Why Isn't Everyone a Bayesian?

B. Efron

The American Statistician, Vol. 40, No. 1 (Feb., 1986), 1-5.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198602%2940%3A1%3C1%3AWIEAB%3E2.0.CO%3B2-3

The American Statistician is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact jstor-info@umich.edu.

http://www.jstor.org/
Tue Jan 13 11:32:37 2004

Why Isn't Everyone a Bayesian?

B. EFRON*

Originally a talk delivered at a conference on Bayesian statistics, this article attempts to answer the following question: why is most scientific data analysis carried out in a non-Bayesian framework? The argument consists mainly of some practical examples of data analysis, in which the Bayesian approach is difficult but Fisherian/frequentist solutions are relatively easy. There is a brief discussion of objectivity in statistical analyses and of the difficulties of achieving objectivity within a Bayesian framework. The article ends with a list of practical advantages of Fisherian/frequentist methods, which so far seem to have outweighed the philosophical superiority of Bayesianism.

KEY WORDS: Fisherian inference; Frequentist theory; Neyman–Pearson–Wald; Objectivity.

1. INTRODUCTION

The title is a reasonable question to ask on at least two counts. First of all, everyone used to be a Bayesian. Laplace wholeheartedly endorsed Bayes's formulation of the inference problem, and most 19th-century scientists followed suit. This included Gauss, whose statistical work is usually presented in frequentist terms.

A second and more important point is the cogency of the Bayesian argument. Modern statisticians, following the lead of Savage and de Finetti, have advanced powerful theoretical reasons for preferring Bayesian inference. A byproduct of this work is a disturbing catalogue of inconsistencies in the frequentist point of view.

Nevertheless, everyone is not a Bayesian. The current era is the first century in which statistics has been widely used for scientific reporting, and in fact, 20th-century statistics is mainly non-Bayesian. [Lindley (1975) predicts a change for the 21st!] What has happened?

2. TWO POWERFUL COMPETITORS

The first and most obvious fact is the arrival on the scene of two powerful competitors: Fisherian theory and what Jack Kiefer called the Neyman–Pearson–Wald (NPW) school of decision theory, whose constituents are also known as the frequentists. Fisher's theory was invented, and to a remarkable degree completed, by Fisher in the period between 1920 and 1935. NPW began with the famous lemma of 1933, asymptoting in the 1950s, though there have continued to be significant advances such as Stein estimation, empirical Bayes, and robustness theory.

Working together in rather uneasy alliance, Fisher and NPW dominate current theory and practice, with Fisherian ideas particularly prevalent in applied statistics. I am going to try to explain why.

*B. Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305.

3. FISHERIAN STATISTICS

In its inferential aspects Fisherian statistics lies closer to Bayes than to NPW in one crucial way: the assumption that there is a *correct* inference in any given situation. For example, if $x_1, x_2, \ldots, x_{20}$ is a random sample from a Cauchy distribution with unknown center θ,

$$f_\theta(x_i) = \frac{1}{\pi[1 + (x_i - \theta)^2]},$$

then in the absence of prior knowledge about θ the correct 95% central confidence interval for θ is, to a good approximation,

$$\hat{\theta} \pm 1.96 \Big/ \sqrt{-\ddot{l}_{\hat{\theta}}},$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE) and $\ddot{l}_{\hat{\theta}}$ is the second derivative of the log-likelihood function evaluated at $\theta = \hat{\theta}$. The (mathematically) equally good approximation

$$\hat{\theta} \pm 1.96/\sqrt{10}$$

(10 being the expected Fisher information), is not correct (Efron and Hinkley 1978).

Fisher's theory is a theory of archetypes. For any given problem the correct inference is divined by reduction to an archetypal form for which the correct inference is obvious. The first and simplest archetype is that of making inferences about θ from one observation x in the normal model

$$x \sim N(\theta, 1). \tag{1}$$

Fisher was incredibly clever at producing such reductions: sufficiency, ancillarity, permutation distributions, and asymptotic optimality theory are among his inventions, all intended to reduce complicated problems to something like (1). (It is worth noting that Fisher's work superseded an earlier archetypical inference system, Karl Pearson's method of moments and families of frequency curves.)

