IN DEFENSE OF THE LIKELIHOOD PRINCIPLE: AXIOMATICS AND COHERENCY

by

James Berger
Purdue University

Technical Report #83-18

Department of Statistics
Purdue University

1983

IN DEFENSE OF THE LIKELIHOOD PRINCIPLE:

AXIOMATICS AND COHERENCY

BY JAMES O. BERGER

Department of Statistics
Purdue University
West Lafayette, Indiana 47907, USA

SUMMARY

Belief in the Likelihood Principle was substantially advanced when A. Birnbaum showed it to be derivable from the apparently more natural Sufficiency and Conditionality Principles. This axiomatic development subsequently came under attack from a number of directions, among the most interesting being a criticism (by D.A.S. Fraser, G. Barnard and others) of the Sufficiency Principle for failure to take into account "structural" knowledge of the performed experiment. This criticism is addressed in this paper from two directions. First, a weak set of alternative axioms for the Likelihood Principle is developed. Second, ideas of coherency are employed to question the validity of knowingly violating the Likelihood Principle. In this development, arguments are presented for basing coherency on decision-theoretic concepts, rather than the more usual betting concepts. The basic conclusions of the paper also apply to other theories which can violate the Likelihood Principle, including many noninformative prior Bayesian theories.

Keywords: LIKELIHOOD PRINCIPLE, CONDITIONALITY, SUFFICIENCY, STRUCTURAL ANALYSIS, PIVOTAL ANALYSIS, COHERENCY, INADMISSIBILITY, COMPLETE CLASSES.

1. INTRODUCTION

The Likelihood Principle (LP) received its major (non-Bayesian) impetus from ideas of R.A. Fisher and G. Barnard. (See Berger and Wolpert (1984) for references.) It essentially states that any decision or inference in a statistical problem should involve the data and experiment only through the likelihood function of the unknowns given the observed data. The implications of the LP are farreaching, and the case for it is strong. The non-Bayesian case rests on various axiomatic developments of the LP from simpler believable principles, such as the Birnbaum (1962) development of the LP from the Conditionality and Sufficiency principles.

Arguments against the LP usually take one of four forms. First, are arguments concerning the "unintuitive" consequences of the LP, such as the consequent irrelevance of stopping rules in the final inference or decision and the incompatibility of the LP with significance testing and randomization analysis. Discussion and references concerning such matters can be found in Basu (1975) and Berger and Wolpert (1984), but in general it is hard to see how anyone believing in logic could reject the LP because of its consequences without rejecting at least one of the precepts upon which it is based. The three remaining arguments against the LP thus focus on either (i) the existence of the likelihood function, (ii) the validity of the Conditionality principle, or (iii) the validity of the Sufficiency principle.

As presented in Birnbaum (1962), the LP is dependent on the existence of a likelihood function (and in fact on the discreteness of the sample space). Thus, it can be argued that in nondiscrete nonparametric problems (and one is rarely completely sure of a parametric model) there may be no clearly defined likelihood function, and even when there is a likelihood function the nondiscreteness makes Birnbaum's arguments questionable. Basu (1975) answers this by arguing that, in reality, data is always discrete (because of limitations on observational accuracy) so that no difficulties arise. The importance of using continuous approximations to discreteness is well recognized, however, and validity of the LP for nondiscrete situations would, therefore, be comforting. A general version of the LP (called the Relative Likelihood Principle) was developed for essentially arbitrary situations in Berger and Wolpert (1984). It was shown to follow from conditionality and sufficiency, and also shown to have essentially the same consequences as the LP. Rejection of the LP on the grounds of nonexistence of likelihood functions can be circumvented, therefore.

The Conditionality Principle (CP) roughly states that, if an independent coin is flipped to decide between performing experiments E_1 and E_2, both pertaining to some unknown quantity θ of interest, then the evidence about θ obtained from the data should not depend in any way on the experiment not actually performed. First stated explicitly by Cox (1958), the CP seems completely obvious, but, rather startlingly, it is in sharp opposition to standard frequentist reasoning in statistics. The experiment not actually performed is another part of the sample space, and hence a frequentist would average over it in determining the performance of his "procedure". Numerous attempts to partially follow the CP and then go frequentist have been advocated, but do not seem to be ultimately justifiable (c.f. Berger and Wolpert (1984)). It is not, on the other hand, illogical for a frequentist to simply reject the CP, arguing (c.f. Neyman and Pearson (1933)) that the notion that one can obtain "evidence about θ from an experiment" is misguided; all one can do is evaluate how well a procedure that will be repeatedly used performs in the long run, and the procedure's performance will be an average over both E_1 and E_2. While logically viable, the artificiality of the position is clear. I doubt if many users of statistics would be willing to accept the point of view that one cannot obtain evidence about θ from an experiment. (Note that the CP and LP presuppose nothing about what this "evidence" is, or even that it is any single quantity.)

