Bayesian Salesmanship[1]

by

James Berger
Purdue University

Technical Report #82-39

Department of Statistics
Purdue University
West Lafayette, IN

November 1982

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

Abstract

Of the many arguments for the Bayesian viewpoint, the conditionality-Likelihood Principle approach seems most effective in "converting" non-Bayesians. When combined with a robust Bayesian perspective, the resulting package can be very convincing to non-Bayesians and answers most of their objections to Bayesian analysis. It also indicates which frequency based (or other classical) modes of analysis have a role to play in Bayesian analysis.

1. Introduction

Bayesian statistics is growing rapidly and vigorously. By now there are extensive Bayesian methods available for use as alternatives to classical statistical techniques. However, non-Bayesian statistics seems to also be growing rapidly, and while there seems to be some trend towards the Bayesian position (c.f. Zellner (1981)), this trend is not yet overpowering. Thus, although the arguments have been going on for over a century, it remains imperative for Bayesians to continue to sell their viewpoint effectively and to attempt to improve their sales techniques.

To relative newcomers in statistics the Bayesian approach seems natural and believable. (It is certainly more natural to talk about actual "probabilities" of hypotheses than about error probabilities of various types.) This article is addressed, not to the problem of convincing such newcomers, but instead to the far more difficult problem of changing the views of established non-Bayesians. We will argue that this can best be accomplished by (i) Arguing for the need to think conditionally on the data and follow (as much as possible) the Likelihood Principle; (ii) Maintaining that, while consideration of prior distributions is necessary for meaningful conditional thought, there is no need to presume (and indeed reasons not to presume) that prior distributions can be specified completely or accurately; (iii) Admitting that, for a variety of reasons, non-Bayesian techniques are (at least for the time being) of interest to a Bayesian.

The more standard arguments for the Bayesian position are based on notions of coherency (or rationality or consistency), and on the importance of using prior information. Although these arguments can be persuasive, we

will argue in section 2 that they are not sufficiently compelling to the majority of statisticians. Of course, statisticians never really accept the Bayesian position until they actually try thinking in a Bayesian fashion and see the clarity and improved results that are usually obtained. It is very hard to "prove" to a non-Bayesian that Bayesian results are really better, however (or, more precisely, that in the substantial majority of problems a Bayesian analysis will give a better answer than a non-Bayesian analysis based on a similar investment of effort). Thus the "success" of Bayesian analysis is typically the gradually perceived clinching argument for the Bayesian viewpoint, rather than the initial argument starting one on the Bayesian path. Also, the perhaps most important component of the Bayesian mode of thought is conditional thinking, which provides further justification for arguing via (i), (ii), and (iii).

There is, of course, nothing new in arguing the Bayesian viewpoint via steps (i), (ii), and (iii), and indeed there will be nothing really new anywhere in the paper. Also, serious efforts will not be made to justify these three steps here; instead the emphasis will be on why this is the best approach to take. Finally, very few references will be given, since the concepts discussed have been developed by a large number of people and this is not intended to be a review paper. Extensive bibliographies concerning some of these issues can be found in Berger (1984) and Berger and Wolpert (1984). As a final caveat, although we repeatedly refer to 'non-Bayesians' as a class, there are, of course, enormous differences in view among non-Bayesians. The arguments we suggest are particularly aimed at frequentists, whose techniques form the bulk of existing statistical methodology. In particular, such theories as fiducial inference (c.f. Wilkinson (1977)), structural inference (c.f. Fraser (1968)), and pivotal inference (c.f. Barnard (1980)) will be ignored, even though they may have major points of disagreement with the advocated form of Bayesian analysis and even though they may

important contributions to the advocated form of Bayesian analysis. Finally, lack of space will also preclude discussion of various pseudo-Bayesian positions. See Berger (1984) and Berger and Wolpert (1984) for some discussion and references.

