even mild misspecification of the shape of $W(\theta)$ can cause an increase of Bayes risk of 40% in the most favorable cases, with much larger increases in unfavorable cases.) The phenomenon observed in this example, of robustness for somewhat large n and extremely large n, but not for n in between, is striking.

B. Γ-minimax and Γ-minimax regret procedures

The Γ-minimax and Γ-minimax regret criteria (see subsection 3.2) are natural criteria to follow if procedure robustness is sought. The basic concepts were originally developed in Robbins (1951, 1964) and Good (1952). Other general discussions can be found in Menges (1966), Blum and Rosenblatt (1967), Kudo (1967) and Berger (1980b).

The Γ-minimax regret criterion seems somewhat more reasonable than the Γ-minimax criterion, in that it is based on the loss in risk by not using the theoretically optimal Bayes rule, rather than the absolute Bayes risk. The danger in using $r(\pi, \delta)$ itself is that there could be an 'unfavorable' prior $\pi_0 \in \Gamma$ with excessively large Bayes risk

$$r(\pi_0) = \inf_\delta r(\pi_0, \delta)\,,$$

in which case the Γ-minimax procedure would be the Bayes rule with respect to π_0. Unless there is some reason to be especially concerned with π_0, however, it would be better to eliminate its prominence by using the Γ-minimax regret criterion. The Γ-minimax regret criterion will, on the other hand, be somewhat more difficult to work with, so if Γ contains no 'unfavorable' prior it might be better to consider Γ-minimaxity.

Recall from subsection 2.3 that Γ should generally be specified in terms of percentiles and relative likelihoods. This has been done in the Γ-minimax literature on testing, multiple decision theory and nonparametrics. The literature on estimation, however, makes unfortunate use of Γ specified by prior moments.

Results on Γ-minimax estimation can be found in Jackson, O'Donovan, Zimmer, and Deeley (1970), Solomon (1972a, 1972b), De Rouen and Mitchell (1974), Watson (1974) and Morris (1982). Testing and multiple decision theory results can be found in Rubin (1965, 1971), Randles and Hollander (1971), Gupta and Huang (1975, 1977), Berger (1979), Gupta and Kim (1980), Gupta and Hsiao (1981), Miescke (1981) and Hsiao (1982). Some nonparametric Γ-minimax studies were done in Doksum (1970), Campbell and Hollander (1979) and Lambert and Duncan (1981).

C. Controlled frequentist risk

As discussed in subsection 3.3, the frequentist risk $R(\theta, \delta)$ can be a good indicator of procedure robustness. In particular, if

$$R(\theta, \delta) \leqslant C \tag{5.3}$$

for all θ, then $r(\pi, \delta) \leqslant C$ for all π, giving an upper bound on the possible harm from use of the procedure δ. Theoretical work finding bounds on the frequentist risk of Bayes estimators can be found in Le Cam (1982), which also contains some earlier references. Studies of particular Bayesian estimators which seem to have good frequentist risk have been done in Novick (1969), Strawderman (1971), Lindley and Smith (1972), Efron and Morris (1972, 1973), Clevenson and Zidek (1975), Leonard (1976), Rubin (1977), Faith (1978), Berger (1979, 1980a, 1982a, 1982b, 1982c), Dey (1980), Dey and Berger (1983), Albert (1981), Berliner (1983), Ghosh and Parsian (1981), Hudson and Tsui (1981), Stein (1981), Berger and Wolpert (1983), Bock (1982), Wolpert and Berger (1982) and Zheng (1982).

A more systematic approach to the robustness problem is the restricted risk Bayes approach, initiated by Hodges and Lehmann (1952), which seeks to minimize the Bayes risk $r(\pi_0, \delta)$ for a chosen prior π_0, subject to the constraint (5.3). This guarantees robustness (in a conservative sense) with respect to the class of all priors. Interestingly, as discussed in subsection 4.5, the restricted risk Bayes problem often corresponds to the true Γ-minimax problem with

$$\Gamma = \{\pi \colon \pi(\cdot) = (1-\varepsilon)\pi_0(\cdot) + \varepsilon P(\cdot),\ P \text{ arbitrary}\},$$

where ε, of course, depends on the C in (5.3). Results for the restricted risk Bayes problem can be found in Efron and Morris (1971), Shapiro (1972, 1975), Masreliez and Martin (1977), Bickel (1979), Marazzi (1980) and Berger (1982c, 1982b).

5.1.3. Robust priors

The difficulty of working with a class Γ of priors makes very appealing the idea of finding prior distributions which give Bayes rules which are naturally robust with respect to reasonable misspecification of the prior. Indeed as Huber says in the discussion of Box (1980)

> "Essentially, by now the Bayesian approach should be concerned not with the ad hoc construction of super models but with deriving reliable guide-lines on how to choose the super model (within the inherent arbitrariness) so as to guarantee robustness, and how to do so in a best possible fashion."

A. Conjugate priors are often not robust

Conjugate priors, by definition, have tails of the same type as the tails of the likelihood function; this can cause robustness problems as indicated in subsections 3.2 and 3.3. Priors with tails flatter than the tails of the likelihood function are generally superior (at least for estimation problems). This observation has been made in Anscombe (1963), Tiao and Zellner (1964), Lindley (1968), Dawid (1973), Hill (1974), Dickey (1974), Meeden and Isaacson (1977), Rubin (1977), Umbach (1978), Ramsay and Novick (1980) and Berger (1980a, 1980b). Rubin (1977) gives an excellent numerical study showing the value of choosing flatter tailed priors.

Incidentally, conjugate priors in estimation problems in exponential families tend to result in linear estimators (see Diaconis and Ylvisaker (1979)), indicating a general lack of procedure robustness of linear estimators (except for those arising from noninformative priors). This can be seen directly by examining risk functions of linear estimators.

Of course, a major advantage of conjugate priors is that they are very easy to work with. Hence if posterior robustness is present, it is often appealing to use conjugate priors. If robustness is of concern, yet simple posteriors are desired, an attractive way to proceed in estimation problems is to use a (robust) flat tailed prior, calculate (usually numerically) moments (or maybe percentiles) of the posterior and then match these to a distribution (usually conjugate) of desired simple form. For instance, if $\bar{X} \sim \mathrm{N}(\theta, 1/n)$ is observed, and it is desired to estimate θ, a Cauchy prior will tend to be robust but will result in an ugly posterior. Calculating (numerically) the first two posterior moments and pretending that the posterior is normal with these moments should be reasonably accurate and will result in a posterior which is easy to communicate and use. Uses of this idea can be found in Bakan and Oleksenko (1977), Morris (1977) and Berger (1980b).

B. Noninformative priors

Noninformative priors are designed to be flat and as uninfluential as possible. They tend to work well (if carefully determined) and can hence be considered to provide robust solutions to problems where very little is known a priori. The literature on this subject is vast. Much of it is summarized (and other references are given) in Jeffreys (1961), Zellner (1971), Box and Tiao (1973), Bernardo (1979) and Berger (1980b).

There are problems with the use of noninformative priors, however, principally the arbitrariness in their definition. (Bernardo (1979) seems to have the most workable definition of what he calls reference priors.) Hence even the user of noninformative priors should be concerned with robustness with respect to the class of reasonable noninformative priors. Also, if a noninformative prior is being used as an approximation to a vague proper prior, it is wise to, at least informally, verify that the results obtained are suitable for vague proper priors.

In testing problems, standard noninformative priors cannot be used when they give infinite mass to one of the hypotheses. Such situations can be handled (in a robust fashion) by use of 'reference informative priors' (cf. Jeffreys (1961) and Zellner (1982)).

C. Priors on the boundary of admissibility

While flat-tailed priors tend to be desirable, priors with tails that are too flat may give rise to inadmissible decision rules, especially in higher dimensions. The most important example is estimation of a p-variate ($p \geqslant 3$) normal mean under quadratic loss (although almost any sensible loss gives similar results). The usual estimator (the vector of sample means or the least squares estimator in a linear regression) is the (generalized) Bayes estimator with respect to the (noninformative) uniform generalized prior on R^p. This estimator is inadmissible, because the prior has tails which are too flat.

Much of the recent work in admissibility has been to find the 'boundary of admissibility' in various problems. Priors with tails flatter than those on the 'boundary' will tend to give inadmissible decision rules, while priors with sharper tails will tend to give admissible decision rules. Since flat tails are desirable for robustness, yet inadmissible decision rules are unappealing, priors on this 'boundary' are natural choices for use. Results of this nature can be found in Stein (1965, 1981), Brown (1971, 1979), Strawderman (1971), Strawderman and Cohen (1971), Berger (1976a, 1976b, 1980a, 1982c), Srinivasan (1980), Berliner (1983), Ghosh and Parsian (1981), Berger, Berliner and Zaman (1982) and Hwang (1982a, 1982b).

D. Maximum entropy and reference priors

An appealing idea when faced with a class Γ of possible priors is to choose that prior which maximizes entropy or some measure of loss, or minimizes some measure of information. Such priors are likely to lead to robustness in that they are as noninformative as possible subject to being in Γ and have been called 'minimax information' priors (Good (1968)), 'maximum entropy' priors (Jaynes (1968, 1981) and Rosenkranz (1977)) and 'reference' priors (Bernardo (1979, 1981)).

The most extensively developed such theory is that of maximum entropy priors, much of the development being due to E.T. Jaynes. While I would call the theory highly successful, there are certain difficulties which are cause for concern. First, when Θ is infinite and the partial prior knowledge is (sensibly) the specification of certain percentiles, the maximum entropy prior does not exist. Even when Θ is bounded, the maximum entropy prior in this situation will have unpleasant jumps. Finally, it is not really clear that a maximum entropy prior will be robust. For example, if $\Theta = R^1$ and the first two prior moments are known (an unrealistic assumption of course), then the maximum entropy prior is normal with the given moments. Although this is not terribly unreasonable when the first two prior moments are exactly known, it still seems preferable to use a flatter tailed prior; say a t-distribution with the given moments and a small number of degrees of freedom.

E. Multistage Bayes priors

Multistage (or hierarchical) priors are priors composed of several stages: at stage one the prior is assumed to be of a given functional form (usually the conjugate prior form) with unknown parameters (called hyperparameters); at stage two these parameters are given a prior distribution with possibly unknown hyperparameters; with the process repeating until the final stage (seldom more than the third stage), at which point a completely specified prior distribution (often noninformative) is given to the hyperparameters of the preceding stage. Such priors are particularly useful in multivariate situations where relationships among the parameters are thought to exist and can be modeled in stages. They are also a useful enrichment of the class of conjugate priors when either robustness or more flexibility is sought, in that Bayesian calculations can be done in stages with these priors and will often be relatively easy if the first stage is of a conjugate form.

A multistage prior can, of course, be thought of as a single stage prior; merely integrate out the multistage prior over all hyperparameters. The robustness of the multistage prior follows from the fact that, virtually always, the single stage version has flat tails. If, for example, $\Theta = R^1$ and the first stage prior is $N(\mu, \tau^2)$, putting a prior on τ^2 and integrating will usually result in a flat tailed prior.

The literature on multistage priors is too large to be mentioned here. Good (1952) was the first to extensively discuss the technique and has a very substantial body of work on the subject and its relationship to Bayesian robustness (cf. Good (1980, 1983)). Lindley and Smith (1972) is also an important landmark.

F. Empirical Bayes priors

If $X_1, \ldots, X_n$ are observed and the X_i have distributions depending on θ_i, where the θ_i can be assumed to be generated from a particular prior distribution π_0, then π_0 can itself often be estimated from the data. This is the empirical Bayes idea, first formalized by Robbins (cf. Robbins (1955, 1964)). The approach is particularly easy if π_0 is chosen to be of a known functional form (say the conjugate form) with unknown hyperparameters, and these hyperparameters are estimated from the data. (This is then actually very closely related to the multistage Bayes approach, with similar answers being obtained under either method.) Providing *all* the data is used to estimate the hyperparameters (as opposed to, say, using just 'past data' to estimate the hyperparameters) the resulting prior seems to be quite robust. This is because 'extreme' data (the bane of nonrobust priors) will tend to give hyperparameter estimates leading to flat priors. For more thorough discussion of this see Berger (1980b).

The empirical Bayes literature is also too large to mention. Good discussions and references can be found in Maritz (1970), Berger (1980b) and Morris (1983b).

5.2. Guidelines

The (woefully) few guidelines that have been discussed for achieving Bayesian robustness are summarized here, with a few additional observations.

5.2.1. General considerations

As stressed in subsection 2.3 and elsewhere, it is very important to consider robustness with respect to reasonable classes of priors. Unfortunately easy-to-work-with-classes, such as classes of conjugate priors and classes based on prior moments, are usually unsuitable.

It cannot be overemphasized that if posterior robustness obtains for the data at hand, then the search is ended. This can often best be discovered by simply varying the prior (over Γ) and seeing how the conclusion changes. Increasingly easy to use interactive computer systems should eventually make this relatively easy to do. It will often suffice to merely check posterior robustness for several, fairly different, priors in Γ. For instance, in Example 3 (subsection 2.3), if posterior robustness with respect to the normal and Cauchy priors is present, then posterior robustness with respect to all of Γ probably also obtains. Two useful indicators of a *lack* of posterior robustness are a flat likelihood function (or more commonly a likelihood function which is flat in certain directions of Θ) and a surprisingly small value of $m(x)$.

When posterior robustness is lacking, the situation must be reconsidered. First of all, one naturally looks for experimental causes or modeling failures accounting for this unpleasant situation. If nothing is turned up, further refinement of Γ is called for. If the limit of the elicitation process has been reached however, then now, and only now, does procedure robustness and the possible use of frequency concepts (see Section 4.4) come into play. (Of course, if one is developing procedures for automatic use by nonsophisticated users, then posterior robustness is relevant from the start. To many this may be deemed to be a major purpose of the theoretical statistician.) One could formally attempt some type of Γ-minimax or Γ-minimax regret analysis but this will tend to prove enormously difficult. Indications of a lack of posterior robustness can be obtained from frequentist measures of the performance of a procedure; if the frequentist measure looks bad for certain Γ which are not completely implausible, concern is indicated.

A natural Bayesian attempt to obtain procedure robustness would be to put a metaprior on Γ itself. Since we are assuming that the elicitation process has ended, this would be merely a technical device to hopefully achieve robustness. Experience indicates that this probably works reasonably well, although it is difficult to do. (For one example, see Dickey and Freeman (1975).) It will usually be nearly impossible to construct a reasonable metaprior with support equal to all of Γ, so careful selection of a representative subset of Γ on which to place the metaprior would be needed. Note that this 'two-stage' prior could be written as a one-stage prior and hence the technique can be interpreted as simply a way of constructing hopefully robust priors.

Due to the difficulties of formally working with Γ for procedure robustness, it may simply be best to investigate the robustness of a procedure with respect to a few carefully chosen disparate priors in Γ.

The material on 'robust priors' in the preceding subsection will not be repeated here, although it is certainly relevant to general guidelines.

5.2.2. Guidelines for particular types of problems

The following essentially obvious comments are not too much better than nothing but may sometimes prove helpful.

A. Estimation

Posterior robustness will typically be obtained when the likelihood function is concentrated in the 'central' portion of the prior. (This 'center' will usually be similar for all π in Γ.) When this is not the case, flat tailed priors will at least give procedure robustness. Note that in multivariate estimation problems it will often be the case that the robustness situation is very different for different coordinates of θ.

B. Testing

In testing problems the tail of the prior will usually be unimportant (in contrast to the estimation situation), in that if the likelihood function is concentrated in the tail of the prior there is usually very strong evidence for a particular hypothesis. This robustness with respect to the tail of the prior is very pleasant. Note however that the posterior odds of the hypotheses can be drastically affected by the tail of the prior (as pointed out by Savage et al. (1962)), so Bayesian measures of the strength of the conclusion are not necessarily robust.

Conclusions in testing problems will, naturally, be frequently sensitive to the prior mass given each hypothesis. This is unavoidable and, to a Bayesian, completely sensible.