Why is so much of applied statistics carried out in a Fisherian mode? One big reason is the *automatic nature* of Fisher's theory. Maximum likelihood estimation is the original jackknife, in Tukey's sense of a widely applicable and dependable tool. Faced with a new situation, the working statistician can apply maximum likelihood in an automatic fashion, with little chance (in experienced hands) of going far wrong and considerable chance of providing a nearly optimal inference. In short, he does not have to think a lot about the specific situation in order to get on toward its solution.

Bayesian theory requires a great deal of thought about the given situation to apply sensibly. This is seen clearly in the efforts of Novick (1973), Kadane, Dickey, Winkler, Smith, and Peters (1980), and many others to at least partially automate Bayesian inference. All of this thinking is admirable in principle, but not necessarily in day-to-day practice. The same objection applies to some aspects of

© 1986 American Statistical Association

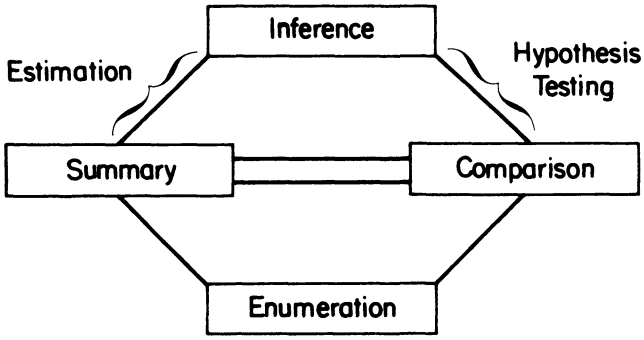

Figure 1. Four Basic Statistical Operations and How They Relate to Estimation. Source: Efron (1982b, fig. 2).

NPW theory, for instance, minimax estimation procedures, and with the same result: they are not used very much.

Not all of statistics is inference. The little diagram of all of statistics in Figure 1 (reprinted from Efron 1982b) starts at the bottom with "enumeration," the collecting and listing of individual datum. The diagram proceeds upward to the reduction of the raw data to more understandable form through the adversarial processes of summary and comparison. This is the part of the analysis where, usually, the statistician decides on a reasonable probabilistic model for the situation. At the top of the diagram is inference. This is the step that takes us from the data actually seen to data that might be expected in the future.

Bayesian theory concentrates on inference, which is the most glamorous part of the statistical world, but not necessarily the most important part. Fisher paid a lot of attention to the earlier steps of the data analysis. Randomization for instance, and experimental design in general, is a statement about how data should be collected, or "enumerated," for best use later in the analysis. Maximum likelihood is a provably efficient way to summarize data, no matter what particular estimation problems are going to be involved in the final inference (Efron 1982b). The NPW school has also contributed to the theory of enumeration, notably in the areas of survey sampling and efficient experimental design.

Fisher's theory culminated in fiducial inference, which to me and most current observers looks like a form of objective (as opposed to subjective) Bayesianism. I will discuss the problems and promise of objective Bayesianism later, but it is interesting to notice that fiducial inference is alone among Fisher's major efforts in its failure to enter common statistical application. In its place, the NPW theory of confidence intervals dominates practice, despite some serious logical problems in its foundations.

4. THE NPW SCHOOL

Unlike Bayes and Fisher, the NPW school does not insist that there is a correct solution for a given inferential situation. Instead, a part of the situation deemed most relevant to the investigator is split off, stated in narrow mathematical fashion, and it is hoped, solved. For example, the correct Bayesian or Fisherian inference for θ in situation (1) leads directly to the correct inference for $\gamma \equiv 1/(1 + \theta)$, but this is not necessarily the case in the NPW formulation. (What is the uniform minimum variance unbiased estimate of γ?)

The NPW piecewise approach to statistical inference has been justly criticized by Bayesians as self-contradictory, inconsistent, and incoherent. The work of Savage, de Finetti, and their successors shows that no logically consistent inference maker can behave in such a non-Bayesian way. The reply of the NPW school is that there is no reason to believe that statistical inference should be logically consistent in the sense of the Bayesians, and that there are good practical reasons for approaching specific inference problems on an individual basis.

As an example consider the following problem: we observe a random sample $x_1, x_2, \ldots, x_{15}$ from a continuous distribution F on the real line and desire an interval estimate for θ, the median of F. The experiment producing the x_i is a new one, so very little is known about F.