The notation that will be used in the paper will be kept simple. It will be assumed that an experiment

$$\mathcal{E} = (X, \{f_\theta: \quad \theta \in \Theta\})$$

is performed, where X is the random variable observed (a particular realization of which will be denoted x), f_θ is the density of X on the sample space $\mathcal{X}$, and θ is the unknown parameter of interest with Θ being the parameter space. (Since we seek to establish nothing rigorously, mathematical niceties will be ignored and densities will be presumed to exist.) Although θ will usually be a parameter in the usual sense, we will have occasion to allow it to be just a general index, such as when $\{f_\theta: \quad \theta \in \Theta\}$ is to be the set of all densities on $\mathcal{X}$ (with respect to an appropriate measure). Prior densities on Θ will be denoted $\pi(\theta)$, and $\pi(\theta|x)$ will denote the resulting posterior density. Although some Bayesians view this standard "model-prior" formulation as misguided fundamentally (c.f. deFinetti (1974,1975), who views only observables as fundamental), we feel that such a formulation is usually an operational necessity, and, in any case, the Bayesian viewpoint can best be sold to non-Bayesians by using their frameowrk in so far as possible.

2. Helpful, but Non-Compelling Bayesian Arguments

Several arguments often used to support the Bayesian position are mentioned here, along with reasons why they seem not to be compelling to most non-Bayesians.

2.1 Prior Information Is Important

Examples abound where prior information is a crucial element of any statistical problem. Indeed, in virtually any realistically complicated statistical setting, massive subjective choices (model assumptions, etc.) are made by any statistician. Even in such a supposedly pure non-Bayesian procedure as a significance test of a single (point) hypothesis, selection of a test statistic requires subjective consideration of the type of alternatives that are of concern. Thus Good (1983) has said

> "The people who don't know they are Bayesians are called non-Bayesians."

Nevertheless, this argument fails to be persuasive to non-Bayesians. They can argue that the model often does have an "objective" reality due to theoretical considerations, and in some situations (such as that of simple Bernoulli trials) this may be the case. We won't pursue this issue, basically because of the empirically observed fact that arguments concerning the lack of objectivity of models seem to have little effect on non-Bayesians.

Non-Bayesians can also, at least, argue that their nonparametric analyses are totally objective, as in the following example (from Efron (1982)) which will be of use later.

Example 1. The experiment $\mathcal{E}$ consists of observing $X_1,\ldots,X_{15}$ which are i.i.d. observations from a completely unknown continuous density f on R^1.

(Here we identify θ with the unknown f, so Θ is the set of all continuous densities on R^1.) Of interest is η, the median of the unknown density. A simple binomial calculation shows that a 96.3% confidence interval for η is given by $[X_{(3)}, X_{(12)}]$, where the $X_{(i)}$ are the order statistics.

2.2 Coherency

Many axiomatic systems of coherent (or consistent or rational) behavior (c.f. Ramsey (1926, 1928), deFinetti (1974, 1975), Savage (1954), and Rubin (1974)) have been created, which show that any coherent method of behavior corresponds to some form of Bayesian analysis. It was indeed these axiomatic developments that convinced many of today's Bayesians to become Bayesians, and they remain a cogent argument.

Those who remain unconvinced by the coherency arguments offer the following criticisms:

(i) The axioms themselves are wrong (c.f. LeCam (1977)).

(ii) The axiomatic systems are only theoretically implementable, in the sense thay they involve infinite numbers of comparisons or, alternatively, involve anticipation of all possible eventualities. Thus C.A.B. Smith (in Savage et. al. (1962)) says

> "Such an absolutely consistent person does not, of course, exist."

(iii) The coherency argument merely states that a coherent method of behavior corresponds to a Bayesian analysis. It doesn't say that one must explicitly use a Bayesian approach to be coherent, as the following example shows.

Example 2. Consider an invariant decision theoretic estimation problem, in which the parameter space Θ is a compact group (say, the group of rotations). For any such problem, the mode of behavior "use the best

invariant estimator" is completely coherent (since it happens to correspond to a Bayesian analysis with the invariant Haar measure on Θ as the prior distribution).

(iv) A Bayesian approach, even though coherent, need not necessarily be better than a non-Bayesian approach. Thus C.A.B. Smith (in Savage et. al. (1962)) says

> "Consistency is not necessarily a virtue: one can be consistently obnoxious."

As an example, for parametric problems where Θ is a compact interval, the formally coherent Bayesian behavior of always using a prior distribution concentrated on the lower endpoint of the interval is clearly obnoxious. A more realistic example is the following.

Example 3. Suppose that $X \sim \mathcal{N}(\theta,1)$ is to be observed, and that a Bayesian determines that his prior density for θ is bell shaped with median 0 and quartiles ± 1. He then chooses the matching conjugate prior ($\mathcal{N}(0,2.19)$) to do the analysis. If now $x = 5$ is observed, the posterior mean can be calculated to be 3.43.