The final criticism of the LP comes from ideas of Barnard (c.f. Barnard (1980, 1982), Barnard and Godambe (1982), and the discussion in Basu (1975)) and Fraser (c.f. Fraser (1963, 1968, 1972, 1979)) concerning the validity of the Sufficiency Principle. The criticism concerns the "sufficiency" of representing the experimental structure solely in terms of probability distributions on the sample space indexed by the unknown θ. This turns out to be a very difficult criticism to answer, and indeed it does not seem answerable in the same self-contained sense as the other criticisms. Although an attempt is made to deal with the issue axiomatically in Section 3, the attempt is something of a failure, one crucial axiom being suspect. Instead, therefore, we turn to Bayesian arguments of coherency and inadmissibility, in an attempt to indicate the difficulties in violating sufficiency. These arguments are, of course, more or less familiar, but because of the importance of the issue and the bearing that these arguments have on theories of inference such as Barnard's Pivotal Inference, Fraser's Structural Inference, and even various noninformative prior Bayesian theories, certain aspects of which may violate the LP, we will include a fair amount of detail and discussion. It should be stated at the outset that, in no sense, do we unequivocally resolve the controversy. The paper should be viewed more as an attempt to carefully state a Bayesian view of the issue.

Some Bayesians may wonder why there is any need to be concerned with the LP, except as a trivial consequence of the Bayesian position. There are essentially two reasons. The first is the purely pragmatic reason that promoting Bayesianism can often most effectively be done by first selling the LP, since this can be done without introducing the emotionally charged issue of prior distributions. The second reason is that the LP shows that Bayesians should be concerned with conditional (posterior) conclusions. This may seem to be a strange statement to most Bayesians, but it is certainly possible to be a Bayesian and not believe this. For instance, one could believe that only frequentist measures of procedure performance have validity, and yet, because of various rationality, coherency, or admissibility arguments, believe that the only reasonable procedures are Bayes procedures, and that the best method of choosing a procedure is through consideration of prior information and application of the Bayesian paradigm (c.f. the discussion by L. Brown in Berger (1983)). The posterior distribution will provide a convenient mathematical device for determining the best procedure, but (from this viewpoint) the overall frequentist (Bayes) performance of the procedure would be the relevant measure of accuracy. The LP directly attacks this view, arguing that thinking "conditional Bayes", not "frequentist Bayes," is important.

In Section 2 the needed notation will be given. Section 3 presents a weak set of principles which imply the LP, and discusses at which point the Barnard-Fraser criticism enters in. Section 4 outlines the scenario through which considerations of coherency and admissibility become relevant. In this section it is argued, as a side issue, that decision theoretic admissibility is a more valid evaluational tool than the more common (to Bayesians) betting coherency. Section 5 argues against violation of the LP because of these considerations, with particular attention given to the Barnard-Fraser criticisms. Section 6 gives some concluding remarks.

2. NOTATION

2.1 The Experiment

We will more or less follow Birnbaum's (1962) notation for reasons of familiarity. The first important issue is - what is an experiment? We will denote an experiment by

$$E = (X, \theta, \{P_\theta\}, (h, \omega)), \tag{2.1}$$

where X (a realization of which is the data and will be denoted x) is a random quantity taking values in a sample space $\mathcal{X}$ according to the probability distribution P_θ, the unknown aspects of this distribution being denoted by θ, an element of the parameter space Θ. (We will not overburden the description of E by including $\mathcal{X}$ or Θ in (2.1).) The unknown θ could include typical unknown parameters and also could index unknown functional features of the distribution. For instance, if $P_\theta(A) = G((A-\mu)/\sigma)$, where μ and σ are unknown and it is only known that $G \in \mathcal{L}$, some set of distributions, then θ would equal (μ,σ,G). We will, however, for simplicity restrict consideration to situations where a likelihood function exists, i.e. where $\{P_\theta, \theta \in \Theta\}$ is dominated by a measure ν, with respect to which we have densities

$$f(x|\theta) = \frac{dP_\theta}{d\nu} .$$

Also, in axiomatic discussions we will implicitly assume that $\mathcal{X}$ is discrete, so that $f(x|\theta)$ is well defined. (Again, as argued in Basu (1975), this assumption reflects ultimate reality; discussion of the philosophical validity of the LP in this setting is thus appropriate.) The density, considered as a function of θ for fixed x, is of course the <u>likelihood function</u> for θ given that x is observed.