C. Design of experiments and sequential analysis

Optimal Bayesian designs are usually robust with respect to small changes in the prior such as changes in the tail. At least this is true when overall average measures of performance (say Bayes risk) are deemed relevant, since such averages are dominated by the contributions from the 'likely' θ, or alternatively the 'likely' x. (Of course, after taking the data and being faced with the need to draw some conclusion, robustness may have to be completely reevaluated.)

Note that, in design, there may be real technical advantages in working with frequentist measures averaged over the prior rather than posterior measures averaged over the marginal distribution of X (which is more instinctively appealing to a Bayesian). This is because the (decision) procedure to be used may be fairly accurately known (say, when the sample size will be moderately large) so that involving the (uncertain) prior only at the last stage can lead to a technically easier robustness analysis.

Sequential analysis is, in a sense, just a design problem, in that the real difficulty is deciding, at a given stage, whether to cease sampling or to continue taking observations. This problem should again be relatively immune to such things as the tail of the posterior distribution (upon which the decision to stop or not is based). Of course, if the likelihood becomes concentrated in the tail of the original prior, then this tail can become relevant through its effect on the posterior. In a very practical sense there may be little problem with robustness in sequential Bayesian analysis since it will often be the case that one simply continues sampling until enough information has been accumulated so that posterior robustness obtains.

5.2.3. Actual practice

In many realistic statistical situations involving complicated Θ, any type of Bayesian approach becomes very difficult. Also the uncertainties in specifying a prior for such Θ are very acute, meaning that Bayesian robustness becomes a very real concern. Unfortunately, one quickly encounters an instance of 'Type II rationality' (cf. Good (1973)) in that if straightforward Bayesian analysis is difficult, then a robust Bayesian analysis might be next to impossible. Type II rationality simply says, in this situation, that if you cannot trust a single prior Bayes analysis and Bayesian robustness results are unavailable, then it is permissible to use some type of non-Bayesian analysis, providing it is deemed to be the lesser evil. In other words, if the dangers of a Bayesian analysis with an ill-specified prior seem large (and cannot be eliminated by robustness considerations) and if an easier non-Bayesian or partially non-Bayesian analysis seems sensible (see subsection (4.4)), then go ahead and abandon ship.

The need to compromise the 'purist' robust Bayesian position was already encountered in subsection 2.4, where post data modification of Γ was discussed. (Of course, allowing such modification somewhat alleviates the current problem since all prior knowledge concerning a complicated Θ need not then be exactingly quantified prior to experimentation.) This compromise was still within the general Bayesian framework, however.

A more significant departure from the usual Bayesian framework. occurs when it is necessary to ignore data. Such situations surely abound in statistics. Survey sampling provides one such situation, in that all sorts of data may be available about the sample, most of which may seem irrelevant to the attribute of interest. Constructing a general Bayesian (superpopulation) model for all the data would be very difficult and, perhaps, would not be trustworthy. The same issue arises in the use of randomization as discussed in subsection 4.2. Hildreth (1963), Pratt (1965, with his discussion on 'insufficient statistics'), Dempster (1968), Hill (1975, 1980a, 1980c), and Good (1976 and earlier with his 'Statistician's Stooge') also contain useful discussions on this issue.

Ignoring data causes no real problem to a Bayesian if the data seems unlikely to have an effect on the posterior distribution of the parameters of interest. Often, of course, this can only be ascertained through, at least informal, Bayesian reasoning. Consider the following examples.

Example 6. (Fraser and Mackay (1976)). Suppose independent observations $X_1, \ldots, X_n$ from an $\mathrm{N}(\mu, \sigma^2)$ distribution are observed, where it is desired to estimate μ but σ^2 is also unknown. Independent observations $Y_1, \ldots, Y_m$ are also available, where Y_i is $\mathrm{N}(\mu_i, \sigma^2)$, μ_i unknown, $i = 1, \ldots, m$. If virtually nothing is known a priori about the μ_i (and they are in no way related to μ), it is certainly reasonable to ignore the Y_i when estimating μ. (A formal Bayesian analysis would certainly show that the Y_i had almost negligible influence on the posterior distribution of μ.)

Example 7. In a medical trial comparing two surgical techniques a significant relationship was found between the time of the day in which the surgery was performed and the success of the surgery. Suppose the relationship was one of the following: (i) the later in the day the surgery occurred the less successful it was; (ii) when surgery began on even hours it was more successful than when it began on odd hours; (iii) when surgery ended on an even minute it was more successful than when it ended on an odd minute. The question before us is: can we ignore the data 'time of day'? The answer in case (i) is almost certainly no and we better hope that the two treatment groups were not unbalanced concerning this covariate. The answer in case (iii) is almost certainly yes; it is hard to believe that this relationship is anything more than a coincidence. In case (ii) the answer is not so certain and indeed some investigation is called for. (Did certain surgeons work at certain times, etc.?)

The decision about ignoring data in Example 7 clearly involves prior opinions. The point, however, is that it *may* be possible to informally reason that certain data can be ignored without having to go through a full blown Bayesian analysis. This is not really a violation of Bayesian principles either since the posteriors obtained by ignoring part of the data are felt to be the same as what would have been obtained by a sound Bayesian analysis with all of the data.

The real difficulty arises when it is necessary to throw away potentially relevant data. The reason for doing this would be an inability to carry out a (robust) Bayesian analysis involving everything. Hill (1975) considers a nonparametric problem of this nature in which a trustworthy complete Bayesian analysis seems almost impossible. Hill says

> "When such a formal analysis simply cannot be made, or even when it is merely very difficult and of dubious validity, then there is little choice but condition on that part of the data that can be effectively dealt with, and rely upon some form of stable estimation argument."

The last part of the comment can be interpreted to mean that if you must ignore data, at least convince yourself that there is no reason to expect it to have a large effect on the posterior (or, more properly, the final conclusion). This point is extensively discussed in Pratt (1965) and also Dempster (1975) which has interesting examples and other references.

The above discussion is not to be interpreted as advocating the frequently encountered viewpoint of 'using whatever approach works well for a given problem'. Indeed the major point in this article is that the only way to ensure that a conclusion being reached is sensible is to verify that it is sensible from a posterior robust Bayesian viewpoint. But if a robust Bayesian analysis is not implementable, then compromises must be made. The robust Bayesian makes this compromise only when he has to, however, and only to the extent necessitated by technical limitations.

Perhaps the most important practical advice the robust Bayesian has to offer is 'think like a robust Bayesian'. (In the same way it has been argued that the most important thing to learn from decision theory is simply the ability to think decision theoretically.) Merely thinking of problems from this perspective, without even doing a formal analysis, will frequently illuminate the truth. Once the truth (or the direction in which it lies) has been discerned, a method of analysis can undoubtedly be found which is acceptable to the relevant audience and leads to this truth.

6. Final comments

In an (obviously unsuccessful) effort to keep from meandering from the central argument, a number of side issues have been deferred to this final section. These include a discussion of various criticisms that can be raised against the robust Bayesian viewpoint (by non-Bayesians, Bayesians, and foundationalists) and a very brief discussion of needed theoretical developments. Many of the criticisms are founded on very deep issues so all that can be done here is to give a superficial view of the arguments and counter arguments.

6.1. Non-Bayesian criticisms

The primary non-Bayesian objection to the robust Bayesian viewpoint is, of course, that Assumption I is wrong. Since an extensive justification of Assumption I was not attempted here, this objection will not be pursued except for one brief comment. Much of the philosophical difference in attitude between Bayesians and non-Bayesians seems to be due to Bayesians being optimistic about the existence of truth and pessimistic about the use of intuition while non-Bayesians are just the opposite. The Bayesian feels there is (at least theoretically) a single correct way of doing things, not many correct ways. Also the Bayesian (and the decision theorist) do not trust intuition to properly combine and relate all relevant factors of a problem to arrive at a conclusion.

Perhaps the most common non-Bayesian objection to anything Bayesian is the Bayesian's lack of objectivity. Bayesian rebuttals range from the subjectivist opinion that 'objectivity' is a myth to the objective Bayesian assertions that objectivity can only be attained if conciously sought from a Bayesian perspective. The former viewpoint is reflected by the following quotation from Good (1973).

> "The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science."

The objective Bayesian viewpoint is that the only way to avoid 'biasing' the analysis is to do a Bayesian analysis with a noninformative prior distribution (see subsection 5.1.3.B for references). Strong support for this view can be obtained from 'Reason (ii)' in subsection 2.1. If a supposedly objective non-Bayesian procedure actually corresponds to a Bayesian procedure for a very biased prior distribution, the claim of objectivity seems somewhat silly. The vehement condemnation of the use of noninformative priors by some non-Bayesians is indeed somewhat mystifying since subjective prior beliefs are not being incorporated. Of course, there are problems in finding and using noninformative priors, but I have seen no better, easier to use and less error prone technique for deriving reasonable objective procedures. (Although when being a purist I would argue against the possibility of objectivity, for a variety of robust Bayesian and Type II rationality reasons the noninformative prior approach seems extremely valuable.)

It should be mentioned that there are several different non-Bayesian theories that reject Assumption I. Besides the various classical theories, these include the fiducial inference of R.A. Fisher (see Wilkinson (1977) for an up-to-date version), the structural inference of D.A.S. Fraser (see Fraser and Mackay (1976) and Fraser (1979) for references), pivotal inference (cf. Barnard (1981)) and the theory of belief functions (cf. Shafer (1982)). Since we are foregoing a serious effort to justify Assumption I, these alternative theories will not be discussed.

6.2. Bayesian criticisms

Many natural Bayesian objections to the viewpoint expressed in this paper, such as the violation of the Likelihood Principle (and to an extent Assumption I) by procedure robustness, have been discussed elsewhere. Several other criticisms can be raised, however. Three are discussed in this subsection.

A. "Just report what the data says."

A very admirable Bayesian desire is to provide a mechanism by which the data can be easily assimilated to allow others to reach a conclusion. The likelihood function of θ, $l_x(\theta)$, is the most basic such mechanism, since anyone can determine his own posterior for θ by simply multiplying $l_x(\theta)$ by his prior (or priors) and normalizing. Thus reporting the likelihood function is definitely reasonable (cf. Box and Tiao (1973)). A similar idea (cf. Bernardo (1979)) would be just to report a noninformative or reference *posterior*, since this will be more meaningful intuitively than $l_x(\theta)$ and anyone can easily still determine his own posterior. Considerable effort has also been spent on finding easier to digest data communication vehicles such as Bayes factors between hypotheses (cf. Dickey (1973, 1974)). A criticism of the robust Bayesian position is that, if the above pursuit is the true job of the statistician, then he need not be concerned with robustness, which afterall only becomes of concern when data-prior interactions are being studied.

I must have stated the criticism unfairly since it seems clearly unworthy. We cannot, after all, abandon the user at the critical point of combining the data with his prior information, particularly when some action or conclusion must be taken. Also it is frequently impossible to even separate data from prior information in a useful way. For example, the usual likelihood function is very model dependent but the model is often unknown and should be considered part of the parameter.

B. "Why single out the prior? Model robustness is just as serious a problem."

First of all, since the data model was allowed to be part of Θ, we did not really ignore model robustness. On the other hand, there *are* sometimes reasons to be more concerned about the parameters than the model. For example, the model may have some theoretical basis while prior opinions about parameters of the model might be much more subjective. Of course, there are many problems in which the reverse is true, where the choice of a model is somewhat arbitrary and will have a much more profound effect on the answer than the choice of a prior on the parameters of the model. Nevertheless, the prevalent statistical attitude is to trust models more than priors and in dealing with this attitude the robust Bayesian viewpoint can be very helpful. Also, even when considerable uncertainty about the model exists, it may cause less of a problem than uncertainty about the prior information, as the following example indicates.

Example 8. Suppose $X_1, \ldots, X_n$ is an independent sample from a location density $f(x-\theta)$ on R^1 where $f(z)$ is symmetric and unimodal. It is deemed reasonable to model f as a t-distribution with quartiles $(\theta \pm 1)$ but specification of the degrees of freedom, α, is judged to be impossible. Prior elicitation reveals that θ is thought to have median zero and quartiles ± 1, with the prior having a symmetric unimodal density. It is desired to estimate θ. Although the model and prior uncertainties seem similar here, the likelihood function will be

$$l_x(\theta) = \prod_{i=1}^{p} f(x_i - \theta),$$

which, even for moderate n, will most likely have sharper tails than the prior. (The tail of $l_x(\theta)$ will be like the nth power of the tail of f.) This indicates that the robustness problem with respect to f will be less serious than that with respect to the prior.

C. "*Robustness is a rare problem and can be dealt with entirely within the Bayesian framework.*"

These issues have been discussed throughout the paper. I have argued that posterior robustness will be lacking in a significant portion of our problems (at least at the present stage of Bayesian development) and that techniques of proceeding, which at least partly lie outside the pure Bayesian framework, can prove useful. Some people may argue that the non-Bayesian components of the robust Bayesian viewpoint will seldom be needed while others will argue that it is these non-Bayesian components which will be of most use. This is exactly what we should be arguing about: what is the best method to achieve the robust Bayesian goal. (See also Berger (1982d).)

From a very pragmatic viewpoint also, the pure Bayesian position strikes me as unwise. The only truly overwhelming problem facing Bayesians is that of convincing non-Bayesians that the Bayesian viewpoint is correct. The major stumbling block in the entire controversy is that Bayesians (as a whole, not individually) have not openly admitted the validity of Assumption II and been willing to accept its consequences. This allows the non-Bayesian to refuse to think about Assumption I, because he feels certain that Assumption II is correct and hence that the Bayesians must be wrong. Again I am talking about the overall image of Bayesian and not necessarily about the viewpoints of particular Bayesians. Even accepting Assumption II but staying within the purely Bayesian framework of posterior robustness will not provide a general enough structure to satisfy many of the criticisms of non-Bayesians. The practicing Bayesian might find that he seldom needs to leave the pure Bayesian structure, and hence that procedure robustness, etc. are concepts only rarely needed, but having them available can never hurt and can, I believe, help substantially in promoting the cause.

6.3. Foundational criticisms

While non-Bayesians attack Assumption I from above (loftily disdaining from grubbing around in subjective probabilities), certain foundationalists attack it from below (urging even deeper submersion in subjectivism). The issue is whether reasoning in terms of a class Γ of (countably additive) prior probability distributions with updating by conditioning (and post data modification of Γ) suffices, or whether more general or more basic concepts are needed. I will basically argue that the above concepts not only suffice but are what we should train ourselves to think in terms of. The robust Bayesian viewpoint is not an attempt to model how intuition works but rather an attempt to create a structure of components which are simple enough to be accessible to intuition and which when combined give the truth. As Good (1976) says

> 'The main merit that I claim for the Doogian philosophy is that it codifies and exemplifies an adequately complete and simple theory of rationality, complete in the sense that it is I believe not subject to the criticisms that are usually directed at other forms of Bayesianism, and simple in the sense that it attains realism with a minimum of machinery."

My rebuttal to the foundational criticisms will thus tend to be that the alternative structures proposed either have components which are not reasonably accessible to intuition or have unnecessarily complicated structures. I, of course, admit that it may be personal taste rather than sound reasoning which leads me to reject these alternative theories. Also, through lack of careful enough study or just general thick-headedness, I may misrepresent certain arguments or be foolish in my response to them, in which case I apologize and look forward to being set straight. In any case, besides being somewhat fun, these foundational issues serve well to illuminate the edges of the robust Bayesian theory.

A. Measurability criticisms

De Finetti (1974, 1975) and Good (1962a) argue that sometimes it is inappropriate to stay within the confines of measurable events. The real concern here is that the data may cause one to desire an enlargement of the σ-field (of measurable events in Θ) that was originally chosen as adequate. This could be subsumed under post-data modification of Γ, of course. In any case technical measurability concerns are certainly not particularly relevant to the validity of the robust Bayesian viewpoint.