As an estimate of θ this can be challenged by a non-Bayesian, since it is 1.57 (sample) standard deviations from the (MLE and MVUE) classical estimate 5. Indeed, the non-Bayesian can point out that, if the Bayesian had fitted his prior information with a Cauchy density instead, then the Bayes estimate would have been about 4.6. The non-Bayesian concludes that he could care less about coherency; he just wants to get a good estimate, and 5 looks like a pretty good estimate of θ to him.

Of course, in the above example, a good Bayesian analysis would have involved a sensitivity study when the surprising (but not impossible) $x = 5$ was observed. Furthermore, in deciding between use of the normal or Cauchy prior, a Bayesian would note that the ratio of the predictive (or marginal) probabilities of $x = 5$ for the normal and Cauchy priors would be very small, indicating that the Cauchy prior is more appropriate for use. The fact that a very careful Bayesian analysis is needed is precisely the point, however; the non-Bayesian can argue that a "good" coherent Bayesian analysis is harder and more fraught with peril than a "good" non-coherent frequentist analysis. Thus LeCam (1977) says

> "To claim that an ideal person could in principle specify such a relation is to beg the question. To claim that since an ideal person could do it, a real person should do it, is to introduce a dogma for which we have little justification."

In principle, a Bayesian might admit the validity of many of the above objections. On the other hand, it certainly seems desirable to strive for coherent behavior, and the axiomatic systems seem to indicate that the direction in which to strive is the Bayesian direction. Thus the Bayesian can argue that most of the effort in statistics should be directed towards extending and sharpening Bayesian tools.

Another extremely valuable consequence of the coherency argument is that if any "complete" theory of statistics exists, it is the Bayesian theory. One can either be a Bayesian, or argue for some form of anarchy.

2.3 The Correspondence Between "Good" Classical Procedures and Bayes Procedures

When non-Bayesians attempt to define "good" classes of statistical procedures, these classes almost invariably end up corresponding to a class of Bayes procedures (or their limits). Most complete class thoerems in frequentist decision theory are of this form, the first and simplest being

the correspondence between most powerful tests and Bayesian tests, when testing between two simple hypotheses.

Although this correspondence could be considered to be a mathematical coincidence, it is hard to argue that a procedure should be selected from the "good" class even if it corresponds to an unreasonable prior distribution.

An example of current research interest in frequentist theory is that of the Stein effect, whereby the usual estimator of a multivariate normal mean can be improved upon (in terms of frequentist risk) in dimensions three or more. Unfortunately, the set of better estimators is enormous, and in choosing an estimator from this set it is impossible to ignore prior information, at least to the extent of deciding where to shrink towards. (See Berger (1980a and 1982a) for more discussion.) While these arguments are certainly telling, they do not seem to be compelling to non-Bayesians.

2.4 Measuring Uncertainty

A very fundamental justification for Bayesian analysis (c.f. Jeffreys (1961), deFinetti (1974, 1975) and Jaynes (1981) which also contain other references) is simply that the goal of statistics is to communicate evidence about uncertainty, and that the correct language to use in measuring 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. Users of statistics want to know the probability (after seeing the data) that a hypothesis is true, or the probability that θ is in a given interval, and yet classical statistics does not allow one to talk of such things. Instead, artificial concepts such as error probabilities and coverage probabilities are introduced as substitutes. It is ironical that non-Bayesians often claim that the Bayesians form a dogmatic unrealistic religion, when instead it

is the non-Bayesian methods that are often founded on elaborate and artificial structures. Unfortunately, those who become used to these artificial structures come to view them as natural, and hence this line of argument tends to have little effect on the established non-Bayesian.

3. The Conditionality-Likelihood Principle-Robust Bayesian Approach

By far the most troubling aspect of frequentist statistics is its unconditional nature, and this usually provides the easiest point of attack. We briefly outline the argument here.

3.1 Conditioning and the Likelihood Principle

Although most frequentists are aware of conditioning problems, the following two examples, taken from Berger and Wolpert (1984), are simple and revealing.