The final element of E in (2.1), the pair (h,ω), is included to allow consideration of the criticisms of Barnard and Fraser. Barnard, with his theory of Pivotal Inference (c.f. Barnard (1980, 1982) and Barnard and Sprott (1983)), and Fraser, with his theory of Structural Inference (c.f. Fraser (1968, 1972, 1979)), argue that it may sometimes be known how X, θ, and P_θ are related, and that this can be important information. Thus it may be known that

$$X = h(\theta,\omega),$$

where ω is an unknown random quantity taking values in Ω according to a known distribution Q, and h is a known function from $\Theta \times \Omega \to \mathcal{X}$. (Often in Structural and Pivotal inference, Q is known only to belong to some class $\mathfrak{Q}$. For simplicity, we assume Q is known.) This is actually more or less the "structural" formulation of the problem. The formulation in Pivotal Inference is based on "pivotals" $\omega = g(X,\theta)$ having known distributions. Typically g will be an appropriate inverse function of h, so the two approaches are very related. We will, for the most part, consider the structural formulation, although comments about differences for the pivotal model will be made. The structural model is sometimes called a functional model (c.f. Bunke (1975) and Dawid and Stone (1982)), but we will stick with Fraser's original term. The following example, from Fraser (1968) (and related to an example in Mauldon (1955)), illustrates the key issue.

Example 1. Suppose $X = (X_1, X_2)$, $\theta = (\sigma_1,\tau,\phi)$, and P_θ is bivariate normal with mean zero and covariance matrix

$$\ddagger = \begin{pmatrix} \sigma_1^2 & \tau\sigma_1 \\ \tau\sigma_1 & (\tau^2+\phi^2) \end{pmatrix} .$$

This could arise from either of the following two structural models:

(i) $\omega = (\omega_1, \omega_2)$ is bivariate normal, mean zero and identity covariance matrix, and

$$X = h(\theta,\omega) = (\sigma_1\omega_1, \quad \tau\omega_1+\phi\omega_2); \tag{2.2}$$

(ii) ω is the same but

$$X = h^*(\theta,\omega) = (\tau'\omega_1+\phi'\omega_2, \ \sigma_2\omega_1), \tag{2.3}$$

where $\sigma_2 = \sqrt{\tau^2+\phi^2}$, $\tau' = \sigma_1\phi/\sigma_2$, and $\phi' = \sigma_1\tau/\sigma_2$. In Barnard's setup one would write (2.2) and (2.3) as

$$\omega = (\omega_1,\omega_2) = (X_1/\sigma_1, \ (X_2-\tau X_1/\sigma_1)/\phi), \tag{2.2'}$$

$$\omega = (\omega_1,\omega_2) = (X_2/\sigma_2, \ (X_1-\tau' X_2/\sigma_2)/\phi'), \tag{2.3'}$$

and ω_1 and ω_2 would be the pivotals with known distribution upon which the inference would be based. In pursuing this example later in the paper, we will assume that independent observations $X^1, \ldots, X^n$ from the model are taken, giving the "sufficient" statistic $S = \sum_{i=1}^{n} (X^i)^t(X^i)$, which has a Wishart $(n, \ddagger)$ distribution.

We do not presuppose that knowledge of (h,ω) is available, although our main concern will be the value of knowing (h,ω) if it is available. When it is available, we will say that we are in the "P-S situation" (for pivotal-structural).

2.2 Evidence and Mixtures

Of interest from an experiment is the "evidence" about θ obtained from knowledge of E and the data x. This will (following Birnbaum) be denoted Ev(E,x), and could be any measure or collection of measures whatsoever (including frequency measures). The completely arbitrary nature of this concept should make it acceptable to most people. In Section 3 we will investigate principles that Ev(E,x) should follow. (Dawid (1977) prefers to replace the concept of evidence by that of an inference pattern, and then talk about principles which inference patterns should follow. This might have some philosophical advantage, but we will stick to the more usual approach.)