At the other extreme Manski (1981) and Lambert and Duncan (1981) argue that, in specifying a prior, the σ-field of measureable events should be restricted to those events about which prior information is to be elicited. This is somewhat appealing intuitively since a single measurable prior with respect to this σ-field would correspond to a class Γ of priors in the usual setup with, say, the Borel σ-field. The difficulties with this approach are that (i) it is very hard to update σ-fields based on the data, surely an essential ingredient of the approach; and (ii) it will generally be much more revealing to investigate robustness by varying π over Γ than to simply make conclusions within the restricted σ-field formulation. The point is that by using the robust Bayesian framework one is often alerted as to what features of the prior need special consideration. Restricted σ-fields may hide these facets of the prior. To some extent my view may be based on simply feeling comfortable with usual probability distributions and hence development of this alternate approach is of interest, but I would be surprised if it led to a more useful framework.

B. "*Finitely additive priors should be allowed.*"

Among Bayesians there is a considerable faction that believes that insisting on countably additive priors is too restrictive in that such conceivably desirable priors as proper 'uniform' priors on R^1 or on the integers are prohibited. De Finetti has long argued this (cf. de Finetti (1972, 1974, 1975)). Other persuasive cases have been made by Dubins (1975), Heath and Sudderth (1978), Kadane, Schervish and Seidenfeld (1981) and Hill (1980b). The case for allowing finitely additive priors also rests on the fact that the 'rationality' or 'coherency' justifications of Bayesian analysis lead only to finitely additive priors, although under slightly stronger axioms countable additivity emerges (cf. Savage (1954), De Groot (1970), Spielman (1977) and Kadane, Schervish and Seidenfeld (1981)).

The arguments for staying within the countably additive framework are that (i) the examples espousing a need for finite additivity are not really convincing; (ii) even if 'uniform' type priors on unbounded spaces are needed, countably additive improper priors can be used; and (iii) finitely additive priors require extremely careful handling.

The first point is that, while convincing 'thought' examples have been constructed of the need for finite additivity, I have not yet seen a real example, involving an actual real world action that must be taken, in which my prior opinions would be uniform on an unbounded set. I certainly admit, however, that situations will exist in which I might want to use a 'noninformative' prior, either as a robust approximation to my true prior beliefs or in a situation in which the appearance of objectivity is deemed necessary. Such situations can be dealt with either by using improper countably additive 'noninformative' priors or by using proper finitely additive noninformative priors. Each approach has certain theoretical drawbacks, which will not be discussed here. They both also have the practical drawback of there being no clearcut definition of noninformative priors and a careless choice can given unsatisfactory results. (The situation for finitely additive priors is particularly bizarre: for instance, there are $2^{2^{\aleph_0}}$ different 'diffuse' finitely additive priors on the positive integers as opposed to the single (constant) countably additive diffuse (improper) prior.)

The main difference, to me, lies in the ease with which they can be used. Finitely additive priors frequently fail to have the Randon–Nikodym property (that conditional probabilities on sets of measure zero can be uniquely defined as limits of conditional probabilities of sets of nonzero measure) and hence frequently do not have well defined conditional (posterior) distributions. (Heath and Sudderth (1978) discuss some common statistical situations involving amenable group structures where meaningful posterior distributions can be defined. See also Dubins (1969).) This makes the typical Bayesian conditional analysis very difficult or impossible in general. Also, unconditional Bayesian analysis can seem very silly if finitely additive priors are used, as the following example shows.

Example 9. It is desired to estimate, under squared error loss, a normal mean θ based on $\bar{X} \sim \mathrm{N}(\theta, 1/n)$. If one were to be repeatedly faced with this problem (with different θ_i) then it might seem reasonable to ask for a procedure with good average Bayes risk $r(\pi, \delta)$. If nothing was felt known about the θ_i, one might be tempted to use a noninformative prior. Use of the improper countably additive uniform prior will give infinite risks, so one is alerted to approach the problem differently. The usual finitely additive uniform prior, however, has finite Bayes risk equal to $1/n$, but there are many estimators which achieve this, among them $\delta^0(\bar{x}) = \bar{x}$ and $\delta^*(\bar{x}) = \bar{x} - 10^{1000}/\bar{x}$. A posterior analysis of the problem (which can be done here using Heath and Sudderth (1978)) shows that δ^0, not δ^*, is correct, but the need for careful unconditional handling of finitely additive priors is, at least, indicated.

It should be noted that there is no foundational reason not to allow finitely additive priors into the robust Bayesian framework so that those who feel comfortable with them are invited to do so.

C. "The use of probability distributions is too restrictive."

The first point, made initially by Kraft, Pratt and Seidenberg (1959) (see also Fine (1973)) is that there may exist 'likelihood orderings' of events that are internally consistent and yet which are not consistent with *any* probability distribution. Although unsettled by this fact, I would argue that it is irrelevant, in that I would myself heavily distrust any likelihood ordering not consistent with some probability distribution. The consistent modes of behavior are those induced by probability distributions, so I would rather take them as my 'primitives' than I would a concept such as 'likelihood orderings'. This is another situation in which I am not concerned with modeling how the mind could work, but rather with developing *a* framework within which the mind can successfully work.

Many foundational theories have been proposed which are based on generalization of probability distributions. Various such attempts can be found in Koopman (1940), Good (1950, 1962a, 1976), Smith (1961), Dempster (1966, 1967, 1968, 1971), Jeffrey (1968), Beran (1972), Huber and Strassen (1973), Fine

(1973), Kyburg (1974), Suppes (1975), Suppes and Zanotti (1977), Levi (1980), De Robertis and Hartigan (1981), Wolfenson and Fine (1982) and Rios and Girón (1980). (Some of these deviate only slightly from the robust Bayesian approach and hence are not really susceptible to the following criticisms.)

A starting point for several of these theories is a rather ill-considered criticism of prior probabilities. They often begin with a 'counterexample' such as the following.

Example 10. Suppose you pull a coin from your pocket and, without looking at it, are interested in the event A that it will come up heads when flipped. Suppose you (reasonably) judge the subjective probability of this event to be close to $\frac{1}{2}$. Next, you contemplate an experiment in which two drugs, about which you know nothing, will be tested, and are interested in the event B that Drug 1 is better than Drug 2. You (reasonably) judge your subjective probability of event B to be $\frac{1}{2}$ also. The argument now proceeds:

"Even though both probabilities were $\frac{1}{2}$, you have a stronger 'belief' in the probability specified for event A, in that if you were told that five flips of the coin were all heads your opinion about the fairness of the coin would probably change very little, while if you were told that in tests on five patients Drug 1 worked better than Drug 2 you would probably change your opinion substantially about the worth of Drug 1." Thus, the argument goes, it is necessary to go beyond probability distributions and have measures of the 'strength of belief' in probabilities.

It is easy to see the flaw in this reasoning. Before getting any data I *would be* equally secure in probabilities of $\frac{1}{2}$ for each A and B, in that I would be indifferent between placing a single bet on either event. My knowledge about the *events* A and B is well described by a probability of $\frac{1}{2}$. However, my knowledge about the overall phenomena being investigated in each case is quite different. A description of my overall knowledge about the situations is more fully described by defining the unknown (and fictitious to a true subjective Bayesian) quantities p_C and p_M, reflecting the 'true' proportion of heads and 'true' proportion of patients for which Drug 1 would work better than Drug 2, respectively, and then quantifying prior distributions (or classes thereof) for p_C and p_M. The prior distributions for p_C will undoubtedly be much more tightly concentrated about $\frac{1}{2}$, than will the prior distributions for p_M. Note that the subjective probabilities of events A and B are just the means of the respective prior distributions. (I first saw an analysis of this common misconception done by D. Lindley, though I cannot recall the reference.)

Thus prior distributions prove to be rich enough to reflect whatever is reasonably desired. Even more interesting is the observation that, in taking account of experimental evidence, one is almost forced to think in the correct fashion. Thus, in Example 10, if at the beginning it was only felt necessary to quantify the probabilities of A and B, reflection on the experiment to be performed reveals that the data information can be combined with prior information via Bayes theorem only if prior information is specified in terms of quantities such as p_C and p_M.

A second, more substantial, reason that alternative theories to Bayesian analysis have been developed is the recognition of the validity of Assumption II, and the *perception* that Bayesian analysis could not incorporate this assumption (although there were numerous works, about 50 by I.J. Good alone I believe, indicating that Assumption II could be incorporated). Some of the approaches do suggest alternate methods of dealing with probabilistic uncertainty, such as using lower and upper probabilities. The robust Bayesian approach seems much more straightforward, however, and does not demand the introduction of all sorts of new and supposedly 'intuitive' criteria. Indeed I have seen no new criterion that is obviously trustworthy and the very same reasoning that forced me to accept the Bayesian viewpoint, as opposed to the 'intuitive' classical viewpoint, argues against the existence of any such other criterion. This is a mild echo of E.T. Jaynes (1976), who said

> "It doesn't matter how many new words you drag into the discussion to avoid having to utter the word 'probability' in a sense different from frequency: likelihood, confidence, significance, propensity, support, credibility, acceptability, indiffidence, consonance, tenability, and so on, until the resources of the good Dr. Roget are exhausted .... It doesn't matter what approach you happen to like philosophically—by the time you have made your methods fully consistent, you will be forced, kicking and screaming, back to the ones given by Laplace."

(Author's note: Laplace argued for noninformative prior Bayesian analysis. We have, of course, allowed ourselves proper subjective priors also, but the following of Assumption I is the most important part of Laplace's methods.)

D. Updating Γ

The fourth reason sometimes proposed for broadening Bayesian analysis is the clear need to sometimes update the prior information by means other than Bayes theorem. This problem was discussed in subsection 2.4.

E. Conclusions

A reading of the above suggests that the espoused robust Bayesian viewpoint was constructed by starting with pure Bayesian analysis and modifying it to handle every meaningful objection raised. This is exactly right. Assumption I is the cornerstone and provides the starting point for the theory. At every stage where additional flexibility was needed it was allowed into the theory, but in a way which minimized the resulting deviation from Assumption I. Any attempt to modify the theory, not satisfying this 'minimum distance from Assumption I' constraint, is unlikely to prove successful.

6.4. Future development

I agree with Dempster (1976) that

> "The ultimate goal of research on Bayesian robustness should be to

> classify applied situations so that a plausible prepackaged robustness analysis within each class will be available. I believe that only the faintest beginnings have been made on this task."

This would enable users to investigate robustness themselves, surely the most desirable goal.

Because it will probably always be the case that many (most?) users of statistics will not have the skill or the inclination to do such analyses however, it behooves researchers to find specific robust Bayesian procedures or families of robust prior distributions (to use in place of conjugate families where warranted) for important situations. Again, relatively little has been done in this area.

Alerting users to situations *lacking* robustness is also very important. They can then know when, and on what, it is necessary to concentrate their prior elicitation.

As a concluding comment, note that a common criticism of Bayesian analysis is that it is too automatic. Thus Kiefer (1977) states that

> "...statistics is too complex to be codified in terms of a simple prescription that is a panacea for all settings..."

As we have seen, robust Bayesian analysis offers no single prescription and instead urges flexibility in thought and methods. It demands only that the proper goal be kept in mind.

Acknowledgements

The list of people who have contributed to my views on this subject would be too large to include. During the writing of the paper I received useful ideas and had helpful discussions with Leon Gleser, Prem Goel, Bruce Hill, Ker-Chau Li, Prem Puri, Herman Rubin, Arif Zaman and participants in a seminar on Foundations of Statistics held at Purdue University. My thanks to all these people. Herman Rubin deserves special mention as my source for clear thinking whenever I get confused.

A preliminary version of the paper was read by José Bernardo, Lawrence Brown, Morris De Groot, Arthur Dempster, Persi Diaconis, Joseph Kadane, Isaac Levi, Dennis Lindley and Melvin Novick, all of whom had many useful comments which greatly improved the paper. I am particularly grateful to Morris De Groot for a very extensive set of insightful comments.

References

[1] ALBERT, J.H. (1981) Simultaneous estimation of Poisson means. *J. Multivariate Anal.* **11**, 400–417.

[2] ALVO, M. (1977) Bayesian sequential estimates. *Ann. Statist.* **5**, 955–968.

[3] ANSCOMBE, F.J. (1963) Bayesian inference concerning many parameters with reference to supersaturated designs. *Bull. Int. Statist. Inst.* **40**, 741–733.

[4] BAKAN, G.J. and OLENSENKO, O.M. (1977) Nonlinear estimation by approximating the a posteriori density by a normal distribution. *Soviet Automat. Control* **10**, 6–10.

[5] BANSAL, A.K. (1978) Robustness of a Bayes estimator for the mean of a normal population with nonnormal prior. *Comm. Statist. A—Theory Methods* **7**, 453–460.

[6] BARNARD, G. (1982) A coherent view of statistical inference. To be published in the proceedings of the statistical symposium held at Waterloo in 1981.

[7] BASU, D. (1971) An essay on the logical foundations of survey sampling. In: V.P. Godambe and D.A. Sprott, eds., *Foundations of Statistical Inference*, Holt, Rinehart and Winston, Toronto.

[8] BASU, D. (1975) Statistical information and likelihood (with discussion). *Sankhyā Ser. A* **37**, 1–71.

[9] BASU, D. (1978) On the relevance of randomization in data analysis (with discussion). In: N.K. Namboodiri, ed., *Survey Sampling and Measurement*, Academic Press, New York.

[10] BASU, D. (1980) Randomization analysis of experimental data: the Fisher randomization test. *J. Amer. Statist. Assoc.* **75**, 575–595.

[11] BERAN, R.J. (1970) Upper and lower risks and minimax procedures. *Sixth Berkeley Symp. Math. Statist. Prob.* **1**, 1–16, University of California Press, Berkeley.

[12] BERAN, R.J. (1971) On distribution-free statistical inference with upper and lower probabilities. *Ann. Math. Statist.* **42**, 157–168.

[13] BERGER, J. (1976a) Inadmissibility results for generalized Bayes estimators of coordinates of a location vector. *Ann. Statist.* **4**, 302–333.

[14] BERGER, J. (1976b) Admissibility results for generalized Bayes estimators of coordinates of a location vector. *Ann. Statist.* **4**, 334–356.

[15] BERGER, J. (1979) Multivariate estimation with nonsymmetric loss functions. In: J.S. Rustagi, ed., *Optimizing Methods in Statistics*, Academic Press, New York.

[16] BERGER, J. (1980a) A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8**, 716–761.

[17] BERGER, J. (1980b) *Statistical Decision Theory: Foundations, Concepts and Methods.* Springer, New York.

[18] BERGER, J. (1982a) Selecting a minimax estimator of a multivariate normal mean. *Ann. Statist.* **10**, 81–92.

[19] BERGER, J. (1982b) Bayesian robustness and the Stein effect. *J. Amer. Statist. Assoc.* **77**, 358–368.

[20] BERGER, J. (1982c) Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III*, Academic Press, New York.

[21] BERGER, J. (1982d) Bayesian salesmanship. To appear in: P.K. Goel and A. Zellner, eds., Bayesian Inference and Decision Techniques with Applications, North-Holland, Amsterdam.

[22] BERGER, J., BERLINER, L.M. and ZAMAN, A. (1982) General admissibility and inadmissibility results for estimation in a control problem. *Ann. Statist.* **10**, 838–856.

[23] BERGER, J. and WOLPERT, R. (1982) The Likelihood Principle: a review and generalizations. Technical Report #82-33, Department of Statistics, Purdue University, West Lafayette, IN.

[24] BERGER, J. and WOLPERT, R. (1983) Estimating the mean function of a Gaussian process and the Stein effect. *J. Multivariate Analysis* **13**.

[25] BERGER, R. (1979) Gamma minimax robustness of Bayes rules. *Comm. Statist.* **8**, 543–560.

[26] BERK, R. (1966) Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37**, 51–58.

[27] BERK, R. (1970) Consistency a posteriori. *Ann. Math. Statist.* **41**, 894–906.

[28] BERLINER, L.M. (1983) Improving on inadmissible estimators in the control problem. *Ann. Statist.* **11**, 814–826.