Example 4. Suppose X_1 and X_2 are independent and identically distributed with $P(X_i = \theta+1) = P(X_i = \theta-1) = 1/2$. (Here $\theta \in \Theta = R^1$.) A 75% confidence set of smallest size for θ is given by

$$C(X_1,X_2) = \begin{cases} \text{the point } \frac{1}{2}(X_1+X_2) & \text{if } X_1 \neq X_2 \\ \text{the point } X_1 - 1 & \text{if } X_1 = X_2 \end{cases} .$$

Notice, however, that when we observe $x_1 \neq x_2$, we are 100% certain that $\theta=\frac{1}{2}(x_1+x_2)$, while if we observe $x_1=x_2$ the data leaves us equally uncertain as to whether θ is x_1+1 or x_1-1. Thus the frequentist 75% confidence arises from the average of the real 100% confidence when $x_1 \neq x_2$ and 50% confidence

when $x_1=x_2$. This example forcefully shows that, at least sometimes, one must think conditionally on the actual data that occurs, rather than use frequentist long run averages.

Example 5. Suppose that a substance to be analyzed can be sent to either a laboratory in New York or a laboratory in California. The two labs seem equally good, so a fair coin is flipped to choose between them. The coin is flipped and comes up tails, which signifies that the California lab is to be used. When the experimental results come back and a conclusion is to be made, should it be taken into account that the coin could have come up heads with the New York lab then being used?

The above example is, of course, a variant of the famous Cox (1958) example, and in symbolic form can be phrased as follows: If with probabilities 1/2 (independent of the unknown quantity θ of interest) either experiment $\mathcal{E}_1$ or experiment $\mathcal{E}_2$ (both pertaining to θ) will be performed, should the analysis depend only on the experiment actually performed or should the possibility of having done the other experiment be taken into account? (Note that we are talking about analysis with the data in hand, after one of the experiments has been done, and not about design, etc.) Uncompromising frequentist theory would suggest that the other experiment must be taken into account, since frequentist measures depend on averaging over all possible experimental outcomes, whether they happen or not. This, of course, strikes almost everyone as nonsensical, and virtually all statisticians would accept that only the experiment actually performed should matter. This has been called (c.f. Basu (1975)) the Weak Conditionality Principle (WCP).

Virtually all frequentists will subscribe to the WCP and admit that

conditioning must sometimes be done, but will then attempt to limit the applicability of conditioning to "relevant subsets" or "ancillary statistics" or "partitions of the sample space". (Brief descriptions and references to such efforts can be found in Berger and Wolpert (1984).) However, Birnbaum (1962) provided a devastating, yet simple, refutation of the viability of limiting the scope of conditioning. He showed that the WCP and the Sufficiency Principle (which states that a sufficient statistic for θ contains all relevant information about θ) together imply the Likelihood Principle, which basically forces conditioning all the way down to the actual observation x. The Likelihood Principle itself traces back to ideas of G. Barnard and R.A. Fisher (see Berger and Wolpert (1984)), and in its basic form is given as follows.

Likelihood Principle. All information or evidence about θ from an experiment $\mathcal{E}$ is contained in the likelihood function for θ, namely

$$\ell_x(\theta) = f_\theta(x).$$

(Here x is the actual observation.) Two likelihood functions (for θ) are equivalent if they are proportional to each other.

A statistician seeking to follow the Likelihood Principle would only trust procedures or measures that depend on $\mathcal{E}$ solely through the observed likelihood function. Classical maximum likelihood estimates are of this form (as are all Bayesian procedures and measures based on the posterior distribution), but frequentist measures such as error probabilities, bias, coverage probability, and P-values, which involve averages over unobserved x, are not of this form, and can hence not be of basic

interest from the conditional viewpoint.

Frequentists can, and will, raise all sorts of objections to the Likelihood Principle and its implications. However, examples and continual application of common sense can be used to counter these objections. (See Basu (1975) and Berger and Wolpert (1984) for a discussion of many common objections. The latter work also shows that powerful versions of the Likelihood Principle hold in essentially complete generality, not requiring densities or even exact knowledge of the model.)

The argument for conditioning leaves one very close to the situation in section 2.4. It basically supports the argument that only the uncertainty about θ for the data, x, actually observed is of interest. For example, coverage probability of a confidence procedure $C(X)$ is something of no intrinsic interest; what is of interest is some measure of how "confident" we are that θ is in $C(x)$ for the actual observed x. (See Example 4.) As Savage succinctly put it in Savage et. al. (1962):

> "The only use I know for a confidence interval is to have confidence in it."

As another example of how artificial frequentist structures can look bad in the light of conditional common sense, consider the following.