The final concept needed is that of a simple mixture of two experiments

$$E_1 = (X_1, \theta, \{P_\theta^1\}, (h_1, \omega_1)) \text{ and } E_2 = (X_2, \theta, \{P_\theta^2\}, (h_2, \omega_2)).$$

(It is important to note that, throughout the paper, θ will be assumed to be the same quantity in the two experiments.) Let J be a random variable (independent of the X_i, ω_i, and any information about θ) taking on the values 1 and 2 with probabilities λ and $(1-\lambda)$, respectively. Then the mixed experiment E^λ is the experiment in which first J is observed, and then experiment E_J is performed. Thus the outcome of E^λ is the pair (J, X_J). We will often imagine, given E_1 and E_2, that we can perform E^λ, and hence can think of $Ev(E^\lambda, (j, x_j))$. Certain objections have been raised (c.f. Durbin (1970) and Kalbfleisch (1974)) concerning the treatment of E^λ as a "real" experiment, but it is clear that E^λ could be performed, and hence any proposed statistical methods should work well for it. Further discussion can be found in Birnbaum (1970), Dawid (1977), and Berger and Wolpert (1984).

3. PRINCIPLES OF EVIDENCE

This section is independent of the rest of the paper, and can be skipped by those less interested in axiomatic arguments. No attempt will be made to survey the wide array of principles that have been discussed and which lead to the LP. References include Birnbaum (1962, 1972), Basu (1975), Dawid (1977), Godambe (1979), Barnard and Godambe (1982), and Berger and Wolpert (1984). The key principles are the Weak Conditionality Principle below, and some version of the Sufficiency Principle. The weakest versions of sufficiency are Mathematical Equivalence (see Birnbaum (1972)) and the Distribution Principle (see Dawid (1977)) which essentially state that (h,ω) in (2.1) is irrelevant to EV(E,x). This section is an attempt to formulate weaker principles which lead to the LP. The effort is not as successful as had been initially hoped. We first list the principles.

Weak Conditionality Principle (WCP): $Ev(E^{\lambda}, (j,x_j)) = Ev(E_j,x_j)$. (Thus the evidence obtained from a simple mixture experiment is simply the evidence obtained from the experiment E_j actually performed.)

Weak Ancillarity Principle (WAP): Suppose E_1 (as in (2.1)) consists of observing the random quantity $X_1 = (Y_1,Z)$, where Z has a known distribution independent of Y_1, ω_1, and θ, and E_2 consists of observing $X_2 = g(Y_1,Z)$, where g is a known function for which there exists a known function g* such that $Y_1 = g^*(X_2)$. Then $Ev(E_1,x_1) = Ev(E_2,x_2)$. (The point here is that it seems clear that only Y_1 contains information about θ in E_1, and we can determine Y_1 exactly from X_2. But since X_2 depends only on Y_1 and the irrelevant Z, it can contain no more information about θ than Y_1. Note that passing from X_1 to X_2 and back to Y_1 is via known functions, so the important structural elements of the experiments are preserved.)

Weak Distribution Principle (WDP): Suppose E_i, i=1,2, consists of observing $X_i = (Y, Z_i)$, where Y is the same random variable and the Z_i are each 1 or 2 with probability $\frac{1}{2}$, independent of Y and θ (but not necessarily of the ω_i). Then $Ev(E_1, (y,j)) = Ev(E_2, (y,j))$ for j = 1,2.

It is the WDP that is potentially inconsistent with P-S analysis, in that the Z_i may contain structural information which is ignored because of the lack of dependence of the distribution of the Z_i on Y or θ. In some sense, the WDP is no more intuitively obvious than Birnbaum's principle of Mathematical Equivalence, but does appeal to a somewhat different intuition, namely the "frequentist" intuition behind sufficiency which states that if Y is a sufficient statistic (of (Y,Z_1)) for θ, then Z_1 could be replaced by any random variable with the same distribution without affecting any conclusions.

Likelihood Principle: If E_1 and E_2 are two experiments and there exist observations x_1' and x_2' in the respective experiments for which $f_1(x_1'|\theta) = cf_2(x_2'|\theta)$ for all θ (and some fixed constant c), then $Ev(E_1,x_1') = Ev(E_2,x_2')$.

Theorem 1. The LP is a consequence of the WCP, WAP, and WDP.