[29] BERNARDO, J.M. (1979) Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **41**, 113–147.

[30] BERNARDO, J.M. (1981) Reference decisions. *Symposia Mathematica* **25**, 85–94.

[31] BICKEL, P.J. (1979) Minimax estimation of the mean of a normal distribution subject to doing well at a point. Technical Report, Dept. of Statistics, Univ. of California at Berkeley.

[32] BICKEL, P.J. and YAHAV, J.A. (1967) Asymptotically pointwise optimal procedures in sequential analysis. In: *Proc. 5th Berkeley Symp. Math. Statist. Prob.* **1**, 401–413, Univ. of California Press, Berkeley.

[33] BICKEL, P.J. and YAHAV, J.A. (1969) Some contributions to the asymptotic theory of Bayes solutions. *Z. Warsch. verw. Gebiete* **11**, 257–276.

[34] BIRNBAUM, A. (1962) On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* **57**, 269–326.

[35] BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.H. (1975) *Discrete Multivariate Analysis: Theory and Practice.* M.I.T. Press, Cambridge.

[36] BLACKWELL, D. and DUBINS, L. (1962) Merging of opinions with increasing information. *Ann. Math. Statist.* **33**, 882–886.

[37] BLUM, J.R. and ROSENBLATT, J. (1967) On partial a priori information in statistical inference. *Ann. Math. Statist.* **38**, 1671–1678.

[38] BOCK, M.E. (1982) Employing vague inequality prior information in estimation of a normal mean vector. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III,* Academic Press, New York.

[39] BOOLE, G. (1854) *An Investigation of the Laws of Thought.* Reprinted by Dover (1958). (Chapters XVII and XVIII.)

[40] BOX, G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143**, 383–430.

[41] BOX, G.E.P. and TIAO, G.C. (1973) *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading.

[42] BROWN, L.D. (1971) Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42**, 855–904.

[43] BROWN, L.D. (1979) An heuristic method for determining admissibility of estimators—with applications. *Ann. Statist.* **7**, 960–994.

[44] BRUNK, H. D. and PIERCE, D.A. (1977) Large sample posterior normality of the population mean. *Commun. Statist. A—Theory Methods* **6**, 1–14.

[45] BURNASEV, M.V. (1979) Asymptotic expansions of the integral risk of statistical estimators of location parameter in a scheme of independent observations. *Soviet Math. Dokl.* **20**, 788–791.

[46] CAMPBELL, G. and HOLLANDER, M. (1979) Nonparametric Bayes estimation with incomplete Dirichlet prior information. In: J.S. Rustagi, ed., *Optimizing Methods in Statistics,* Academic Press, New York.

[47] CASSEL, C.M., SÄRNDAL, C.E. and WRETMAN, J.H. (1977) *Foundations of Inference in Survey Sampling.* Wiley, New York.

[48] CHAMBERLAIN, G. and LEAMER, E.E. (1976) Matrix weighted averages and posterior bounds. *J. Roy. Statist. Soc. Ser. B.* **38**, 73–84.

[49] CHERNOFF, H. (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23**, 493–507.

[50] CHERNOFF, H. (1956) Large-sample theory: Parametric case. *Ann. Math. Statist.* **27**, 1–22.

[51] CHERNOFF, H. (1959) Sequential design of experiments. *Ann. Math. Statist.* **30**, 755–770.

[52] CHERNOFF, H. (1970) *Sequential Analysis and Optimal Design.* SIAM.

[53] CLEVENSON, M. and ZIDEK, J. (1975) Simultaneous estimation of the mean of independent Poisson laws. *J. Amer. Statist. Assoc.* **70**, 698–705.

[54] DAVIS, W.A. (1979) Approximate Bayesian predictive distributions and model selection. *J. Amer. Statist. Assoc.* **74**, 312–317.

[55] DAWID, A.P. (1979) On the limiting normality of posterior distributions. *Proc. Camb. Phil. Soc.* **67**, 625–633.

[56] DAWID, A.P. (1973) Posterior expectations for large observations. *Biometrika* **60**, 664–666.

[57] DE FINETTI, B. (1937) Foresight: Its logical laws, its subjective sources. In: H.E. Kyburg and H.E. Smokler, eds., *Studies in Subjective Probability* (1964), Wiley, New York.

[58] DE FINETTI, B. (1972) *Probability, Induction and Statistics.* Wiley, New York.

[59] DE FINETTI, B. (1974, 1975) *Theory of Probability*, Volumes 1 and 2. Wiley, New York.

[60] DE GROOT, M.H. (1970) *Optimal Statistical Decisions.* McGraw-Hill, New York.

[61] DEMPSTER, A. P. (1966) New Methods of reasoning toward posterior distributions based on sample data. *Ann. Math. Statist.* **37**, 355–374.

[62] DEMPSTER, A.P. (1967) Upper and lower probabilities induced by multivalued maps. *Ann. Math. Statist.* **38**, 325–339.

[63] DEMPSTER, A.P. (1968) A generalization of Bayesian inference. *J. Roy. Statist. Soc. Ser. B.* **30**, 205–248.

[64] DEMPSTER, A.P. (1971) Model searching and estimation in the logic of inference. In: V.P. Godambe and D.A. Sprott, eds., *Foundations of Statistical Inference*, Holt, Rinehart and Winston, Toronto.

[65] DEMPSTER, A.P. (1975) A subjectivist look at robustness. *Bull. of the International Statistical Institute* **46**, 349–374.

[66] DEMPSTER, A.P. (1976) Examples relevant to the robustness of applied inferences. In: S.S. Gupta and D.S. Moore, eds., *Statistical Decision Theory and Related Topics II*, Academic Press, New York.

[67] DE ROBERTIS, L. and HARTIGAN, J.A. (1981) Bayesian inference using intervals of measures. *Ann. Statist.* **9**, 235–244.

[68] DE ROUEN, T.A. and MITCHELL, T.J. (1974) A G_1-minimax estimator for a linear combination of binomial probabilities. *J. Amer. Statist. Assoc.* **69**, 231–233.

[69] DEY, D. (1980) On the choice of coordinates in simultaneous estimation of normal means. Technical Report 80-32, Dept. of Statistics, Purdue University, West Lafayette, IN.

[70] DEY, D. and BERGER, J. (1983) Combining coordinates in simultaneous estimation of normal means. *J. Statist. Planning and Inference* **8**, 143–160.

[71] DIACONIS, P. and FREEDMAN, D. (1982) Bayes rules for location problems. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III*, Academic Press, New York.

[72] DIACONIS, P. and FREEDMAN, D. (1983) Frequency properties of Bayes rules. In: G.E.P. Box, T. Leonard, and C.F. Wu, eds., *Scientific Inference, Data Analysis and Robustness*, Academic Press, New York.

[73] DIACONIS, P. and YLVISAKER, D. (1979) Conjugate priors for exponential families. *Ann. Statist.* **7**, 269–281.

[74] DIACONIS, P. and ZABELL, S. (1982) Updating subjective probability. *J. Amer. Statist. Assoc.* **77**, 822–830.

[75] DICKEY, J.M. (1973) Scientific reporting. *J. Roy. Statist. Soc. Ser. B* **35**, 285–305.

[76] DICKEY, J.M. (1974) Bayesian alternatives to the F-test and least-squares estimator in the normal linear model. In: S.E. Fienberg and A. Zellner, eds., *Studies in Bayesian Econometrics and Statistics*, North Holland, Amsterdam.

[77] DICKEY, J.M. (1976a) Discussion of 'Strong inconsistency from uniform priors' by M. Stone, *J. Amer. Statist. Assoc.* **71**, 119–125.

[78] DICKEY, J.M. (1976b) Approximate posterior distributions. *J. Amer. Statist. Assoc.* **71**, 680–689.

[79] DICKEY, J.M. and FREEMAN, P. (1975) Population—distributed personal probabilities. *J. Amer. Statist. Assoc.* **70**, 362–364.

[80] DOKSUM, K. (1970) Decision theory for some nonparametric models. In: *Proc. Sixth Berkeley Symp. Math. Statist. Prob. 1*, 331–343. University of California Press, Berkeley.

[81] DUBINS, L.E. (1969) An elementary proof of Bochner's finitely additive Radon–Kikodym theorem. *Amer. Math. Monthly* **76**, 520–523.

[82] DUBINS, L.E. (1975) Finitely additive conditional probabilities, conglomerability and disintegrations. *Ann. Probability* **3**, 89–99.

[83] EDWARDS, W., LINDMAN, H. and SAVAGE, L.J. (1963) Bayesian statistical inference for psychological research. *Psychological Review* **70**, 193–242. Reprinted as Part I of this book.

[84] EFRON, B. and HINKLEY, D. (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65**, 457–482.

[85] EFRON, B. and MORRIS, C. (1971) Limiting the risk of Bayes and empirical Bayes estimators—Part I: the Bayes case. *J. Amer. Statist. Assoc.* **66**, 807–815.

[86] EFRON, B. and MORRIS, C. (1972) Limiting the risk of Bayes and empirical Bayes estimators—Part 2: the empirical Bayes case. *J. Amer. Statist. Assoc.* **67**, 130–139.

[87] EFRON, B. and MORRIS, C. (1973) Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117–130.

[88] FABIUS, J. (1964) Asymptotic behavior of Bayes estimates. *Ann. Math. Statist.* **35**, 846–856.

[89] FAITH, R. (1978) Minimax Bayes estimators of a multivariate normal mean. *J. Multivariate Anal.* **8**, 372–379.

[90] FINE, T. (1973) *Theories of Probability.* Academic Press, New York.

[91] FISHBURN, P.C. (1965) Analysis of decisions with incomplete knowledge of probabilities. *Operations Research* **13**, 217–237.

[92] FISHBURN, P.C., MURPHY, A.H. and ISAACS, H.H. (1968) Sensitivity of decisions to probability estimation errors: a re-examination. *Operations Research* **16**, 253–268.

[93] FORTUS, R. (1979) Approximations to Bayesian sequential tests of composite hypotheses. *Ann. Statist.* **7**, 579–591.

[94] FRASER, D.A.S. (1979) *Inference and Linear Models.* McGraw-Hill, New York.

[95] FRASER, D.A.S. and MACKAY, J. (1976) On the equivalence of standard inference procedures. In: W.L. Harper and C.A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science, Vol. II*, Reidel, Boston.

[96] FREEDMAN, D. (1963) On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386–1403.

[97] FREEDMAN, D. (1965) On the asymptotic behavior of Bayes estimates in the discrete case II. *Ann. Math. Statist.* **35**, 454–456.

[98] GEISSER, S. and EDDY, W.F. (1979) A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74**, 153–160.

[99] GHOSH, J.K., SINHA, B.K. and JOSHI, S.N. (1982) Expansions for posterior probability and integrated Bayes risk. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III*, Academic Press, New York.

[100] GHOSH, M. and PARSIAN, A. (1981) Bayes minimax estimation of multiple Poisson parameters. *J. Multivariate Anal.* **11**, 280–288.

[101] GLESER, L.J. and KUNTE, S. (1976) On asymptotically optimal sequential Bayes interval estimation procedures. *Ann. Statist.* **4**, 685–711.

[102] GODAMBE, V.P. (1982) Estimation in survey sampling: robustness and optimality. *J. Amer. Statist. Assoc.* **77**, 393–406.

[103] GODAMBE, V.P. and THOMPSON, M.E. (1971) The specification of prior knowledge by classes of prior distributions in survey sampling estimation. In: V.P. Godambe and D.A. Sprott, eds., Foundations of Statistical Inference. Holt, Rinehart and Winston, Toronto.

[104] GODAMBE, V.P. and THOMPSON, M.E. (1977) Robust near optimal estimation in survey practice. *Bulletin of the International Statist. Inst.* **47**, 127–170.

[105] GOLDSTEIN, M. (1974) Approximate Bayesian inference with incompletely specified prior distributions. *Biometrika* **61**, 629–631.

[106] GOLDSTEIN, M. (1980) The variance modified Bayes estimator. *J. Roy. Statist. Soc. Ser. B* **41**, 96–100.

[107] GOLDSTEIN, M. (1980) The linear Bayes regression estimator under weak prior assumptions. *Biometrika* **67**, 621–628.

[108] GOOD, I.J. (1950) *Probability and the Weighing of Evidence.* Griffin, London.

[109] GOOD, I.J. (1952) Rational decisions. *J. Roy. Statist. Soc. Ser. B* **14**, 107–114.

[110] GOOD, I.J. (1962a) Subjective probability as the measure of a nonmeasurable set. In: *Logic, Methodology and Philosophy of Science.* University Press, Stanford.

[111] GOOD, I.J. (1962b) How rational should a manager be? *Management Science* **8**, 383–393.

[112] GOOD, I.J. (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press, Cambridge.

[113] GOOD, I.J. (1968) The utility of a distribution. *Nature* **219**, 1392.

[114] GOOD, I.J. (1973) The probabilistic explication of evidence, surprise, causality, explanation and utility. In: V.P. Godambe and D.A. Sprott, eds., *Foundations of Statistical Inference*, Holt, Rinehart and Winston, Toronto.

[115] GOOD, I.J. (1976) The Bayesian influence, or how to sweep subjectivism under the carpet. In: W.L. Harper and C.A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science II*, Reidel, Boston.

[116] GOOD, I.J. (1980) Some history of the hierarchical Bayesian methodology. In: J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith, eds., *Bayesian Statistics*, University Press, Valencia.

[117] GOOD, I.J. (1983) The robustness of a hierarchical model for multinomials and contingency tables. In: G.E.P. Box, T. Leonard, and C.F. Wu, eds., *Scientific Inference, Data Analysis and Robustness*, Academic Press, New York.

[118] GOOD, I.J. and CROOK, J.F. (1974) The Bayes/non Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711–720.

[119] GUPTA, S.S. and HSIAO, P. (1981) On Γ-minimax, minimax and Bayes procedures for selecting populations close to a control. *Sankhyā Ser. B.* **43**, 291–318.

[120] GUPTA, S.S. and HUANG, D.Y. (1975) On Γ-minimax classification procedures. *Proc. of the 40th Session of the International Statistical Institute* **46**, Book 3, 330–335.

[121] GUPTA, S.S. and HUANG, D.Y. (1977) On some Γ-minimax selection and multiple comparison procedures. In: S.S. Gupta and D.S. Moore, eds., *Statistical Decision Theory and Related Topics II*, Academic Press, New York.

[122] GUPTA, S.S. and KIM, W.C. (1980) Γ-minimax and minimax rules for comparison of treatments with a control. In: K. Matusita, ed., *Recent Developments in Statistical Inference and Data Analysis*, North-Holland, Amsterdam.

[123] HÁJEK, J. (1981) *Sampling from a Finite Population.* Dekker, New York.

[124] HARTIGAN, J.A. (1969) Linear Bayesian methods. *J. Roy. Statist. Soc. Ser. B* **31**, 446–454.

[125] HEATH, D.C. and SUDDERTH, W.D. (1978) On finitively additive priors, coherence and extended admissibility. *Ann. Statist.* **6**, 333–345.

[126] HEYDE, C.C. and JOHNSTONE, I.M. (1979) On asymptotic posterior normality for stochastic processes. *J. Roy. Statist. Soc. Ser. B* **41**, 184–189.

[127] HILDRETH, C. (1963) Bayesian statisticians and remote clients. *Econometrica* **31**, 422–438.

[128] HILL, B. (1965) Inference about variance components in the one-way model. *J. Amer. Statist. Assoc.* **60**, 806–825.

[129] HILL, B. (1969) Foundations for the theory of least squares. *J. Roy. Statist. Soc. Ser. B* **31**, 89–97.

[130] HILL, B. (1970) Some contrasts between Bayesian and classical inference in the analysis of variance and in the testing of models. In: D.L. Meyer and R.O. Collier, Jr., eds., *Bayesian Statistics*, Peacock Publ., Itasca, IL.