Example 6. Frequentist measures are claimed to have the very desirable property of guaranteeing "long run" success rates. For instance, it is claimed to be scientifically desirable to test established theories (the null hypotheses) with classical tests having low probability of Type I error (say $\alpha = .01$), since then one can be assured of incorrectly rejecting only 1% of correct theories. A conditionalist would argue that this is an artificial concept

that does not really measure what is sought, namely the conditional chance that one is making a mistake when the null hypothesis is rejected. To see the difference, suppose that the observation X is either 1 or 2, with probabilities under f_0 and f_1 given by

	X = 1	X = 2
f_0	.01	.99
f_1	.01001	.98999

and it is desired to test H_0: $f = f_0$ versus H_1: $f = f_1$. The most powerful $\alpha = .01$ level test rejects H_0 when $X = 1$ is observed. From the conditional viewpoint, however, when $x = 1$ is actually observed (i.e. rejection of H_0 actually occurs) there is very little evidence against H_0 (because the likelihood ratio is so close to 1), so the chance of being in error is about 1/2 (based on the data alone). Thus $\alpha = .01$ is providing a very misleading and false sense of security when rejection actually occurs.

One can present the case for conditioning, as outlined above, without heavy involvement of subjective prior distributions. As alternatives to coverage probability, one can present posterior probability of coverage for "objective" noninformative priors (c.f. Box and Tiao (1973)). As alternatives to error probabilities in testing, one can argue simply for the use of likelihood ratios of hypotheses (at least when the hypotheses are simple). Indeed, fostering the attitude that one can often simply look at the likelihood function itself and obtain most things of interest is important.
The point, of course, is to avoid introduction

of subjective prior distributions until the frequentist has already accepted the conditional viewpoint. Indeed, it can be argued that acceptance of the conditional viewpoint is a far more drastic and important step than is the formal use of Bayesian methodology to implement the viewpoint. Thus L.J. Savage said (in the Discussion of Birnbaum (1962))

> "I, myself, came to take... Bayesian statistics...seriously only through recognition of the likelihood principle."

3.2 Robust Bayesian Analysis

Although there do exist statisticians who adopt a mainly conditional viewpoint, and yet stop short of being Bayesians, most statisticians find the step from conditioning to Bayesian analysis to be a small one. The basic problem for a non-Bayesian conditionalist is that of interpreting and using the likelihood function, and cogent arguments can be presented (c.f. Basu (1975) and Berger and Wolpert (1984)) that in, other than very simple applications, the likelihood function should be interpreted as a probability density with respect to a prior distribution π for the unknown θ. (The coherency arguments can be used as an aid to support the necessity of such an interpretation.) What prior π to use is a matter much debated, even among Bayesians. In Jeffreys (1961), Zellner (1971), and Box and Tiao (1973) it is argued that, for applications in which little prior information is available or an objective analysis is deemed to be desirable, π should be a noninformative prior, or its close relative in testing problems, a "reference informative prior" (c.f. Zellner (1982)). The pure subjectivist argues that a proper subjective prior should virtually always be used. The robust Bayesian (c.f. Berger (1980b, 1984)) argues that no single prior should be chosen, but that instead one learns by passing various plausible priors, π, over the likelihood function. While not disputing the value of any of these approaches (the noninformative prior approach is very good if "objectivity" is demanded or little

subjective information is available, and the single prior subjective approach rarely gives bad answers), we would argue for the robust Bayesian approach on pragmatic and scientific grounds. (We do not mean to imply that most Bayesians are "non-robust," but rather that we are not vocal enough about robustness.)

The pragmatic reason for taking a robust Bayesian approach is obvious. By admitting to non-Bayesians that we cannot determine exactly a completely trustworthy prior distribution, and yet arguing that only by trying various plausible priors and seeing what happens can true understanding be gained, we in one stroke overcome the biggest perceived objection to Bayesian analysis.

A concerted effort to justify the robust Bayesian method on scientific grounds is given in Berger (1984), in which earlier references to the method can also be found. (The method is sometimes called the Doogian approach because of the long advocacy of robust Bayesian techniques by I.J. Good.) Example 3 provides a simple illustration of the need to consider robust Bayesian analysis (or sensitivity analysis). The philosophical justification for the viewpoint is clear; subjective probabilities can be specified only to a certain degree of accuracy by finite minds and, even worse, prior distributions on infinite sets require the specification of an infinite number of subjective probabilities. Indeed, the resistance of many Bayesians to public acknowledgement of this seemingly obvious conclusion that we will always be working, in reality, with prior distributions that have been arbitrarily chosen (to at least some extent)is somewhat surprising. (Of course, as mentioned in Example 3, good Bayesian techniques do exist for helping to select a prior that is likely to work well.)