A genuine Bayesian solution seems difficult here, since it requires a prior distribution on the space of all distributions on the real line. Frequentist theory produces a simple solution in terms of a confidence interval based on the order statistics of the sample,

$$\theta \in [x_{(3)}, x_{(12)}]$$

with probability .963, no matter what F may be. The fact that this solution, unlike a Bayesian one, does not also solve the corresponding problem for say $\phi \equiv 50\%$ trimmed mean of F does not dismay the frequentist, particularly if a satisfactory Bayesian solution is not available.

The Bayesian accusation of incoherency of the frequentist cuts both ways: in order to be coherent Bayesians have to solve all problems at once, an often impossible mental exercise.

As another example consider "rejecting at the .05 level." The inconsistencies of this practice are well documented in the Bayesian literature (see Lindley 1982). On the other hand it is one of the most widely used statistical ideas. Its popularity is founded on a crucial practical observation: it is often easier to compare quantities than to assign them absolute values. In this case the comparison is between the amount of evidence against the null hypothesis provided by different possible outcomes of the data. For testing H_0: $x \sim N(0, 1)$ versus H_1: $x \sim N(2, 1)$, we know that a larger observed x provides greater evidential value against H_0 and in favor of H_1, even if we cannot absolutely quantify "evidence."

A Bayes solution to this problem, "the aposteriori odds ratio is 7 to 1 in favor of H_1," is more satisfactory than "the data are significant at the .05 level," but it also requires more input. In fact, it tacitly implies that we have assigned an *absolute* measure of evidence to every possible outcome. Absolute here means that the meaning of 7 to 1 is the same no matter what experiment it came from. [Good's (1965) Bayes–non-Bayes compromise suggests using Bayesian ideas in a comparative mode, but this is the only example I know.]

The heart attack decision tree (Fig. 2) illustrates another difficult situation for the honest Bayesian. The tree purports to predict coronary patients with high risk of dying (population 2) on the basis of variables observed at hospital admission. A series of dichotomous observations are made, for example, high or low kinase level, which result in a final prediction. The nodes marked "2" on the tree predict death. Of the 389 patients classified by the tree, only 1 out

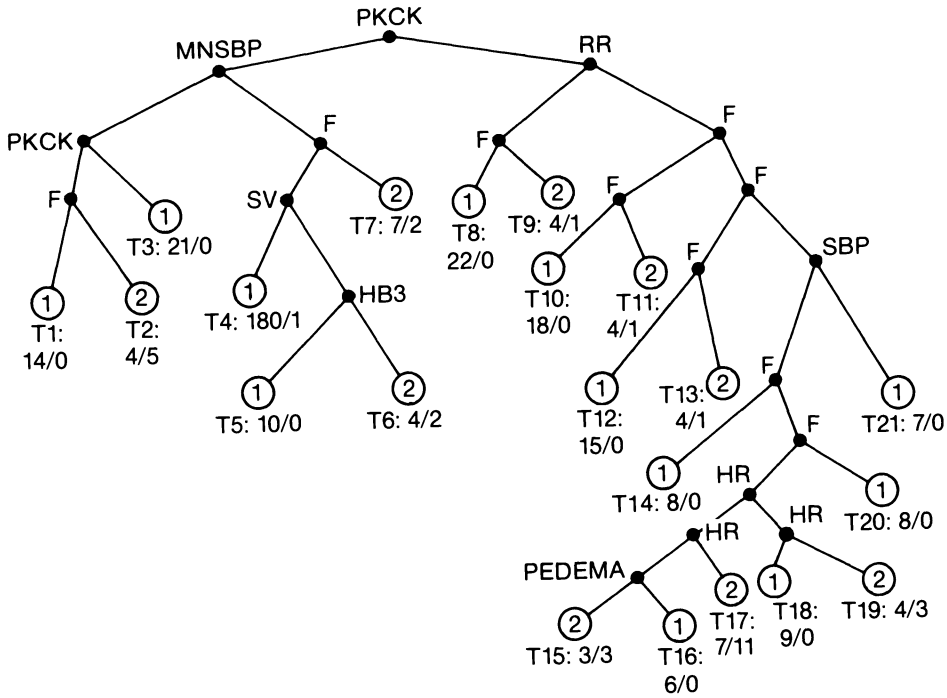

Figure 2. A Decision Tree for Classifying Heart Attack Patients Into Low Risk of Dying (population 1) or High Risk of Dying (population 2). Smaller values of the decision variables go to the left. Circled numbers at terminal nodes indicate population prediction. For example, 6 of the 389 patients in the training set end up at T6, 4 from population 1, 2 from population 2; these patients would all be predicted to be in population 2. Abbreviations: PKCK, peak creatinine kinase level; MNSBP, minimum systolic blood pressure; SBP, systolic blood pressure; FF, respiration rate; HR, average heart rate; SV, superventricular arrhythmia; HB3, heart block 3rd degree; PEDEMA, peripheral edema; F, Fisher linear discriminant function, differing from node to node. Source: Efron (1982a, fig. 7.1).