Proof. Consider the mixed experiment E^λ, where $\lambda = 1/(1+c)$. By the WCP we have that

$$Ev(E^\lambda, (j,x_j')) = Ev(E_j,x_j'). \tag{2.4}$$

Next, define $Y = (J,X_J)$ (the outcome of E^λ) and

$$V_1 = \begin{cases} Y & \text{if } X_J \neq x_1' \text{ or } x_2' \\ 0 & \text{otherwise,} \end{cases} \tag{2.5}$$

and

$$Z_1 = \begin{cases} 1 & \text{if } X_J = x_1', \text{ or } V_1 \neq 0 \text{ and } Z = 1 \\ 2 & \text{if } X_J = x_2', \text{ or } V_1 \neq 0 \text{ and } Z = 2, \end{cases} \tag{2.6}$$

where Z is a random variable (independent of everything, probabilistically and structurally) taking values 1 and 2 with probability $\frac{1}{2}$. Consider the experiment E^* with observation (Y,Z), and note that an application of the WAP (with $g(Y,Z) = Y$) shows that

$$Ev(E^\lambda, y) = Ev(E^*, (y,z)). \tag{2.7}$$

Also, defining E^{**} as the experiment of observing $X^* = (V_1, Z_1)$, noting that (2.5) and (2.6) define $X^* = g(Y,Z)$, and observing that

$$g^*(X^*) \equiv \begin{cases} V_1 & \text{if } V_1 \neq 0 \\ (Z_1, x_{Z_1}') & \text{if } V_1 = 0 \end{cases} = Y,$$

the WAP can be applied to conclude that

$$Ev(E^*, (y,z)) = Ev(E^{**}, (v_1, z_1)). \tag{2.8}$$

Note that

$$P_\theta(Z_1 = 1 | v_1) = \begin{cases} P(Z=1) = \frac{1}{2} & \text{if } v_1 \neq 0 \\ \dfrac{\lambda f_1(x_1'|\theta)}{\lambda f_1(x_1'|\theta) + (1-\lambda) f_2(x_2'|\theta)} = \frac{1}{2} & \text{if } v_1 = 0, \end{cases}$$

so that Z_1 is independent of both V_1 and θ. Application of the WDP to E^{**} shows that Z_1 could be replaced by any (structurally independent) random variable with the same distribution, and use of the WAP (as in (2.7)) then shows that $Ev(E^{**}, (v_1, z_1))$ does not depend on z_1. Combining this with (2.4), (2.7), and (2.8) yields the desired conclusion.||

Since the WDP is in as much conflict with P-S analysis as Mathematical Equivalence, not much progress has been made. The point of developing alternative simple principles is the hope that they will spur those who question the LP into finding a clear counterexample to at least one of the principles. Unfortunately,"counterexamples" so far developed are either of the extermely involved variety (such as the Stopping Rule Paradox), or are of the form - here is an example of where 'My Method' clashes with Principle A - without a quantified demonstration of harm that would result in following Principle A. Of course, we don't believe valid counterexamples will be found. The reason is simply that repeated use of any method violating the LP seems likely to itself be demonstrably inferior. We turn now to this issue.

4. LONG RUN PERFORMANCE

The practical importance of considering the long run performance of statistical procedures or methods is certainly a matter open to debate, but one feature of long run performance seems clear: it cannot be right (philosophically) to recommend repeated use of a method if the method has "bad" long run properties. There have been two main approaches proposed for long run evaluations: decision theory and betting schemes. We will argue that the decision theoretic approach is the more satisfactory of the two (even for "inference" problems), although either approach strongly contraindicates violation of the LP. It is interesting that frequentist decision theory emerges (even for Bayesians) as an important testing ground for statistical theories.

4.1 Decision Theoretic Evaluations

The decision theoretic approach supposes that the result of the statistical investigation is to take an action $a \in \mathcal{A}$ (which could conceivably be the action to take a particular "inference"), the consequence of which, for given data x and when θ obtains, is the loss $L(x,a,\theta)$. It is also supposed that the statistical method being evaluated provides an action to take for each possible x, thus defining a statistical procedure $\delta(\cdot):\mathcal{X}\to\mathcal{A}$. (For the most part we will stick to nonrandomized procedures for simplicity.) As usual in frequentist decision theory, we define the frequentist risk and the Bayes risk (with respect to a prior distribution π on Θ) as, respectively,

$$R(\theta,\delta) = E_\theta L(X,\delta(X),\theta), \text{ and } r(\pi,\delta) = E^\pi R(\theta,\delta).$$

Of interest will be the following standard definitions. The procedure δ^1 is strictly inadmissible if there exists a δ^2 with $R(\theta,\delta^2) < R(\theta,\delta^1)$ for all θ, and is extended inadmissible if $R(\theta,\delta^2) < R(\theta,\delta^1) - \varepsilon$ for all θ and some $\varepsilon>0$. If $r(\pi,\delta^2) < \infty$ for all countably additive π, the above risk inequalities can be replaced by $r(\pi,\delta^2) < r(\pi,\delta^1)$ and $r(\pi,\delta^2) < r(\pi,\delta_1) - \varepsilon$ for all countably additive π.