To obtain a theory suitable for general use in practice, Bayesians must go beyond even the pure robust Bayesian position, wherein before performing the experiment a class Γ of plausible prior distributions is envisaged, which is then employed through standard Bayesian updating by the likelihood function. Among the "sins" which it seems necessary to practice are the following:

(i) Delaying at least some of the prior specification until the likelihood function (for the observed data) is available (or alternatively

allowing data-based choice of the prior or class of priors). The reason for this is simply that, in realistically high dimensional problems, it will be very difficult to specify believable prior distributions; looking first at the likelihood function can indicate where prior elicitation efforts need to be concentrated (namely, where the data provides inconclusive evidence). The viewpoint, that one is merely passing plausible priors over the likelihoood function to see what happens, makes this heresy seem more palatable.

(ii) Some of the data may be ignored, if its incorporation would involve considerable extra (and uncertain) prior specification, and if it seems that the posterior distribution of the quantity of interest would not be excessively changed by incorporation of this data. Thus Hill (1975) 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..."

Further discussion of this point can be found in Pratt (1965), who calls it use of "insufficient statistics".

(iii) Use of non-Bayesian (even frequentist) measures may be helpful in certain situations, as will be argued in the next section.

All of the above points are discussed at considerable length in Berger (1984). When combined with an acknowledgment that, in Bayesian reporting, one must clearly report the likelihood function and the priors separately (along with the Bayesian conclusions), virtually any objection to the Bayesian approach can be dealt with.

3.3 Potential Uses of Frequency Measures

One of the major objections to Bayesian analysis is that there are many

problems that seem to be solvable in a non-Bayesian fashion (c.f. Example 1) and yet have no trustworthy Bayesian solution. Of course, to a large extent this is because far more research has been conducted on frequentist methods than on Bayesian methods; the existence of such problems is to be expected. Indeed, there is perhaps doubt that frequentist methods will prove to be of lasting value; thus L.J. Savage in Savage et. al. (1962) states

> "I used to be bowed by critics who said, with apparent technical justification, that certain popular nonparametric techniques apply in situations where it seems meaningless even to talk of a likelihood function, but I have learned to expect that each of these techniques either has a Bayesian validation or will be found to have only illusory value as a method of inference."

Nevertheless, at the moment, certain frequentist methods are certainly of value to a Bayesian. We briefly discuss these here.

A. Design, Prediction, and Sequential Analysis

It is indisputable that many questions in statistics involve "looking ahead" at the data that can be expected to occur. Frequentist averaging over the sample space is certainly necessary in such problems, 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, the

Bayesian will have the goal of obtaining good conditional performance, which may lead to a quite different design than a classical design.

B. Procedures for Nonspecialists

It is probably the case that most of the people who will actually be using statistics will not have been well enough trained to perform a careful conditional-robust Bayesian analysis. Hence a major portion of theoretical work should be devoted to providing relatively simple and easy to use Bayesian procedures with built in robustness. Since these procedures will be used repeatedly, averages over the sample space become relevant. Consider the following.

Example 3 (continued). Suppose the simple Bayesian procedure envisioned consists of eliciting the prior median and quartiles from the user, and then calculating the posterior mean for the fitted prior. If performance is measured by the squared error loss in estimation, it is easy to see that use of δ^N, the estimate from the conjugate normal prior π^N, can result in an infinite overall expected loss

$$r(\pi,\delta^N) = E(\theta - \delta^N(X))^2$$

(the expectation being taken over both θ and X), when the true prior is the Cauchy prior π^C. On the other hand, if δ^C, the Bayes estimate for the fitted Cauchy prior, is used, $r(\pi^N,\delta^C)$ is quite small. Indeed, for any reasonable prior π with median 0 and quartiles ± 1, $r(\pi,\delta^C)$ will be very satisfactory; thus, in repeated use, Cauchy priors should prove more satisfactory (for this situation) than normal priors. (Even for sensible bounded losses, the overall performance of δ^N can be bad.)

Another much more sophisticated situation, involving the Stein effect in estimating a multivariate normal mean, can be found in Berger (1982b). Here procedures are developed, whose Bayesian (conditional) performance is excellent, yet which are extremely insensitive to departures from prior assumptions.