of 30 deaths was misclassified, that is, predicted to live. Can we believe that the tree has 96.7% probability of successfully predicting deaths?

Because the medical investigators had little prior knowledge of the situation, the tree was constructed by an elaborate data-fitting procedure, which in fact was designed to maximize the apparent success rate. At each stage the dichotomous variable to be used and the splitting point defining "high" or "low" were chosen to give the maximum apparent difference between populations 1 and 2. A bootstrap analysis, much like a cross-validation, gave an unbiased estimate of successful prediction of death of about 70%, rather than 96.7%, for this tree. (Details appear in sec. 7.6 of Efron 1982a.)

The fact that the observed data were used to construct the tree, and how they were used, makes no difference to the Bayesian, since it has no effect on the likelihood function. This is similar in spirit to the fact that the stopping rule used in a sequential procedure has no Bayesian consequence. It makes a world of difference to the frequentist. If exactly the same tree had been constructed by a less flexible rule, the unbiased estimate would move closer to the observed value 96.7%. This is incoherent behavior. The Bayesian estimate, whatever it is, would not change.

"Ad hoc" is a pejorative adjective in Bayesian descriptions of frequentist statistics. On the other hand, ad hoc reasoning produces a reasonable answer here, in a problem that seems far too complicated for a full Bayesian solution. The right to split off the simple part of a complicated inference problem should not be the exclusive property of the frequentists, but I am not aware of much Bayesian activity along these lines. The coherency approach of Savage and de Finetti seems to have discouraged it. (For a counterexample to this statement see Boos and Monahan, in press.)

The NPW school invented decision theory, but it is not decision theory that separates them from the Bayesians. In fact, Bayesians have made good use of decision theory. The parting of the ways occurs on the crucial issue of θ averages, expectations taken with the state of nature θ fixed. In other words, frequentist calculations. Controlling, or at least computing, θ averages is central to the NPW approach and irrelevant to the Bayesians. This brings us to the topic of objectivity, in my opinion the linchpin of non-Bayesian success with statistical practitioners.

5. OBJECTIVITY

So far I have been careful not to define the kind of Bayesian theory under criticism. The dominant Bayesian school, and the one with the legitimate claim to philosophic coherency, is the subjective Bayesianism of de Finetti and Savage. Now by definition one cannot argue with a subjectivist, so I will just state the obvious fact: though subjectivism is undoubtedly useful in situations involving personal decision making, for example, business and legal decisions, it has failed to make much of a dent in scientific statistical practice. The nature of scientific communication makes me doubt that it ever will.

"Scientific objectivity" is more than a catch-phrase. Strict objectivity is one of the crucial factors separating scientific thinking from wishful thinking. Complete objectivity about one's own work is a little much to expect from a human being, even a scientist, but it is not too much to expect from

one's colleagues. A prime requirement of any statistical theory intended for scientific use is that it reassures oneself *and others* that the data have been interpreted fairly.

With this in mind, it is not surprising that intuitively "fair" statistical ideas, like unbiasedness, confidence intervals, and .05 significance, are immensely popular with the statistical public. Ideas that do not pass the test of objectivity are not much used. This includes NPW ideas as well as Bayesian ones, for example, James–Stein estimation. (An interesting borderline case, which could go either way depending on how it develops, is robust estimation.)

Of course there is no scientific law that says that objectivity must be interpreted in a frequentist sense, and in fact, there is another line of Bayesian thought that attempts to deal directly with the issue of fairness. I call this "objective Bayes theory" to differentiate it from the Savage–de Finetti approach. Bayes and Laplace were objective Bayesians, and in this century, Jeffreys (1961) wrote a famous book on the subject. The goal of objective Bayesianism is to produce prior distributions that capture the idea of objectivity.