[131] HILL, B. (1974) On coherence, inadmissibility and inference about many parameters in the theory of least squares. In: S. Fienberg and A. Zellner, eds., *Studies in Bayesian Econometrics and Statistics*, North-Holland, Amsterdam.

[132] HILL, B. (1975) A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163–1174.

[133] HILL, B. (1977) Exact and approximate Bayesian solutions for inference about variance

components and multivariate inadmissibility. In: A. Aykac and C. Brumat, eds., *New Developments in the Applications of Bayesian Methods*, North-Holland, Amsterdam.

[134] HILL, B. (1980a) Robust analysis of the random model and weighted least squares regression. In: *Evaluation of Econometric Models.* Academic Press, New York.

[135] HILL, B. (1980b) On some statistical paradoxes and non-conglomerability. In: J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith, eds., *Bayesian Statistics*, University Press, Valencia.

[136] HILL, B. (1980c) Invariance and robustness of the posterior distribution of characteristics of a finite population, with reference to contingency tables and the sampling of species. In: A. Zellner, ed., *Bayesian Analysis in Econometrics and Statistics, Essays in Honor of Harold Jeffreys*, North-Holland, Amsterdam.

[137] HINKLEY, D.V. (1983) Can frequentist inference be very wrong? A conditional 'Yes'. In: G.E.P. Box, T. Leonard, and C.F. Wu, eds., *Scientific Inference, Data Analysis and Robustness*, Academic Press, New York.

[138] HODGES, J.L. and LEHMANN, E.L. (1952) The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* **23**, 396–407.

[139] HSIAO, P. (1982) Γ-minimax procedures for selecting good location parameters in some multivariate distributions. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III*, Academic Press, New York.

[140] HUBER, P.J. (1972) Robust statistics: a review. *Ann. Math. Statist.* **43**, 1041–1067.

[141] HUBER, P.J. (1973) The use of Choquet capacities in statistics. *Bulletin of the International Statist. Inst.* **45**, 181–191.

[142] HUBER, P.J. and STRASSEN, V. (1973) Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* **1**, 251–263.

[143] HUDSON, H.M. and TSUI, K. (1981) Simultaneous Poisson estimators for a priori hypotheses about means. *J. Amer. Statist. Assoc.* **76**, 182–187.

[144] HWANG, J.T. (1982a) Semi tail upper bounds on the class of admissible estimators in discrete exponential families with applications to Poisson and negative binomial distributions. *Ann. Statist.* **10**, 1137–1147.

[145] HWANG, J.T. (1982b) Certain bounds on the class of admissible estimators in continuous exponential families. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III*, Academic Press, New York.

[146] ISAACS, H.H. (1963) Sensitivity of decisions to probability estimation errors. *Operations Research* **11**, 536–552.

[147] JACKSON, D.A., DONOVAN, T.M., ZIMMER, W.J. and DEELY, J.J. (1970) Γ_2-minimax estimators in the exponential family. *Biometrika* **57**, 439–443.

[148] JACKSON, P., NOVICK, M. and DEKEYREL, D. (1980) Adversary preposterior analysis for simple parametric models. In: A. Zellner, ed., *Bayesian Analysis in Economics and Statistics*, North-Holland, Amsterdam.

[149] JAYNES, E.T. (1968) Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics* **4**, 227–241.

[150] JAYNES, E.T. (1976) Confidence intervals versus Bayesian intervals. In: W.L. Harper and C.A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science II*, Reidel, Boston.

[151] JAYNES, E.T. (1981) The intuitive inadequacy of classical statistics. Presented at the International Convention on Fundamentals of Probability and Statistics, Luino, Italy.

[152] JEFFREY, R. (1968) Probable knowledge. In: I. Lakatos, ed., *The Problem of Inductive Logic*, North-Holland, Amsterdam.

[153] JEFFREYS, H. (1961) *Theory of Probability, 3rd Ed.* Oxford University Press, Oxford.

[154] JOHNSON, B.R. and TRUAX, D.R. (1978) Asymptotic behavior of Bayes procedures for testing simple hypotheses in multiparameter exponential families. *Ann. Statist.* **6**, 346–361.

[155] JOHNSON, R.A. (1967) An asymptotic expansion for posterior distributions. *Ann. Math. Statist.* **38**, 1899–1906.

[156] JOHNSON, R.A. (1970) Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41**, 851–864.

[157] KADANE, J.B. and CHUANG, D.T. (1978) Stable decision problems. *Ann. Statist.* **6**, 1095–1110. Reprinted as Part IV of this book.

[158] KADANE, J.B., DICKEY, J.M., WINKLER, R.L., SMITH, W.S. and PETERS, S.C. (1980) Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.* **75**, 845–854.

[159] KADANE, J.B., SCHERVISH, M. and SEIDENFELD, T. (1981) Statistical implications of finitely additive probability. Technical Report no. 206, Carnegie–Mellon University, Pittsburgh.

[160] KIEFER, J. (1977) The foundations of statistics—are there any? *Synthese* **36**, 161–176.

[161] KIEFER, J. and SACKS, J. (1963) Asymptotically optimum sequential inference and design. *Ann. Math. Statist.* **34**, 705–750.

[162] KLEYLE, R. (1975) Upper and lower probabilities for discrete distributions. *Ann. Statist.* **3**, 504–511.

[163] KOOPMAN, B.O. (1940) The axioms and algebra of intuitive probability. *Ann. Math.* **41**, 269–278.

[164] KRAFT, C.H., PRATT, J.W. and SEIDENBERG, A. (1959) Intuitive probability on finite sets. *Ann. Math. Statist.* **30**, 408–419.

[165] KUDŌ, H. (1967) On partial prior information and the property of parametric sufficiency. *Proc. Fifth Berkeley Symp. Prob. Statist.* **1**. University of California Press, Berkeley.

[166] KYBURG, H. (1974) *The Logical Foundations of Statistical Inference.* Reidel, Dordrecht.

[167] KYBURG, H.E. (1976) Statistical knowledge and statistical inference. In: W.L. Harper and C.A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science II,* D. Reidel, Boston.

[168] LAMBERT, D. and DUNCAN, G. (1981) Bayesian learning based on partial prior information. Technical Report no. 209, Department of Statistics, Carnegie–Mellon University.

[169] LEAMER, E.E. (1978) *Specification Searches.* Wiley, New York.

[170] LE CAM, L. (1953) On some asymptotic properties of the maximum likelihood estimates and related Bayes estimates. *University of California Pub. Statist.* **1**, 277–330.

[171] LE CAM, L. (1956) On the asymptotic theory of estimation and testing hypotheses. *Proc. 3rd Berkeley Symp. Math. Statist. Probability* **1**, 129–156. University of California Press, Berkeley.

[172] LE CAM, L. (1982) On the risk of Bayes estimates. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III,* Academic Press, New York.

[173] LEONARD, T. (1976) Some alternative approaches to multiparameter estimation. *Biometrika* **63**, 69–76.

[174] LEVI, I. (1974) On indeterminate probabilities. *J. of Philosophy* **71**.

[175] LEVI, I. (1980) *The Enterprise of Knowledge.* MIT Press, Cambridge.

[176] LINDLEY, D.V. (1960) The use of prior probability distributions in statistical inference and decisions. *Proc. Berkeley Symp. Math. Statist. Probability* **1**, 453–468. University of California Press, Berkeley.

[177] LINDLEY, D.V. (1968) The choice of variables in multiple regression (with discussion). *J. Roy. Statist. Soc. Ser. B* **30**, 31–66.

[178] LINDLEY, D.V. (1982) Scoring rules and the inevitability of probability. *International Statistical Review* **50**, 1–26.

[179] LINDLEY, D.V. and NOVICK, M. (1981) The role of exchangeability in inference. *Ann. Statist.* **9**, 45–58.

[180] LINDLEY, D.V. and SMITH, A.F.M. (1972) Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34**, 1–41.

[181] MANSKI, C.F. (1981) Learning and decision making when subjective probabilities have subjective domains. *Ann. Statist.* **9**, 59–65.

[182] MARAZZI, A. (1980) Robust Bayesian estimation for the linear model. Research Report no. 27, Fachgruppe fuer Statistik, Eidgenoessische Technische Hochschule, Zurich.

[183] MARITZ, J.S. (1970) *Empirical Bayes methods.* Methuen, London.

[184] MASRELIEZ, C.J. and MARTIN, R.D. (1977) Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *IEEE Trans. on Automatic Control* **22**, 361–371.

[185] MEEDEN, G. and ISAACSON, D. (1977) Approximate behavior of the posterior distribution for a large observation. *Ann. Statist.* **5**, 899–908.

[186] MENGES, G. (1966) On the Bayesification of the minimax principle. *Unternehmensforschung* **10**, 81–91.

[187] MIESCKE, K.J. (1981) Γ-minimax selection procedures in simultaneous testing problems. *Ann. Statist.* **9**, 215-220.

[188] MORRIS, C. (1977) Interval estimation for empirical Bayes generalizations of Stein's estimator. The Rand Paper Series, Rand Corp., Santa Monica.

[189] MORRIS, C. (1982) Natural exponential families with quadratic variance functions. *Ann. Statist.* **10**, 65–80.

[190] MORRIS, C. (1983a) Parametric empirical Bayes confidence intervals. In: G.E.P. Box, T. Leonard and C.F. Wu, eds., *Scientific Inference, Data Analysis and Robustness*, Academic Press, New York.

[191] MORRIS, C. (1983b) Parametric empirical Bayes inference: Theory and applications (with Discussion). *J. Amer. Statist. Assoc.* **78**, 47–65.

[192] NOVICK, M.R. (1969) Multiparameter Bayesian indifference procedures (with discussion). *J. Roy. Statist. Soc. Ser. B* **31**, 29–64.

[193] PIERCE, D.A. and FOLKS, J.L. (1969) Sensitivity of Bayes procedures to the prior distribution. *Operations Research* **17**, 344–350.

[194] POLASEK, W. (1983) Multivariate regression systems: Estimation and sensitivity analysis of two-dimensional data. (This Volume.)

[195] POTTER, J.M. and ANDERSON, B.D.O. (1980) Prior information and decision making. *IEEE Trans. Syst., Man., Cybern.* **10**, 125–133.

[196] PRATT, J.W. (1965) Bayesian interpretation of standard inference statements (with discussion). *J. Roy. Statist. Soc. Ser. B* **27**, 169–203.

[197] PRATT, J.W., RAIFFA, H. and SCHLAIFER, R. (1965) *Introduction to Statistical Decision Theory.* McGraw-Hill, New York.

[198] RAIFFA, H. and SCHLAIFER, R. (1961) *Applied Statistical Decision Theory.* Harvard University, Boston.

[199] RAMSAY, J.O. and NOVICK, M.R. (1980) PLU robust Bayesian decision theory: point estimation. *J. Amer. Statist. Assoc.* **75**, 901–907.

[200] RANDLES, H.R. and HOLLANDER, M. (1971) Γ-minimax selection procedures in treatment versus control problems. *Ann. Math. Statist.* **42**, 330–341.

[201] RIOS, S. and GIRÓN, F.J. (1980) Quasi-Bayesian behavior: a more realistic approach to decision making? In: J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith, eds., *Bayesian Statistics*, University Press, Valencia.

[202] ROBBINS, H. (1951) Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statist. Prob.* **1**. Univ. of California Press, Berkeley, pp. 131–148.

[203] ROBBINS, H. (1955) An empirical Bayes approach to statistics. *Proc. Third Berkeley Symposium Math. Statist. Prob.* **1**. University of California Press, Berkeley, pp. 157–164.

[204] ROBBINS, H.E. (1964) The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35**, 1–20.

[205] ROSENKRANTZ, R.D. (1977) *Inference, Method and Decision: Towards a Bayesian Philosophy of Science.* Reidel, Boston.

[206] ROYALL, R.M. and PFEFFERMANN, D. (1982) Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika* **69**, 401–409.

[207] RUBIN, D.B. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6**, 34–58.

[208] RUBIN, H. (1971) A decision-theoretic approach to the problem of testing a null hypothesis. In: S.S. Gupta and J. Yackel, eds., *Statistical Decision Theory and Related Topics*, Academic Press, New York.

[209] RUBIN, H. (1972) On large sample properties of certain nonparametric procedures. *Proc. Sixth Berkeley Symp. Math. Statistics and Prob.* University of California Press, Berkeley, pp. 429–435.

[210] Rubin, H. (1977) Robust Bayesian estimation. In: S.S. Gupta and D. Moore, eds., *Statistical Decision Theory and Related Topics II*, Academic Press, New York.

[211] Rubin, H. and Sethuraman, J. (1965) Bayes risk efficiency. *Sankhyā A* **27**, 347–356.

[212] Savage, L.J. (1954) *The Foundations of Statistics.* Wiley, New York.

[213] Savage, L.J. (1961) The foundations of statistics reconsidered. *Proc. Fourth Berkeley Symp. Math. Statistics and Prob.* University of California Press, Berkeley, pp. 575–586.

[214] Savage, L.J. (1962) Bayesian statistics. In: R.E. Machol and P. Gray, eds., *Recent Developments in Information and Decision Processes*, Macmillan, New York.

[215] Savage, L.J. et al. (1962) *The Foundations of Statistical Inference.* Methuen, London.

[216] Schneeweiss, H. (1964) Eine Entscheidungsregel für den Fall partiell bekannter Wahrscheinlichkeiten. *Unternehmensforschung* **8** no. 2, 86–95.

[217] Schwartz, L. (1965) On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4**, 10–26.

[218] Schwarz, G. (1962) Asymptotic shapes of Bayes sequential testing regions. *Ann. Math. Statist.* **33**, 224–236.

[219] Schwarz, G. (1968) Asymptotic shapes for sequential testing of truncation parameters. *Ann. Math. Statist.* **39**, 2038–2043.

[220] Shafer, G. (1976) *A Mathematical Theory of Evidence.* Princeton University Press, Princeton.

[221] Shafer, G. (1979) Two theories of probability. In: *PSA 1978,* Vol. 2, Philosophy of Science Association, East Lansing, Michigan.

[222] Shafer, G. (1981a) Jeffrey's rule of conditioning. *Philosophy of Science* **48**, 337–362.

[223] Shafer, G. (1981b) Constructive probability. *Synthese* **48**, 1–60.

[224] Shafer, G. (1982) Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B* **44**, 322–352.

[225] Shapiro, S.H. (1972) A compromise between the Bayes and minimax approaches to estimation. Technical Report no. 31, Department of Statistics, Stanford University.

[226] Shapiro, S.H. (1975) Estimation of location and scale parameters—a compromise. *Commun. Statist.* **4**(12), 1093–1108.

[227] Skibinski, M. and Cote, L. (1963) On the inadmissibility of some standard estimates in the presence of prior information. *Ann. Statist.* **34**, 539–548.

[228] Smith, C.A.B. (1961) Consistency in statistical inference and decision. *J. Roy. Statist. Soc. Ser. B* **23**, 1–25.

[229] Smith, G. and Campbell, F. (1980) A critique of some ridge regression methods (with discussion). *J. Amer. Statist. Assoc.* **75**, 74–103.

[230] Solomon, D.L. (1972a) Λ-minimax estimation of a multivariate location parameter. *J. Amer. Statist. Assoc.* **67**, 641–646.

[231] Solomon, D.L. (1972b) Λ-minimax estimation of a scale parameter. *J. Amer. Statist. Assoc.* **67**, 647–649.

[232] Spielman, S. (1977) Physical probability and Bayesian statistics. *Synthese* **36**, 235–269.

[233] Srinivasan, C. (1980) Admissible generalized Bayes estimators and exterior boundary value problem. *Sankhyā.* To appear.

[234] Stein, C. (1965) Approximation of improper prior measures by prior probability measures. In: *Bernouilli-Bayes-Laplace Festchr.*, 217–240. Springer, New York.

[235] Stein, C. (1981a) Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–1151.

[236] Stein, C. (1981b) On the coverage probability of confidence sets based on a prior distribution. Technical Report no. 180, Dept. of Statistics, Stanford University.

[237] Stone, M. (1963) Robustness of nonideal decision procedures. *J. Amer. Statist. Assoc.* **58**, 480–486.