C. Frequency Measures Which Imply Good Average Conditional Performance

We will illustrate this notion with the concept of coverage probability. (See Pratt (1965) for more general development and other applications.) Thus suppose $C(X)$ is a confidence procedure at level $1-\alpha$, so that

$$P_\theta(C(X) \text{ contains } \theta) \geq 1 - \alpha \quad \text{for all } \theta \in \Theta .$$

Then, for any prior distribution π on Θ,

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

However,

$$\begin{aligned} E^\pi P_\theta(C(X) \text{ contains } \theta) &= E^m E^{\pi(\theta|X)} P(\theta \in C(X)) \\ &\equiv E^m \lambda_\pi(X), \end{aligned}$$

where m is the marginal distribution of X and $\lambda_\pi(x)$ is clearly the posterior probability that θ is in $C(x)$ (which is the ideal conditional measure). It follows that

$$E^m \lambda_\pi(X) \geq 1 - \alpha,$$

which implies that $C(x)$ has a very good chance of containing θ (if α is

small) according to a good conditional measure (and no matter what π is). For the above reason, many Bayesians would accept the classical answer in Example 1 as the best approximate estimate of the chance that η is in $[x_{(3)}, x_{(12)}]$ that is currently obtainable. Care should still be taken, however, to make sure that there are no obvious conditional problems (as in Example 4). Thus Good (1976) states that

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

D. Frequency Measures as an Aid to Robust Bayesian Analysis

When the prior elicitation process has ended, and one is left with a class Γ of possible priors, it is to be hoped that the Bayesian conditional answer is clearcut (i.e. passing the different priors in Γ over the likelihood function results in essentially the same conclusion). When this is not the case, and an answer must be produced, there is no clearcut way to proceed. A natural Bayesian possibility is to put some (arbitrary) metaprior on Γ (i.e. a prior distribution on the elements of Γ). This is merely a formal way to proceed, however, since it is being assumed that no further subjective prior elicitation is possible. As a possible alternative, one could select from among the possibilities by frequentist criteria. Discussion of this can be found in Berger (1980b) and Berger (1984). It should be emphasized that there is no foundational reason to prefer any particular method here, since further prior elicitation is not possible.

The above view is not really a completely practical view, since it will rarely be the case that a constructed Γ is impossible to further refine, often by the data interactive process of looking at the predictive likelihoods (i.e. $m(x)$) for the various priors in Γ and observed x. It may still be the

case, however, that the best pragmatic way to proceed is to choose among the possible priors by frequency methods. If, for instance, a prior in Γ results in a procedure with good frequency properties, then for reasons indicated in part C it may be reasonable to just use that prior instead of attempting further refinement of Γ. (This is especially true in those situations where Γ is very large and Bayesian calculations are very difficult.) This is not to be interpreted as an advocacy of the routine use of frequency concepts in such situations in preference to the more natural Bayesian methods, but instead is another justification for allowing consideration of frequency measures when a Bayesian analysis is unclear or difficult.

E. Other Uses of Frequency Concepts in Bayesian Analysis

Classical significance testing of a null hypothesis is of some value, as a quick and dirty indicator of a situation needing deeper investigation (i.e. of a situation in which formulation and examination of alternatives, hopefully by Bayesian measures, is in order). As with all techniques violating the Likelihood Principle, however, care must be taken in attempting to interpret P-values literally. (See Berger and Wolpert (1984) for discussion and other references.)

Invariance theory in statistics is closely related to the Bayesian approach with noninformative priors (c.f. Berger (1980b)), and indeed often suggests appropriate choices for a noninformative prior.

Asymptotics is, of course, frequently relevant, providing often needed simplifications. Indeed, there is a very substantial literature on Bayesian asymptotics. (See Berger (1984) for references.)

The study of admissibility, with its close tie-ins with Bayesian theory, has led to a number of significant Bayesian advances, particularly in Bayesian robustness (via study of the Stein effect).

This list is by no means exhaustive, or even representative. Much of frequentist statistics has helpful things to say to Bayesians.

4. Conclusions

The essential features of the approach presented for selling the Bayesian view are (i) to break the argument up into easily digestible (and defendable) parts, and (ii) to allow a great deal of flexibility in methods. Arguing first for the adoption of a conditional viewpoint, and then for the introduction of priors to implement this viewpoint, reduces the need to deal with the initial anti-prior bias of many non-Bayesians. It also focuses the issue on the fact that the goal is to obtain good conditional probabilistic measures of uncertainty.