Consider situation (1) again. The obvious prior here is the improper one, spreading probability mass for θ uniformly from $-\infty$ to $+\infty$, often denoted simply by "$d\theta$." A Bayesian using this prior obtains good frequentist results. The central 90% aposteriori interval for θ, for example, agrees exactly with the standard 90% confidence interval.

Next consider the situation where we observe

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, I\right), \tag{2}$$

which is just two independent copies of (1). In some recent work it was necessary for me to make inferences about the parameter $\lambda = \theta_1\theta_2$. It seemed intuitively reasonable, and objective, to use the improper prior $d\theta_1\, d\theta_2$, which spreads probability mass for (θ_1, θ_2) uniformly over the entire plane.

As Table 1 shows, the aposteriori central 90% probability interval for λ derived from the prior $d\theta_1\, d\theta_2$ does not have good frequentist properties. For values of (θ_1, θ_2) in the first quadrant, it gives overly low probabilities of missing λ on the left and overly high probabilities of missing λ on the right. From a frequentist viewpoint we have not been very objective at all, having biased the interval estimates toward the origin.

Recently Charles Stein has given a method of constructing priors that have better frequentist properties (Stein 1982).

Table 1. Theoretical Versus Actual Probability of Not Covering $\lambda = \theta_1\theta_2$ Using the Central 90% Aposteriori Interval Based on the Improper Prior $d\theta_1\, d\theta_2$

	Theoretical	
(θ_1, θ_2)	.050	.050
(0, 0)	002 (006)	002 (006)
(2, 2)	023 (036)	087 (065)
(3, 3)	029 (041)	074 (053)
(4, 4)	031 (046)	070 (047)
(5, 5)	036 (048)	064 (043)
(1, 10)	045 (053)	051 (053)

NOTE: Figures in parentheses are, essentially, the same probabilities using the improper prior $(\theta_1^2 + \theta_2^2)^{1/2}\, d\theta_1\, d\theta_2$.

For the parameter λ, Stein's theory suggests using the improper prior $(\theta_1^2 + \theta_2^2)^{1/2}\, d\theta_1\, d\theta_2$, and Table 1 shows that this does indeed give better frequentist coverage probabilities. [In fact, these figures were derived for the bias-corrected percentile intervals, sec. 10.7 of Efron (1982a), and it was then verified that these intervals were almost the same as Stein's.]

The point of the example is that the theory of Bayesian objectivity cannot be a simple one. The correct objective prior seems to depend on which parameter we want to estimate. In higher dimensions when we have several parameters rather than just two, these problems become acute (Efron 1982b). This does not mean that the situation is hopeless. Even a partial solution to the problem of Bayesian objectivity would likely be a valuable contribution to statistical theory and practice. As a hopeful prototype, the Bayesian explanation of the James–Stein estimator has deepened our understanding of this potentially wonderful tool. The whole subject of empirical Bayes can be thought of as an exercise in Bayesian objectivity—trying not to put more information than necessary into the prior—and more progress in this area can be expected.

6. SUMMARY

A summary of the major reasons why Fisherian and NPW ideas have shouldered Bayesian theory aside in statistical practice is as follows:

1. Ease of use: Fisher's theory in particular is well set up to yield answers on an easy and almost automatic basis.
2. Model building: Both Fisherian and NPW theory pay more attention to the preinferential aspects of statistics.
3. Division of labor: The NPW school in particular allows interesting parts of a complicated problem to be broken off and solved separately. These partial solutions often make use of aspects of the situation, for example, the sampling plan, which do not seem to help the Bayesian.
4. Objectivity: The high ground of scientific objectivity has been seized by the frequentists.

None of these points is insurmountable, and in fact, there have been some Bayesian efforts on all four. In my opinion a lot more such effort will be needed to fulfill Lindley's prediction of a Bayesian 21st century.

[*Received July 1985.*]

REFERENCES

Boos, D., and Monahan, J. (in press), "The Bootstrap for Robust Bayesian Analysis: An Adventure in Computing," *14th Symposium on the Interface*.

Efron, B. (1982a), *The Jackknife, the Bootstrap, and Other Resampling Plans*, CBMS-NSF Monograph 39, Philadelphia: SIAM.

——— (1982b), "Maximum Likelihood and Decision Theory," *Annals of Statistics*, 10, 340–356.

Efron, B., and Hinkley, D. V. (1978), "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information," *Biometrika*, 65, 457–487.