[238] Stasser, H. (1981) Consistency of maximum likelihood and Bayes estimates. *Ann. Statist.* **9**, 1107–1113.

[239] Strawderman, W.E. (1971) Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42**, 385–388.

[240] STRAWDERMAN, W. and COHEN, A. (1971) Admissibility of estimators of the mean vector of a multivariate normal distribution with quadratic loss. *Ann. Math. Statist.* **42**, 270–296.

[241] SUPPES, P. (1975) Approximate probability and expectation of gambles. *Erkenntnis* **9**, 153–161.

[242] SUPPES, P. and ZANOTTI, M. (1977) On using random relations to generate upper and lower probabilities. *Synthese* **36**, 427–440.

[243] TELLER, P. (1976) Conditionalization, observation and change of preference. In: W. Harper and C.A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Reidel, Dordrecht.

[244] THEIL, H. (1963) On the use of incomplete prior information in regression analysis. *J. Amer. Statist. Assoc.* **58**, 401–414.

[245] TIAO, G.C. and ZELLNER, A. (1964) On the Bayesian estimation of multivariate regression. *J. Roy. Statist. Soc. Ser. B* **26**, 277–285.

[246] UMBACH, D. (1978) On the approximate behavior of the posterior distribution for an extreme multivariate observation. *J. Multivariate Anal.* **8**, 518–531.

[247] VARDI, Y. (1979a) Asymptotic optimality of certain sequential estimators. *Ann. Statist.* **7**, 1034–1039.

[248] VARDI, Y. (1979b) Asymptotic optimal sequential estimation. *Ann. Statist.* **7**, 1040–1051.

[249] WALKER, A.M. (1969) On the asymptotic behavior of posterior distributions. *J. Roy. Statist. Soc. Ser. B* **31**, 80–88.

[250] WATSON, S.R. (1974) On Bayesian inference with incompletely specified prior distributions. *Biometrika* **61**, 193–196.

[251] WEERHANDI, S. and ZIDEK, J.V. (1981) Multi-Bayesian statistical decision theory. *J. Roy. Statist. Soc. Ser. A* **144**, 85–93.

[252] WELCH, B.L. and PEERS, H.W. (1963) On formulas for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **25**, 318–329.

[253] WEST, M. (1981) Robust sequential approximate Bayesian estimation. *J. Royal Statist. Soc. Ser B* **43**, 157–166.

[254] WEST, S. (1979) Upper and lower probability inferences for the logistic function. *Ann. Statist.* **7**, 400–413.

[255] WILKINSON, G.N. (1977) On resolving the controversy in statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 119–171.

[256] WILLIAMS, P.M. (1976) Indeterminate probabilities. In: M. Przelecki, K. Szaniawski and R. Wójciki, eds., *Formal Methods in the Methodology of Empirical Sciences.* Reidel, Dordrecht.

[257] WOLFENSON, M. and FINE, T. (1982) Bayes-like decision making with upper and lower probabilities. *J. Amer. Statist. Assoc.* **77**, 80–88.

[258] WOLPERT, R. and BERGER, J. (1982) Incorporating prior information in minimax estimation of the mean of a Gaussian process. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III.* Academic Press, New York.

[259] WOODROOFE, M. (1980) On the Bayes risk incurred by using asymptotic shapes. *Commun. Statist. A—Theory Methods* **9**, 1727–1748.

[260] ZAMAN, ASAD (1982) Quasitransitive preferences over lotteries. Technical Report, University of Pennsylvania.

[261] ZELLNER, A. (1971) *An Introduction to Bayesian Inference in Econometrics.* Wiley, New York.

[262] ZELLNER, A. (1976) Bayesian analysis of the regression model with multivariate Student-t error terms. *J. Amer. Statist. Assoc.* **71**, 400–405.

[263] ZELLNER, A. (1982) Applications of Bayesian analysis in econometrics. Technical Report, H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago, Chicago.

[264] ZELLNER, A. and GEISEL, M. (1968) Sensitivity to control of uncertainty and form of the criterion function. In: D.G. Watts, ed., *The Future of Statistics.* Academic Press, New York.

[265] ZHENG, Z. (1982) A class of generalized Bayes minimax estimators. In: S.S. Gupta and J. Berger, eds., *Statistical Decision Theory and Related Topics III.* Academic Press, New York.

[266] ZIDEK, J.V. (1982) Aspects of multi-Bayesian theory. Technical Report No. 82-11, Dept. of Mathematics, University of British Columbia, Vancouver.

PART III

COMMENTS ON "THE ROBUST BAYESIAN VIEWPOINT" BY JAMES BERGER

Robustness of Bayesian Analyses, edited by Joseph B. Kadane
© Elsevier Science Publishers B.V. (1984) 125–144

1. Comment by *Lawrence D. Brown**

Professor Berger's robust Bayesian proposal ranks among the most important statistical ideas of the past decade. This idea has already been described in several of his papers, which Berger cites in his references. Now he has tried to supply a detailed pragmatic/philosophic rationale for this proposal. It is an enormous task and he has done an admirable job.

My perspective toward the proposal differs somewhat from Berger's. Because of this I tend to give greatest emphasis to different aspects of it, and to different unsettled issues. However, I believe I am in agreement with Berger regarding the general structure of the proposal, its path of development, and its broad potential applicability. I emphasize these facts now and will do so again in closing, in order to try to make clear that the following comments are not meant as criticisms but are offered in part as an additional path of support for the theory and in part as suggestions concerning questions which should be answered in order to solidify its foundations.

My perspective can be described in a few words as that of a pure, flexible, collective, pragmatic frequentist. I will now try to briefly explain this mouthful of words!

Here is the basic principle: Statistical (and other) procedures being used today and being proposed for future use should be judged collectively and realistically according to their long term expected consequences.

The preceding is a statement of general philosophy, not a mathematical axiom. This principal is to some degree connected with an acceptance of fundamental axioms of probability theory such as those of Kolmogorov. It must be emphasized, however, that it is not a reexpression of these axioms nor a logical antecedent or consequence of them. It does not confirm or deny the possibility of structuring statistical practice on the basis of this system, or of any particular axiomatic system. To this extent the perspective is pure in the sense of being as yet unadulterated by additional layers of structure but it is flexible in that it solicits the construction of further principles or axiomatic systems consistent with its basic premise.

In operational terms the principle suggests attempting to codify as carefully as possible the types of statistical reasoning and consequent procedures to be used for all different classes of statistical problems. These procedures should then be judged according to the question: If they are used as specified for the next 5 (500? 50 000?) years how well will they do collectively. Since there is, it

*Cornell University, Ithaca, NY, U.S.A. These comments were supported in part by N.S.F. Research Grant MCS 8200035.

would seem, no universal standard, it is only possible to answer the question in a relative sense: Do these procedures work better over the long run than some others which have been proposed? The question should be interpreted in a collective sense. It is undoubtedly true that each statistical situation has its unique features. But this does not mean that each problem should be treated in isolation. Theories of probability still apply collectively to the wide range of statistical situations. (This does not deny the possible existence of issues—for example epistemological questions—which properly lie outside this wide range of situations.) This point may be clearly illustrated in relation to classical theories of probability: The law of large numbers applies not only to independent identically distributed events, but to independent nonidentically distributed events as well. (And also to certain dependent events.) The only stipulation is the Liapunov condition which specifies that no single event stands out in relation to the others.

In viewing the collective aspect of this argument an analogy can be made with Mill's theory of utilitarianism. Mill's subject was of course ethics, not statistics, and the analogy is in other important respects not perfect. But there are revealing parallels in the following quotation and elsewhere. Mill wrote in his *Utilitarianism*:

> "...the standard of what is right in conduct is not the agents own happiness but the happiness of all concerned."

The frequentist framework as so far elucidated is potentially consistent with various Bayesian prescriptions. Indeed some of them viewed axiomatically seem to be structured precisely to guarantee optimum long term behaviour. Many Bayesian theories do not have the collective feature built into them, but some do have or can be modified in an attempt to contain it.

There is however a further aspect to the frequentist point of view being espoused here. It requires also firmly based pragmatic judgments concerning proposed methodology. The question needs to be answered, "How well will the proposed methodology actually perform in practical situations (when used by intelligent well-educated practitioners)?" It is here, of course, that Berger's 'Assumption II', with which I completely agree, gains its relevance and force.

Let me sketch why I think the frequentist orientation leads inexorably through Assumption II to a robust Bayesian viewpoint; and also to point out what I found to be the biggest surprise in Berger's paper.

There are statistical estimation situations possessing a loss function and others for which it is reasonable to act as if certain conventional loss functions (such as squared error) are valid. In such a circumstance the frequentist orientation leads inexorably to the possibility of proposing various decision rules, calculating their respective risk function and eliminating those rules which prove to be inadmissible. (One can reasonably decide to make minor corrections and use some rule which is inadmissible by a numerically insignificant amount if it possesses some other virtue, such as convenience.) But this activity does not lead to a unique choice of statistical procedure. There are many admissible procedures from which to choose.

There may have been a brief period of euphoria during the development of Wald's (1950) Theory of Statistical Decision Functions when it was thought that the minimax principle would provide a unique satisfactory answer in a variety of situations. If it was ever thought so, the period was brief, for Wolfowitz who collaborated with Wald on aspects of the theory wrote (1951) that the minimax principle "might be the course of a very conservative statistician" and then proceeded to examine the structure and appeal of a variety of nonminimax rules.

Whatever the outlook historically, it is now clear that there are important problems in which an otherwise plausible (and minimax) procedure is inadmissible by a significant margin. The most outstanding example is of course that involving the Stein effect. The situation is described in Berger's Section 4.5 from the point of view of the frequentist position. In brief, there are many possible minimax rules. The only sensible way to choose among them seems to be to construct some crude prior distribution (Berger's π_A) and then minimize the posterior risk among the minimax or C-minimax rules as proposed in Berger's (4.6). This leads to the procedure δ^* described below (4.7), or possibly to some slightly better, smooth alternative procedure such as that described in Berger (1980a). Because of the restriction to (C-)minimax rules the determination of the prior need only be rather crude. It suffices in this special example to determine only its mean and variance. (Perhaps, more realistically, one should think in terms of more robust measures such as an α-trimmed mean and variance.)

The same recommended procedure, δ^*, can be arrived at through robust Bayesian reasoning based on the class Γ, or some similar class. This is not surprising since there seems to be a close mathematical link, though not a philosophical one, between imposition of (C-)minimaxity and minimization of maximum expected risk over a broad class of priors. (It would be interesting to see if there is a valid, precise theorem to this effect.) To my taste the robust Bayesian approach seems ultimately preferable since it removes the artificial choice of C from the prescription. Furthermore C-minimaxity can only lead to sensible answers in certain types of problems. On the other hand we need to know much more about the possible consequences of various types of choices for Γ before the robust Bayesian method can be used with confidence in a variety of situations, especially those where it may not lead to proposal of a (C-)minimax procedure.

The C-minimax and robust Bayesian proposals are both described above as methodologically frequentist suggestions. One calculates risk functions first, then integrates them against prior distributions, and choose that procedure leading to the best value(s) of $r(\pi, \delta)$. (Best for the given π over the allowable (C-minimax) δ and/or 'best' as π ranges over the allowable priors ($\pi \in \Gamma$).) This methodology—suggested by the frequentist argument—leads automatically to a procedure acceptable within the frequentist philosophy (subject of course to the long term validity of the remaining assumptions concerning the probability model and loss function, etc., and a sufficiently careful choice of π (or Γ)).

Here is the surprise: Although the above proposal is philosophically and methodologically frequentist it can be implemented *most of the time* in a Bayesian fashion. This is convincingly the case in Example 4 (Section 3.1) for $x = 0, 1, 2$ and in similar examples. Berger calls this phenomenon 'posterior robustness'. When posterior robustness occurs even the dogmatic frequentist should act as if he were a dogmatic Bayesian.

I think it would be revealing and useful to have at hand mathematical results connecting the structure of the class Γ with the probability of occurrence of posterior robustness. For example suppose $X \sim N(\theta, 1)$, $L(\theta, a) = (\theta - a)^2$ as in Example 4 and Γ is specified by (2.1) with $\pi_A = N(\theta, 1)$ and $\varepsilon = 0.1$ which seems like a plausible value. Then for $x^2 = 8$ one has

$$\inf_{a_0(x)} \sup_{\pi \in \Gamma} |\rho(\pi, \tfrac{1}{2}x) - \rho(\pi, x, a_0(x))| \sim 0.6 \ .$$

(For $x^2 = 4$ this quantity is $\sim$0.1.) The value 0.6 seems unsatisfactory (but 0.1 may be O.K.) since $\rho(\pi, x, a_0(x)) < \rho(\pi, x, x) = 1$. Now $\Pr\{X^2 \geqslant 8\} \sim 0.05$ (and $\Pr\{X^2 \geqslant 4\} \sim 0.32$). Hence posterior robustness holds 70–95% of the time in this problem. (It may be that this type of calculation is unfortunately sensitive to the choice of Γ and to the choice of loss function. This too needs to be investigated.)

At the risk of repeating what Berger has written, let me summarize the preceding arguments as concisely as possible. I do this because I think there may be at least a difference of emphasis between my position and that in Jim Berger's paper and this summary may help to isolate this difference.

(1) The frequentist approach leads naturally to the quest for a priori information in order to choose among admissible procedures.

(2) Pragmatic considerations make it clear that a priori information cannot be specified with perfect precision. (Assumption II.) Furthermore, Bayes procedures for priors which are very close to each other (when viewed a priori) may have very different long term (frequentist) performance.

(3) Therefore, expected risk ($r(\pi, \delta)$) must in principle be investigated over all priors in the range of those which seem plausible. (This range is the class Γ.)

(4) When posterior robustness over Γ is present direct use of Bayes rule gives a satisfactory answer without computation of $r(\pi, \delta)$. Thus, although the robust Bayesian proposal is in my view frequentist, it can be implemented a large proportion of the time in a methodologically Bayesian fashion.

(5) There is much work to be done. Theoretical and numerical work is needed in order to delineate what types of classes Γ we should really be looking for in particular problems (see B below), what procedures are really robust over these classes, and how posterior robustness is really related to robustness

of $r(\pi, \delta)$. I think we also need to continue the process admirably begun by Jim Berger of examining in depth the logical and philosophical aspects of a comprehensive robust Bayesian philosophy.

In a spirit of furthering this continuing examination and in the hope of giving Professor Berger something specific to comment on, let me mention three specific issues which concerned me in reading his paper.

A. My first concern and only real criticism of the paper is what I view as the imprecision of Assumption I. I suspect that it does not literally say what it is intended to convey. Literally, its principle conclusion is that "the only trustworthy and sensible measures of this [posterior] information are Bayesian posterior measures". Does this convey implicity that these measures must be derived by the conventional Bayesian methodology? I think not. But if not, then what is the meaning and content of these measures—and are they really "trustworthy and sensible"? Thus suppose the statistician quotes a posterior distribution $P(\cdot)$ after observing some data x. This is presumably the type of measure that Assumption I is promoting. If it is calculated according to Bayes rule in the conventional fashion then it is indeed "trustworthy and sensible" (so long as the prior was) and may be correctly interpreted in the same terms as the prior. (For example, if the prior were a true (frequential) prior then $P(B)$ would be a statement about the (frequential) probability that $\theta \in B$.) However, suppose $P(\cdot)$ is calculated in a robust Bayes fashion. (Note that no precise prescription has been given for doing this and perhaps none can be given.) Then how can $P(\cdot)$ correctly have the same interpretation as would a prior? In fact, does it even have any correct meaningful interpretation? It probably can be given a meaningful interpretation through certain frequency calculations. If so, however, this seems to me to place these pseudo-Bayesian measures as objects within a frequentist setting rather than as primitive notions of an essentially axiomatic nature. Until I feel I understand better the content of this 'assumption', I remain somewhat open as to its possible interpretations and its position in the overall construction of a rationale for a robust Bayesian methodology.