Following Hill (1974), and in a similar manner to many betting scenarios, we consider the following game.

Evaluation Game. Player 1 proposes use of δ^1 and Player 2 proposes δ^2. A master of ceremonies will choose a sequence $\underset{\sim}{\theta} = (\theta_1,\theta_2,\ldots)$, and for each θ_i the experiment E will be independently performed yielding an observation X_i (from the distribution P_{θ_i}, or equal to $h(\theta_i,\omega_i)$). Player j will use $\delta^j(x_i)$, paying to the other player his "loss" $L(x_i,\delta^j(x_i),\theta_i)$. After n plays, Player 2 will have won

$$S_n = \sum_{i=1}^{n} [L(x_i,\delta^1(x_i),\theta_i) - L(x_i,\delta^2(x_i),\theta_i)].$$

<u>Theorem 2</u>. In the situation of the above game,

(a) If δ^1 is strictly inadmissible and $r(\pi,\delta^2) < \infty$ for all countably additive π, then

$$P_\pi\,(\liminf_{n\to\infty} \frac{1}{n} S_n > 0) = 1, \tag{3.1}$$

where P_π denotes the joint probability distribution of the X_i and θ_i.

(b) If δ^1 is extended inadmissible and the random variables $Z_i = [L(X_i,\delta^1(X_i),\theta_i) - L(X_i,\delta^2(X_i),\theta_i)]$ have uniformly bounded variances (i.e. $E_{\theta_i}[Z_i - E_{\theta_i} Z_i]^2 < K < \infty$ for all θ_i), then

$$P_{\underset{\sim}{\theta}}\,(\liminf_{n\to\infty} \frac{1}{n} S_n > \varepsilon > 0) = 1, \tag{3.2}$$

for <u>any</u> sequence $\underset{\sim}{\theta} = (\theta_1,\theta_2,\ldots)$.

(c) If δ^1 is strictly inadmissible, Θ is closed, $R(\theta,\delta^1)$ and $R(\theta,\delta^2)$ are continuous in Θ, and the moment condition in (b) holds, then (3.2) is valid for any <u>bounded</u> sequence $\underset{\sim}{\theta}$ (although ε could depend on the bound).

<u>Comment 1</u>. For bounded losses, the moment conditions in the theorem are all clearly satisfied. Even unbounded losses rarely cause a problem.

<u>Comment 2</u>. If (3.2) holds, then it also holds for $P_{\underset{\sim}{\theta}}$ replaced by P_π for any prior π, including finitely additive π. Also, Heath and Sudderth (1978) show that δ^1 is extended admissible only if it is Bayes with respect to some (possibly finitely additive) prior π.

<u>Proof of Theorem 2</u>. (a). If $r(\pi,\delta^1) < \infty$, then $E_\pi Z_i$ (E_π being expectation over the X_i and θ_i, and Z_i being as in part (b)) has finite expectation $\Delta_\pi = r(\pi,\delta^1) - r(\pi,\delta^2) > 0$. The result follows from the strong law of large numbers. If $r(\pi,\delta^1) = \infty$, the result follows by truncating the loss at a suitably large level.

(b). Define

$$\psi(\theta_i) = E_{\theta_i}(Z_i) = R(\theta_i,\delta^1) - R(\theta_i,\delta^2) > \varepsilon.$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}[Z_i - \psi(\theta_i)] \to 0 \text{ almost surely,}$$

and the result follows easily. The proof of part (c) is similar.||

The Evaluations Game seems to be a reasonably fair way of testing the performance of a procedure. If δ^1 is certain to lose an arbitrarily large amount in comparison with δ^2, as occurs in the situations of Theorem 2, then δ^1 would seem to be theoretically inferior. (The word "theoretically" is inserted, because the practical difference in a realistic finite number of uses may be negligible.) Extended inadmissibility seems very serious, in that δ^1 would always have a long run loss. Strict inadmissibility is less compelling, in that δ^1 is only guaranteed to lose against a countably additive prior π, or a bounded sequence $\underset{\sim}{\theta}$ (in the situation of part (c) of the theorem). But the fact that it will lose for any such π (even one of Player 1's choosing) or for any such $\underset{\sim}{\theta}$, strikes us as sufficient reason to perceive δ^1 as not being fundamentally sound. Note that it is not necessary to know the bound on $\underset{\sim}{\theta}$ in case (c) to conclude that δ^1 loses. It is only necessary to know that there is some bound. (And, in reality, $\underset{\sim}{\theta}$ will be bounded; unbounded Θ are typically used only because one is not sure what the bound on θ should be.) Again, it may be that δ^1 is justifiable as a good approximate rule, even if it is strictly or extended inadmissible, but we would certainly hesitate to call any statistical method which led to δ^1 a fundamentally sound method.