One can then address the concern of how best to achieve this goal. Here, the Bayesian salesman does best to avoid being too dogmatic. Adopting a robust Bayesian position, i.e. admitting uncertainty in the prior, is only sensible and realistic. (Of course, frequentist models, etc., are every bit as uncertain, but we are not trying to sell the frequentist viewpoint). Also, admitting that there are presently many problems where believable Bayesian answers are not available and reasonable frequentist answers are, is only prudent. It must be stressed, however, that a frequentist answer is not inherently sensible; it must have some plausible relationship to a meaningful conditional measure. It would be very nice to have the arguments and debates come down to this level; namely, that of which methodology best achieves the conditional goal.

There are substantial rewards, to the statistician adopting a conditional viewpoint, that should also be pointed out. First of all, many difficult problems involving stopping rules and censoring disappear. (See Berger and Wolpert (1984) for discussion and references.) Also, time need not be wasted on extremely difficult frequentist problems, such as solution of the Behrens-Fisher problem; conditional Bayesian solutions (including objective noninformative prior solutions) are readily available. Finally, it can lead the statistician to avoid frequentist analysis or research which is clearly ridiculous conditionally, even if the statistician is unwilling to become a Bayesian.

It is possible to object to the flexible robust Bayesian position, advocated above, on the grounds that it will frequently be incoherent, and indeed it often will be formally incoherent. The word "formally" is stressed, since the approach does strive for coherency to the extent obtainable. Our own view is that, as Bayesian methodology expands, there will be less and less need to leave the formal Bayesian paradigm; we will learn what types of prior distributions are inherently robust and work well for given problems. There is no need to make such a strict position a part of the Bayesian argument, however.

Acknowledgment. I am grateful to Prem Goel and Arnold Zellner for suggestions which substantially improved the manuscript.

References

Barnard, G.A. (1980). Pivotal inference and the Bayesian controversy (with Discussion). In Bayesian Statistics, University Press, Valencia.

Basu, D. (1975). Statistical information and likelihood (with discussion). Sankhyā, Ser. A 37, 1-71.

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. (1984). The robust Bayesian viewpoint. In: Robustness in Bayesian Statistics, J. Kadane (ed.). North Holland, Amsterdam.

Berger, J. and Wolpert, R. (1984). The Likelihood Principle: a Review and Generalizations. Monograph Series of the Institute of Mathematical Statistics.

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

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

Cox, D.R. (1958). Some problems connected with statistical inference. Ann. Math. Statist. 29, 357-372.

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

Efron, B. (1982). Why isn't everyone a Bayesian? Presented at the conference 'Reflections on Bayesian Approaches in Operations Research, Probability, and Statistics', Blacksburg, Virginia (1982).

Fishburn, P.C., 1981. Subjective expected utility: a review of normative theories. Theory and Decision 13, 139-199.

Fraser, D.A.S. (1968). The Structure of Inference. Wiley, New York.

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

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.). Reidel, Boston.

Good, I.J. (1983). The robustness of a hierarchical model for multinomials and contingency tables. In: Scientific Inference, Data Analysis, and Robustness, G.E.P. Box, T. Leonard, and C. F. Wu (eds.). Academic Press, New York.

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

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

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

LeCam (1977). A note on metastatistics or 'an essay toward stating a problem in the doctrine of chances'. Synthese 36, 133-160.

Lindley, D.V. (1971). Bayesian Statistics Review. S.I.A.M., Philadelphia.

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

Ramsey, F.P. (1926, 1928). 'Truth and probability' (1926) and 'Further considerations' (1928), in The Foundations of Mathematics and Other Logical Essays. Harcourt, Brace, and Co., New York (1931).

Rubin, H. (1974). Axiomatic development of rational behavior. Technical Report, Department of Statistics, Purdue University.

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

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

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

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

Zellner, A. (1981). The current state of Bayesian econometrics. Address at the Canadian Conference on Applied Statistics. Concordia University, Montreal, April 29 - May 1, 1981.

Zellner, A. (1982). Applications of Bayesian analysis in econometrics. Address at the Institute of Statisticians International Conference on Practical Bayesian Statistics, St. John's College, Cambridge, July 21-24, 1982.