In this connection I note that there exist non-Bayesian ways to make sense of measures which appear to be formally analogous to Bayesian posterior distributions, but I do not think these have the universal applicability and appeal of the more classical interpretations. There are, for example, betting paradigms such as those expressed in Robinson (1979a, 1979b), Bondar (1977) and Heath and Sudderth (1978). There is also the decision theoretic approach expressed in Gatsonis (1982a, 1982b) and in Eaton (1982).

B. A related concern involves the class Γ. If one begins with a class such as Γ of (2.1), makes an observation and examines the class of resulting posteriors, one finds that it is a somewhat strange looking conglomeration. It is certainly not a set satisfying a restriction like the original (2.1). Further, one sees that the class of posteriors exhibits a wider range of contamination than the priors. If a further observation is taken, with this class of posteriors being used as the priors, an even stranger class results, etc. (Of course, with probability one as this process continues, $\theta - \bar{x} \rightarrow 0$ in probability uniformly over the original class of priors.) From several points of view it would be more reassuring if an efficient and plausible model of prior indeterminacy could be found which would be stable under repeated sampling. Is there such a model?

C. The third issue I want to raise concerns Example 10 in Section 6.3.C. Let me preface my remarks with a general statement. The reasoning throughout Berger's paper is inductive. Certain special examples (i.e. the Stein Effect) are examined in detail. Something is found to be true in the example (i.e. a convenient and reasonably efficient robust Bayesian procedure can be found). A general conclusion is then magnified from the specific example to a broader context (i.e. "Bayesian robustness will be valuable in less ideal situations"). Even when the example is carefully chosen and carefully presented there is always a danger in this type of analysis that an inaccurate conclusion will be drawn because the example does not represent all instances represented in the general conclusion. I think this has happened in Example 10.

The conclusion drawn from the example is that, "[ordinary] prior distributions prove to be rich enough to reflect whatever is reasonably desired". It may be that the example shows that ordinary prior to posterior calculations suffice when it is desired only to draw inferences about potential future independent identically distributed observations on the random variable which has been observed in the sample. (I believe this to be true, though of course isolated examples such as Example 10 do not really provide a precise formulation or proof.) However other things may be "reasonably desired". The following scenario gives one possible instance. I presume there are many others.

Suppose the two drugs in Example 10 were two different drugs which are generally recognized to be equally effective (e.g. Aspirin and Tylenol for reducing a fever). A priori you entertain two concepts: (α) Since they are recognized after long experience to be equally effective they must indeed be so, or (β) general experience is very imprecise, so they may very well differ in effectiveness by a moderate amount.

For simplicity let us say conception (α) gives probability 1 to the value $p_M = \frac{1}{2}$. (A prior tightly concentrated in the region about $\frac{1}{2}$ would be more realistic, but the preceding is easier to calculate with and the qualitative conclusions are the same.) Let us suppose conception (β) gives probability $\frac{1}{3}$ to each of $p_M = \frac{2}{5}, \frac{1}{2}, \frac{3}{5}$. It would undoubtedly be very hard to provide a confident statement concerning the relative probabilities of (α) and (β). (This of course is one of the reasons compelling the use of robust Bayesian analyses.) But suppose you could arrive at a figure of, say, 0.3 for the probability of (β). The ordinary prior, p, which corresponds to this two-tiered scenario is one which gives probabilities 0.1, 0.8, 0.1 to $p_M = \frac{2}{5}, \frac{1}{2}, \frac{3}{5}$, respectively.

If you then observe that Drug 1 works better in 5 of 5 trials you may indeed use an ordinary Bayesian analysis with this prior. The resulting posterior probabilities are approximately 0.03, 0.74 and 0.23 for $p_M = \frac{2}{5}, \frac{1}{2}, \frac{3}{5}$. You can correctly state within this Bayesian framework that the probability is $0.03(\frac{2}{5})+0.74(\frac{1}{2})+0.23(\frac{3}{5})=0.52$ that Drug 1 will work better on the next patient than will Drug 2.

Suppose that a second similar question now arises (concerning, say, two medicines for athletes' foot). If you consider *only* the ordinary prior P, derived previously, you would, I think, be led to the decision that relative effectiveness of Desitin has nothing in common with the relative effectiveness of Aspirin tested earlier and so the prior P is again appropriate for the athletes' foot problem.

However, consideration of the two-stage prior and earlier experimental results indicate a probability on (β) of 0.56. This yields new probabilities P', in the athletes' foot problem of 0.19, 0.62, 0.19 for the values $p'_M=\frac{2}{5}, \frac{1}{2}, \frac{3}{5}$. Passage from the ordinary distribution P in the first problem to the ordinary distribution P' in the second is inconsistent solely on the basis of the ordinary distributions P, P' and the 5 of 5 observation. It only becomes consistent behavior when one permits use of two-tiered distributions. (There is an alternate method of analysing the preceding situation. One can describe an a priori joint distribution of p_M and p'_M in which p_M and p'_M are correlated and each has marginal distribution P. However this method is ultimately more complex than the two-tiered method. In principle it requires a priori specification of the joint distribution of parameters for all potential statistical problems, and simultaneously hides the two-tier structure upon which this specification can sensibly and economically be constructed.)

Let me close these comments with a final remark concerning my fundamental agreement with Berger's arguments. I note that in Section 3.2 Berger remarks that "a statistician should be responsible for the long run performance of his methodology". He also observes in Section 4.4, concerning robust Bayesian procedures, that "their long run frequency performance is definitely relevant". These are reasonable paraphrases of the frequentist position I have tried to describe in the first part of my comments. The major difference is only one of its emphasis within the overall theory. I place this responsibility for long run performance as the primary goal and test for any further theory. Instead, Berger places Assumption I first and only later comes to the frequentist test. We both agree on Assumption II and on the consequent importance of developing robust Bayesian methodologies; and this of course is the main lesson.

References

[1] BONDAR, J. (1977) A conditional confidence principle. *Ann. Statist.* **5**, 881–891.

[2] EATON, M. (1982) A method for evaluating improper prior distributions. *Third Purdue Symp. on Statist. Dec. Theory I*, 329–352.

[3] GATSONIS, C. (1982a) Deriving posterior distributions for a normal mean—a decision theoretic approach. Technical report, Rutgers University. Submitted for publication.

[4] GATSONIS, C. (1982b) Deriving posterior distributions for a location parameter. Technical report, Rutgers University. Submitted for publication.

[5] HEATH, D. and SUDDERTH, W. (1978) On finitely additive priors, coherence, and extended admissibility. *Ann. Statist.* **6**, 333–345.

[6] MILL, J.S., *Utilitarianism.* (Quoted in Castell, A. (1943) *An Introduction to Modern Philosophy.* Macmillan Co. (see p. 326).)

[7] ROBINSON, N. (1979a) Conditional properties of statistical procedures. *Ann. Statist.* **7**, 742–755.

[8] ROBINSON, N. (1979b) Conditional properties of statistical procedures for location and scale parameters. *Ann. Statist.* **7**, 756–771.

[9] WALD, A. (1950) *Statistical Decision Functions.* Wiley, New York.

[10] WOLFOWITZ, J. (1951) On ε-complete classes of decision functions. *Ann. Math. Statist.* **22**, 461–465.

2. Comment by *Bruce M. Hill*[*]

There are many important and interesting questions raised by this article. The primary question that I would like to discuss concerns the approach to robustness that Jim Berger recommends. This seems to be a compromise between frequentist concepts, such as minimaxity, and the Bayes risk approach. Although I think this can be a useful way to formulate the robustness question, I think that it is important also to understand the subjective Bayesian content implicit in such a formulation. For me it is pleasant but not quite enough to know that δ^* of page 44 is minimax and also has the Bayes risk values of Table 2. I would like also to know what underlying subjective probability assessments lead to δ^* as an approximation to my posterior expectation of θ. This is not just a matter of idle curiosity because I hope, by such considerations, to learn also when δ^* is not a good approximation and then how to improve upon it. On the one hand this point of view reflects the ordinary subjective Bayesian concern to bring to bear as much as possible of the relevant information, so that things fit together. On the other hand, it reflects a concern that if there were not such subjectivistic content, then the apparent advantages of δ^* might prove to be illusory.

The way that I would try to understand the situation is to think in terms of a mixture model corresponding to two or more hypotheses. Let H_1 be a hypothesis about the world such that given H_1, one would approximately describe his opinions about θ by the prior density π^N of Berger. This might be viewed, more generally, as a prior distribution for θ that one would use as a first stab at the problem, or under 'ordinary' circumstances. But one would not want to pretend that such a distribution ever completely describes ones opinions and so we might want to think, perhaps only informally, about alternative real world hypotheses to which one attaches some credibility and conditional upon which one might have a very different opinion about θ. In particular it seems often appropriate to choose the prior distribution for θ, given not H_1, to be more diffuse than given H_1. This is not always the case, but often occurs because H_1 is a relatively clearly formulated hypothesis, while not H_1 has not yet been clearly delineated, and thus corresponds to a more confused situation. See for example the discussion of 'statistic acid' by L.J. Savage (1961, p. 4.1). It is important to observe that in this approach the components of θ may be conditionally independent, given a hypothesis such as H_1, but unless $\Pr(H_1) = 1$ or 0 they are marginally dependent. As discussed in Hill (1974, Section 4) there are some interesting philosophical and mathematical aspects of the problem seen from this vantage point, and the Stein 'paradox' disappears.

The estimator δ^* behaves roughly like the posterior expectation of θ under the mixture model. If π_2 is the prior density for θ under H_2, and if π_2 satisfies very weak conditions that reflect the relative diffuseness of opinion about θ, given H_2, then as $\|x\| \uparrow \infty$, the posterior probability of H_2 goes to 1 and

[*] University of Michigan.

$E(\theta \mid x) \sim E(\theta \mid H_2, x) \sim x$. See Hill (1974, Sections 4 and 7) for precise conditions under which this occurs. On the other hand sufficiently small values of $\|x\|$ tend to support H_1 (although in a more limited way) and thus tend to yield $E(\theta \mid x)$ close to $E(\theta \mid H_1, x)$. For me this is the essential part of the behavior of δ^*, with the precise cutoff point $4(p-2)$ of no real importance, and in fact even misleading since at best such a cutoff point must reflect very complicated subjective considerations. Even from a risk function point of view it now becomes clear how, under this mixture model, δ^* may be seriously inadequate. For in the transition zone, between the large values of $\|x\|$ that strongly support H_2 and the small values of $\|x\|$ that tend to support H_1, there will be an interval of values of $\|x\|$ such that both H_1 and H_2 will be given substantial posterior probability. In turn this suggests that the risk function for δ^* is such that real improvement should be possible for 'moderate' values of $\|\theta\|$. From a subjective Bayesian point of view this stems from the fact that δ^* 'acts' as though one or the other of H_1 and H_2 were 'true', just as conventional significance tests either reject or do not reject a null hypothesis and yet the data may be such that the posterior probability of each hypothesis may be substantial. A related circumstance arises in the Bayesian analysis of random effects models (Hill (1980, p. 203)), where no matter how small MSB/MSW may be, the posterior expectation of θ_i gives nonnegligible weight to the corresponding row average $Y_{i\cdot}$.

All in all I expect there is very little real disagreement between Berger and myself. I tend to think there are aspects of the subjective Bayesian approach that are more important than either Bayes or frequentist risk properties, but I would not want to ignore the risk function either. I think there are plausible and convenient families of prior distributions for the Stein problem as in Hill (1974, Remark 8, pp. 572, 578) and Hill (1980) and that the analysis is not so messy as Berger says (second last paragraph of Section 4), so I would tend to carry the Bayesian analysis somewhat further than he does, but of course one has to stop somewhere. Where we differ most is perhaps with regard to the desirability of self-imposed restrictions on the way in which one formulates and solves problems. I try to see things from as many different perspectives as I can in order to be as sure as I can that I have not lost something in my formulation and that all the pieces do fit together. For me a robust Bayesian is simply one who thinks carefully about his opinions, to whatever degree he thinks is appropriate, and then analyzes the data to reflect such considerations.

References

[1] SAVAGE, L.J. (1961) The subjective basis of statistical practice. The University of Michigan (unpublished notes).

[2] HILL, B.M. (1975) On coherence, inadmissibility and inference about many parameters in the theory of least squares, In: S. Fienberg and A. Zellner, eds., *Studies in Bayesian Econometrics and Statistics.* North-Holland, Amsterdam.

[3] HILL, B.M. (1980) Robust analysis of the random model and weighted least squares regression. In: J. Kmenta and J. Ramsey, eds., *Evaluation of Econometric Models.* Academic Press, New York, pp. 197–217.

3. Comment by *Joseph B. Kadane**

I have little trouble with the idea of a class Γ of prior distribution and of studying how the likelihood function transforms each to a posterior distribution in some class Γ^*. I do have trouble, however, with the idea that one can change Γ after seeing the data. By allowing such a change, nearly every posterior class Γ^* can be obtained with nearly every data set. If the point is only to identify sensitive areas in which the posterior depends critically on features of the prior not carefully considered beforehand, then I am less troubled. But Berger proposes to go beyond this, without setting limits to the amount of change in Γ he would permit.

What kind of reporting would he suggest? Is the original Γ relevant to a reader, in Berger's view? If the Bayesian viewpoint is to maintain its vitality, it must insist on limits to the kinds of change permitted in Γ before transformation to Γ^*. Perhaps some applied work providing examples of inference done in Berger's style would be enlightening.

Can a set of likelihoods be far behind?

*Carnegie-Mellon University.

4. Comment by *Dennis V. Lindley*[*]

This admirable paper with its fine bibliography is so crowded with important ideas that it is hard to know where to begin any commentary. Assumption I is the cornerstone of the argument, though to me it is not an assumption but a consequence that follows from other, much simpler and intuitively appealing assumptions: essentially justification (v). I would express the cornerstone rather differently. To me, it says that any inferential or decision procedure must be equivalent to one that uses probabilities for all uncertain quantities—and if decisions are involved uses utilities and the maximization of expected utility. You can obtain the procedure how you wish but, to be sensible, it must match a Bayesian procedure. (It is for this reason that all samplingtheory procedures must be rejected; because there is no prior that could possibly reproduce them.) There is no obligation to take a prior and a likelihood, multiply, normalize and to obtain a posterior: you may do it as you wish only observing the rules of the probability calculus. In fact, it seems easiest to do it in the usual way: but see Sturrock (1973).

In this view, what role does robustness play? Simply that one does not want to be in a situation where a small shift in a probability here causes a large shift in a probability there. (This remark applies equally to likelihoods as priors: we have got so wedded to the idea of likelihoods being 'known' and priors not that it is hard to be impartial between them.) Whether this consideration is expressed by posterior or procedure robustness depends on whether or not the data are to hand.

In the illuminating normal-Cauchy example (Example 4), if the Cauchy prior is used, numerically large values of θ will occur which, because of the normal likelihood, will produce numerically large values of x. If the normal prior is used, these values will not occur and the normal procedure will not take them into account: that is, it will do extremely badly with the large values. Consequently, it will be sensible to guard against them by using the Cauchy prior. Similar considerations apply to the likelihood, a Cauchy one being more robust than normal; illustrating the point that it is only the probability calculus governing both prior and likelihood that matters.

A vigorous protest at the use of minimax ideas is in order. A minimax procedure does not accord with any Bayes procedure and to use it as an antidote for robustness is to invite unsatisfactory procedures. If elicitation leads to ambiguities, then further refinement seems appropriate. I met an example recently where the decision maker had to provide a mean and a variance: Lindley (1983). The former presented no problem but the decision maker only felt comfortable with a range of variances, so a distribution was placed on the variance. This is surely the better way out of the problem of a class of priors: put a prior over the class. Or, expressed differently, use a hierarchical model. If

[*]Minehead, England.

normal or Cauchy seems doubtful put probability α on one and $1-\alpha$ on the other: or better use t with a distribution on the degrees of freedom. It should always be remembered that we do not need $\pi(\theta)$ but, in Jeffreys' notation, $\pi(\theta \mid H)$ and that H can change by introspection or deeper elicitation, as with the variance.