Adopting a decision-theoretic viewpoint for evaluation can be criticized, especially for inference problems in which losses (if they exist at all) are vague or hard to formulate. This is not the place to argue the case for a decision-theoretic outlook, and indeed a justification of decision theory is not needed for our purpose here. Our goal is to judge the claim in P-S analysis (and other approaches) that the LP is invalid, because it ignores important features of the experiment. We will essentially try to argue that, in any decision problem, repeated violation of the LP will result in long run loss. Most statisticians would probably have qualms about trying to argue that, even if the LP should be followed in any decision problem, it need not be followed in inference problems. Essentially such an argument would be of the variety - "I know I'm right, but will not allow any quantifiable evaluation of my methods."

We will avoid the "unfair" possibility of taking an inference procedure and evaluating it with respect to a particular loss function. It is somewhat more fair to evaluate it with respect to a very wide range of loss functions (indeed, if a wide enough range of loss functions is allowed many "inadmissible" inference procedures become admissible, c.f. Brown (1973)), and strict inadmissibility for a wide range of reasonable losses should be a serious concern. More commonly, however, we will consider particular losses as given, and see where the following of P-S reasoning might lead us. Criticizing P-S reasoning (in particular, possible violation of the LP) in decision settings for which it was never intended is, of course, an uncertain undertaking, especially since it is not clear what P-S reasoning in decision contexts would be. Of relevance here is the following comment of Hill (1974):

> "But no matter what is meant by inference, if it is to be of any value, then somehow it must be used, or acted upon, and this does indeed lead back to the decision-theoretic framework. I suspect that for some 'inference' is used as a shield to discovery that their actions are incoherent."

As a final comment, it should be mentioned that even Bayesians should be willing to submit their procedures (usually derived conditionally) to long run performance evaluations, especially when robustness is a serious concern or when improper prior distributions were used.

4.2 Betting Evaluations

Studying coherence in betting has a long tradition in statistics, especially Bayesian statistics. The typical scenario deals with evaluation of methods (usually inference methods) which produce, for each x, either a probability distribution for θ, say $q_x(\theta)$ (which could be a posterior distribution, a fiducial distribution, a structural distribution, etc.), or a system of confidence statements $\{C(x), \alpha(x)\}$ with the interpretation that θ is felt to be in $C(x)$ with probability $\alpha(x)$. For simplicity, we will restrict ourselves to the confidence statement framework; any $\{q_x(\theta)\}$ can be at least partially evaluated through confidence statements by choosing $\{C(x)\}$ and letting $\alpha(x)$ be the probability (with respect to q_x) that θ is in $C(x)$.

The assumption is then made (more on this later) that, since $\alpha(x)$ is thought to be the probability that θ is in $C(x)$, the proposer of $\{C(x), \alpha(x)\}$ should be willing to make both the bet that θ is in $C(x)$ at odds of $(1-\alpha(x))$ to $\alpha(x)$, and the bet that θ is not in $C(x)$ at odds of $\alpha(x)$ to $(1-\alpha(x))$. An evaluations game, as in Section 4.1, is then proposed, where the master of ceremonies again generates θ_i and X_i, Player 1 stands ready to accept bets on $\{C(x), \alpha(x)\}$, and Player 2 bets $s(x)$ at odds determined by $\alpha(x)$. Here, $s(x) = 0$ means no bet is offered; $s(x) > 0$ means that an amount $s(x)$ is bet that $\theta \in C(x)$; and $s(x) < 0$ means that the amount $|s(x)|$ is bet that $\theta \notin C(x)$. (As discussed in Robinson (1979a), restricting $s(x)$ to satisfy $|s(x)| \leq 1$ is also sensible.) The winnings of Player 2 at the ith play are

$$W_i = [I_{C(x_i)}(\theta_i) - \alpha(x_i)]\, s(x_i),$$

where $I_A(\theta)$ is 1 if $\theta \in A$ and 0 otherwise, and of interest is again the limiting behavior of $\frac{1}{n} \sum_{i=1}^{n} W_i$. If