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

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford: Clarendon Press.

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," *Journal of the American Statistical Association*, 75, 845–854.
Lindley, D. V. (1975), "The Future of Statistics—A Bayesian 21st Century," *Supp. Adv. Appl. Prob.*, 7, 106–115.
——— (1982), "Scoring Rules and the Inevitability of Probability," *ISI Review*, 50, 1–26.
Novick, M. (1973), "High School Attainment: An Example of a Computer-Assisted Bayesian Approach to Data Analysis," *ISI Review*, 41, 268–271.
Stein, C. (1982), "The Coverage Probability of Confidence Sets Based on a Prior Distribution," Technical Report 180, Stanford University.

Comment

HERMAN CHERNOFF*

This discussion of Efron's elegant paper reminds me of the story of the reaction of a jealous mathematician to a new theorem. "This theorem is irrelevant. Moreover, it is wrong. Besides, I derived it long ago." I wish (a) to add some reasons why not everyone is a Bayesian, (b) to explain why everyone is, should be, or will soon be a Bayesian, and (c) to claim that (objective) Bayesianism is wrong.

1. WHY NOT EVERYONE IS A BAYESIAN

Teaching

Statistical practice changes slowly because teaching of elementary statistics changes slowly. Most statistical practice is at a rather elementary level, although frequently applied to complex problems. It is dedicated to communicate with people with meager statistical training and constrained by conventions in elementary texts. Therefore, Bayes's theorem is relatively abstract and less intuitive than applications of proportions, means, and standard deviations.

Once sociologists and physicians have learned about significance levels well enough to use them, a major reorganization of the thought processes is required to adapt to decision theoretic or Bayesian analysis. Guttman (1978) has railed for years about the stargazing habits (associated with significance levels .01 and .05) of his fellow sociologists, which he attributes to poor teaching on the part of the statistical profession.

Practice Versus Theory

The popularity of Fisherian theory stems in part from the byproducts of the analysis suggested by the theory. Analysis of variance is one of the most important data analytic tools in statistics. This tool stems from a hypothesis-testing formulation that is difficult to take seriously and would be of limited value for making final conclusions. Its importance stems from the fact that most scientific and statistical practice is concerned, not with grand final conclusions, but with many small steps gradually contributing clarity and order in moderately confused situations. Theory serves as a guide and enhances intuition, and it should not be used as a prescription requiring optimal behavior. Fisherian theory provides relatively simple, effective tools, the robustness of which seems apparent in circumstances in which the Bayesian and decision theorist may find it difficult to operate.

By the time clarity is attained, there is little additional need for sophisticated statistics or for statisticians. By then the issues are clear, the appropriate experiments are evident, the noise factors are reduced to minor significance, and the conclusions are obvious to any intelligent *observer*. At this point, it is usually forgotten that a statistician ever played a useful role.

2. WHY EVERYONE IS, SHOULD BE, OR WILL BE A BAYESIAN

How I Am a Bayesian

In doing applied statistics, I feel relatively uninhibited and engage in many practices that might be frowned on by careful dogmatists. With the help of theory, I have developed insights and intuitions that, I believe, prevent me from giving undue weight to generalizations drawn from excessive data dredging or other forms of statistical heresy. This feeling of ease and freedom, however, does not exist until I have formulated some decision theoretic and Bayesian view of the problem. Until then, there is discomfort and the feeling that the problem is not well stated or understood. In this sense, I am a Bayesian and decision theorist in spite of my use of Fisherian tools. Is this a unique position or are most of us closet Bayesians and decision theorists in this sense?

When Bayesianism Is the Only Way

Whereas most applications of statistics are to small science and technology, there are cases in which great issues are involved. Makers of public policy are sometimes forced to make decisions on issues for which the scientific foundations are unclear and fundamental data are lacking (e.g., acid rain, ozone depletion, the safety of nuclear power, and carcinogenic effects of diesel fumes). Decisions made now in considerable ignorance have important consequences. Delaying such decisions is also potentially costly. Should we rely on the vagaries of coin tossing, or uninformed public discussions, or the possibly slightly better scientific analysis? But scientific analysis in nuclear safety may, for lack of any real alternative, require a Bayesian fault tree analysis with priors based more on imagination than on real data. One hopes that numbers so derived will not have undue influence on a public impressed by quantification.

*Herman Chernoff is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138.

© 1986 American Statistical Association