My own view about Assumption II is that we should learn to measure probabilities. Physicists, presented with Newtonian mechanics for the first time, did not dismiss it because they could not measure accelerations; they learnt to do so. Surveyors do not deplore Euclidean geometry because they cannot measure distances without error: they use techniques like least-squares. And they discover that angles are easier to 'elicit' than distances: perhaps log-odds are better than probabilities. We need to recognize real man trying to imitate normative man and to develop the equivalent of the surveyor's least squares. This view has been described in Lindley, Tversky and Brown (1979).

There are many references in the paper to 'long-run frequency' properties. The difficulty with these is to decide what 'long-run' is appropriate. We are typically interested in an inference or decision that is specific to a single occasion when 'long-run' ideas are irrelevant. If the decision is part of an obvious sequence, as with routine quality control, then the 'long-run' idea is germane but is adequately considered within the personalistic, nonrepetitive framework when discussion of the sampling plan (rather than the handling of the data) arises. Probability has nothing to do with 'long-run': it is only when allied to exchangeability that frequency ideas surface and then the relevant 'long-run' is well-specified as being the set judged exchangeable. It is hard for those trained in the Berkeley tradition to discard the inappropriate ideas when changing to the Bayesian paradigm. It took me 25 years and even now I have to be careful not to be beguiled by some apparently plausible argument. Strict adherence to the alternative canon always provides a rescue.

References

[1] LINDLEY, D.V. (1983) Reconciliation of probability distributions. To appear in *Operations Research.*

[2] LINDLEY, D.V., TVERSKY, A. and BROWN, R.V. (1979) On the reconciliation of probability assessments (with discussion). *J. Roy. Statist. Soc. Ser. A* **142**, 146–180.

[3] STURROCK, P.A. (1973) Evaluation of astrophysical hypotheses. *Astrophys. J.* **182**, 569–580.

5. Reply to the comments by *Jim Berger*

It is a privilege to have four such distinguished statisticians as discussants for the paper. Their comments are extremely valuable in illustrating different related perspectives and in focusing a number of key issues. Foremost among these issues is that of the use of frequency measures in Bayesian robustness. Since this is perhaps the most involved and unclear consideration, I will delay discussion of it until the end, first dealing with the other issues raised.

A number of Professor Brown's comments are addressed to important questions concerning the development and implementation of robust Bayesian methods. He makes the crucial observation that "when posterior robustness occurs even a dogmatic frequentist should act as if he were a dogmatic Bayesian" (assuming of course that the posterior robustness is obtained for a realistically large class Γ of priors), but points out the need for theoretical development to deal with situations lacking posterior robustness. His proposal for examining the probability that posterior robustness obtains is interesting, since it is certainly important to get some feel as to the extent of the problem of a lack of robustness.

Professor Brown also expresses concern over the choice of Γ. This is a crucial and delicate problem, since there will constantly be the competing desires to choose a large Γ to ensure that nothing is left out and to choose a small Γ to make posterior robustness more attainable. The ε-contamination class in (2.1) will usually be large enough (for suitable ε) to provide a feeling of security, but it includes many undoubtedly unreasonable priors that could destroy posterior robustness. In this respect, I view the process of robust Bayesian analysis from the somewhat data interactive viewpoint in which one starts with a perhaps crudely large Γ, checks for posterior robustness, and progressively refines Γ (if needed) until posterior robustness is achieved. (Some further comments related to this will be given later.) In any case, finding suitably rich and reasonably easy to work with Γ is important.

In desiring a class Γ which is stable (after transformation to a class of posteriors) under repeated sampling, Professor Brown is probably asking for too much. As he points out, when $n \to \infty$ the class of posteriors will usually degenerate to a point mass at θ, and it is rarely reasonable to have all point masses as priors in the original Γ. Also, the class of posteriors resulting from (2.1) will evidence a wider range of contamination than the original Γ only when the data is itself inconclusive and raises doubts about the plausibility of π_A.

I found Brown's discussion of Example 10 quite interesting and have no real disagreement with it (except that my calculator would not reproduce a probability of 0.56 for (β)). Indeed, the discussion actually seems to support the contention that the prior distributions provide a rich enough structure. If I feel that the aspirin–tylenol experiment tells me something about the situation in the second experiment (through its effect on my prior conceptions concerning the validity of 'general experience'), then I should indeed attempt to incorporate this information. Attempts to do this could, as Brown mentions, take the form of a complicated joint a priori specification for both situations or take the form of a two stage prior. I have nothing against hierarchical priors as conceptual tools, even though they always correspond to single priors. (Perhaps I misleadingly came across in Example 10 as recommending only consideration of single stage priors.) Thus Brown's discussion reinforces the contention that, when the *need* to involve certain information is recognized, a Bayesian will be forced to consider an appropriately rich prior structure.

Professor Kadane expressed concern about allowing modification of Γ after seeing the data. I share the concern but, as argued in the paper, feel there is little choice. In a large economic model or weather model there may be hundreds of variables, and any attempt to construct accurate priors before seeing the data seems hopeless and a possibly great waste of time. (Only some of the variables may turn out to be important.) Furthermore, scientific surprises and correctable modeling errors must be allowed for. (When looking at the data it is not at all uncommon to realize that the model or prior being considered is clearly inappropriate due to some oversight.) On the other hand, indiscriminate changing of Γ is clearly unacceptable. I made no attempt to say what changes are, and are not, permissible, for two reasons. First it is unlikely that such limits could be codified. Second, in a certain sense, the problem seems moot, precisely becasue of reporting requirements. As an outsider evaluating a statistical analysis I will look at Γ and the likelihood function and decide if they seem reasonable. Where Γ came from is almost immaterial: either it is plausible or not. Thus a prior chosen simply to be 'compatible' with the data will hopefully look suspicious. There is obviously a certain danger in allowing post data modification of Γ, but it seems a necessary evil in attaining realism.

Professor Kadane also mentions that perhaps a set of likelihoods, not just priors, is in order. Certainly this is true, and the brief discussion in Section 4.3 is admittedly inadequate. Although one can formally subsume uncertainty in the likelihood into uncertainty in the prior (just enlarge Θ), the paper essentially ignores model robustness. This was done in an attempt to keep the paper within reasonable bounds, but was perhaps an error. As Professor Lindley comments, treating the model and the prior separately is often a mistake—they are both just an attempt to impose a subjective structure on the situation.

I was rather surprised at Professor Lindley's expression of the cornerstone of Bayesian analysis. From his statement, he would apparently have no objection to a frequentist decision theorist choosing any admissible decision rule, since such usually correspond to Bayes rules. I presume Lindley left out, for the sake of brevity, the qualification that the prior to which the chosen decision rule corresponds had better be an accurate representation of subjective beliefs. I would, in any case, argue (see Berger (1982d)) that the conditional approach to Assumption I tends to be more convincing than the rationality approach, but the more reasons for being a Bayesian, the better.

Professor Lindley comments that the impact of Assumption II can be reduced by developing improved methods of eliciting subjective probabilities. While certainly true, this doesn't mean that Assumption II should be ignored. The smaller Γ is, the less Bayesian robustness will be a concern, but a nonsingular Γ will always be present. Perhaps Lindley was trying to say that, even if Assumption II is true, it does not follow that the Bayesian perspective is wrong, with which I would, of course, wholeheartedly agree.

Turning finally to the main issue raised by the discussants, namely the validity of the use of frequency measures, I will begin with some thoughts on Professor Brown's very interesting position. His logical and admirable presentation that even a frequentist decision theorist should be highly interested in posterior Bayesian robustness will, I hope, have a significant impact on the frequentist school of statistics. Furthermore, it appears that our practical positions may be close to identical: one should strive to attain posterior robustness and (possibly) involve frequency calculations only when posterior robustness is unattainable. Further delineation of our differences might thus seem to be mere quibbling, but I feel that the difference between basing one's outlook on Assumption I or on Brown's frequency principle does have a profound effect on actual statistical practice, particularly in the extent to which one becomes satisfied in an actual investigation that posterior robustness obtains. (Formal verification of posterior robustness over a class such as (2.1) will frequently be impossible.)

Professor Brown views Assumption I as somewhat imprecise and indeed it is only later in the paper that my views on the subject are stated more clearly. My interpretation of Assumption I is essentially that the only way to ascribe meaning to a set of data is to see its effect on one's prior opinions (in which I include specification of the model) through the prior to posterior transformation. Thus there is a set of possible 'meanings' corresponding to the set of posteriors obtained from Γ. The reasons for saying that these are the only valid meanings are (in my opinion) the conditionality arguments given in some detail in Berger and Wolpert (1982). Though this is not the place for a discussion of such arguments, I cannot resist one example.

Example. Suppose X is to be observed and is known to have either the distribution P_0 or P_1 given by Table 1.

Table 1

$x =$	1	2	3
P_0	0.005	0.005	0.99
P_1	0.00501	0.98499	0.01

What 'meaning' should be given to the observation $x = 1$. From a common sense viewpoint this observation indicates virtually nothing concerning the truth of P_0 or P_1. I know of no way to say this in a frequency sense, however. The most powerful $\alpha = 0.01$ level test of P_0 versus P_1 (which also has probability of Type II error equal to 0.01) fails to distinguish in meaning between the observations $x = 1$ and $x = 2$. And, since conditional frequency approaches are untenable here (a 3-point sample space cannot be divided into two parts on each of which a conditional frequency analysis can be performed), it seems that there is no frequency resolution to the difficulty. Even embedding the problem in a decision theoretic setting will not help. If one must either decide P_0 or P_1 under, say, 0–1 loss, then the *procedure* of basing one's decision on the most powerful $\alpha = 0.01$ level test has the very attractive risk 0.01, which seems completely misleading when $x = 1$ is actually observed (and P_1 is then concluded to be true). Again, there is no conditional frequentist decision theoretic solution to this inadequacy. Of course, the chance of observing $x = 1$ is very small (good frequentist procedures do not give obviously silly conclusions with high probability), but nevertheless the view that one should always have a valid frequency interpretation of a conclusion seems suspect.

The above example demonstrates the important point that, even though there may be a 1–1 correspondence between admissible frequentist decision procedures and Bayes procedures, there can be a major difference in what is reported as the accuracy of the procedure (the frequentist risk or the posterior expected loss). It could, of course, be argued that if the 'real' accuracy of the procedure is important to know because of further decisions that might have to be made, then these further decisions should have been incorporated in an expanded original decision problem. This path out of the dilemma quickly leads to the practical absurdity, however, of having to imagine and simultaneously solve all future decision problems which will be faced. Unless the contradictions between frequency analysis and what could be termed 'conditional common sense' can be resolved, I do not see how the frequency principle can be used as the foundation for statistics.

While rejecting frequency criteria as fundamental, I am certainly willing to defend them as sometimes valuable tools for use in achieving the conditional Bayesian goal. In this light, let me now turn to the criticisms (or at least warnings) concerning the use of frequency measures given in the discussions of Professors Hill and Lindley. It will be assumed in the following that posterior robustness is found to be lacking (for Γ), and the issue is how then to proceed.

It is useful to begin with a statement of the *philosophical* situation. Ideally, Γ has been constructed through utilization of *all* available subjective information. When this is the case, Bayesian techniques of dealing with Γ (such as putting a metaprior on Γ) have no rational basis; if all subjective information has really been utilized, any metaprior put on Γ is completely arbitrary. In such a vacuum, frequency measures are as reasonable as anything else as a basis for proceeding and may even be preferred for reasons such as those in Section 4.4(B). I do *not*, however, give blanket endorsement to the use of frequency measures in such situations. Essentially one is in a state of irresolvable ignorance and I am not so bold as to propose a foolproof method of resolving irresolvable ignorance.

Although the philosophical scenario above is important for conceptual reasons, it does not really settle the issue. In reality, Γ will never be a finalized summary of all subjective information and there may be good pragmatic reasons for dealing with Γ in a Bayesian fashion as proposed by Hill and Lindley. For instance, after formulation of Γ, a frequently remaining subjective belief may be that the elements of Γ are roughly thought to be 'equally likely' to represent the 'true prior'. This feeling would provide at least a partial justification for putting a 'uniform' metaprior on Γ. There are a number of potential problems in attempting to do this, but the idea is appealing. An even easier approach to dealing with the 'equally likely' situation is to choose (after observing x) that prior in Γ which maximizes the marginal distribution $m(x)$. (This is actually related to the use of a 'uniform' metaprior, the prior maximizing $m(x)$ being interpretable as the most likely prior, a posteriori, for the uniform metaprior.) There may, of course, be other kinds of 'residual' subjective information after Γ has been specified, such as smoothness constraints on the priors. Different Bayesian methods may be useful in dealing with such cases.

On the other hand, there may be substantial pragmatic reasons for proceeding in a frequentist fashion. In particular, if a Bayesian approach is technically too difficult (or too difficult for the ability level of the user) and an easier frequency analysis is possible and can be given some kind of Bayesian motivation (as discussed in Section 4.4), then there is justification for adopting the frequency analysis.

The example discussed in Section 4.5, and referred to in Professor Hill's discussion, is a good case in point. I am entirely in sympathy with Hill's attempt to intuitively understand δ^* from a Bayesian perspective and indeed essentially agree that his understanding is the right one. However, in actually implementing this understanding, it is necessary to choose some particular prior on Hill's H_1 and this is not easy to do. For instance, there is no clear reason to choose a prior on H_1 that destroys the independence of the θ_i. (For the Γ in (4.4), for instance, if independent metapriors are put on the π_i, then, in the resulting Bayes rule, $\delta_i^\pi(x)$ will only depend on x_i.) There are metapriors (leading to particular priors) which give estimators similar to δ^*, but they are not significantly more intuitively plausible than others which lead to estimators much less attractive than δ^*. Of course, attractive is here being judged relative to the Γ in (4.4), which may be unreasonably large (in allowing arbitrary P_i), but the fact that one can get such attractive behavior for such a large Γ by frequency methods also supports the pragmatic justification for allowance of frequency methods. (Using a too large Γ may entail a sacrifice of significant possible gain that could result from a realistically smaller specification—it is in this light that I understand and agree with Hill's concerns about the subjective basis of δ^*.) Also, in saying that δ^* has 'attractive' performance, I am implicitly including its conditional performance, based on the argument that if an estimator has such good $r(\pi, \delta^*)$ for all $\pi \in \Gamma$, then its conditional Bayesian posterior performance (which is of paramount importance) must also be pretty good.

I must admit, however, to not being entirely convinced as to the ultimate pragmatic necessity of employing frequentist measures in Bayesian robustness. I am continually surprised at the success of even very simple Bayesian methods (such as choosing a prior in Γ based on maximizing $m(x)$), at the substantial degree of robustness typically attained by Bayes rules resulting from ad hoc metapriors on Γ and at how forcing further refinement or involvement with Γ is generally more revealing than switching to a frequentist measure. The existing examples where involvement of a frequentist measure proves easier may essentially be due to the extensive frequentist theory upon which we can draw. Of course, there is no reason not to draw on this theory when it does prove helpful, while at the same time encouraging the more natural Bayesian approaches. Also, the frequentist viewpoint is one of the 'different perspectives' from which a problem can be viewed and may sometimes provide insights as argued in Section 4.4.

In the above light, it is interesting to consider Professor Lindley's final comments. He is certainly correct about the difficulty in overcoming an inappropriate tradition and I would be happy to support the position that strict adherence to the alternative canon (Bayesian analysis) *usually* provides a rescue (providing robustness is considered). I am uncomfortable with 'always', however, and feel that there are sufficient philosophical and pragmatic grounds to justify sometimes leaving the canon, providing this departure can be given some justification in terms of the canon.

In conclusion, I would like to thank the discussants for their stimulating comments, resulting in further clarification of my thoughts (and hopefully my comments) concerning the robust Bayesian viewpoint. Our areas of agreement strike me as being much larger than our areas of disagreement, and these discussions hopefully bring us even closer.