$$P_{\underset{\sim}{\theta}} (\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i > \varepsilon > 0)$$

for all sequences $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots)$, then $\{C(x), \alpha(x)\}$ is called incoherent, or alternatively $s(x)$ is said to be a super relevant betting strategy. If it is merely the case that for θ_i generated according to any countably additive π,

$$P_{\pi}(\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i > 0),$$

then $\{C(x), \alpha(x)\}$ is weakly incoherent or $s(x)$ is weakly relevant. (These concepts can be found in this or related form in such works as Buehler (1959, 1976), Wallace (1959), Freedman and Purves (1969), Cornfield (1969), Pierce (1973), Bondar (1977), Heath and Sudderth (1978), Robinson (1979a, 1979b), and Lane and Sudderth (1983). Other general Bayesian works on coherency include Ramsey (1926), deFinetti (1937, 1974), Savage (1954), and Levi (1980).)

If $\{C(x), \alpha(x)\}$ is incoherent or weakly incoherent, then Player 1 will for sure lose money in the appropriate evaluations game, which certainly casts doubt on the validity of the probabilities $\alpha(x)$. A number of objections to the scenario can, and have, been raised, however, and careful examination of these objections is worthwhile.

Objection 1. Player 1 will have no incentive to bet unless he perceives the odds as slightly favorable. This turns out to be no problem if incoherence is present, since the odds can be adjusted by $\varepsilon/2$ in Player 1's favor, and

Player 2 will still win. If only weak incoherence is present, it is still often possible to adjust the odds by a function g(x) so that Player 1 perceives that the game is in his favor, yet will lose in the long run, but this is not clearly always the case.

Objection 2. Weak incoherence has been deemed not very meaningful, since a sequence $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots)$ could be chosen so that Player 1 is not a sure loser. However, the fact that Player 1 is a sure loser for any π (even one selected by himself) or any bounded $\underset{\sim}{\theta}$ (under certain reasonable assumptions) seems quite serious.

Objection 3. Of course, frequentists who quote a confidence level α for $\{C(x)\}$ remove themselves from the game, since they do not claim that α is the probability that θ is in $C(x)$, and hence would find the betting scenario totally irrelevant.

Objection 4. The game is unfair to Player 1, since Player 2 gets to choose when, how much, and which way to bet. Various proposals have been made to "even things up." The possibility mentioned in Objection 1 is one such, but doesn't change the conclusions much. A more radical possibility, suggested by Fraser (1977), is to allow Player 1 to decline bets. This can have a drastic effect, but strikes us as too radical, in that it gives Player 1 license to state completely silly $\alpha(x)$ for some x. It is after all $\{\alpha(x)\}$ that is being tested, and testing should be allowed for all x.

Objection 5. The most serious objection we perceive to the betting game is that $\{\alpha(x)\}$ is generally not selected for use in the game, but rather to communicate information about θ. It may be that there is no better choice of $\{\alpha(x)\}$ for communicating the desired information. Consider the following example, which can be found in Buehler (1971), and is essentially successive modifications by Buehler and H. Rubin of an earlier example of D. Blackwell.

Example 2. Suppose $X = \theta + \omega$, where $P(\omega=1) = P(\omega=-1) = \frac{1}{2}$, and $\theta \in \Theta = \{\text{integers}\}$. We are to evaluate the confidence we attach to the sets $C(x) = \{x+1\}$ (the point $(x+1)$), and a natural choice is $\alpha(x) = \frac{1}{2}$ (since θ is either $x-1$ or $x+1$, and in the absence of fairly strong prior information about θ, either choice seems equally plausible). This choice can be beaten in the betting game, however, by betting that θ is not in $C(x)$ with probability $g(x)$, where $0 < g(x) < 1$ is an increasing function. (Allowing Player 2 to have a randomized betting strategy does not seem unreasonable.) Indeed, the expected gain per bet of one unit, for any countably additive π on Θ, is $E^{\pi}[g(\theta+1) - g(\theta-1)] > 0$, so that $\alpha(x) = \frac{1}{2}$ is weakly incoherent. (A continuous version of this example, mentioned in Robinson (1979a), has ω normal $(0,1)$, $\Theta = \mathbb{R}^1$, $C(x) = (-\infty, x)$, and $\alpha(x) = \frac{1}{2}$. Earlier examples of similar phenomenon include the usual Student -t intervals, c.f. Stein (1961), Buehler and Fedderson (1963), and Brown (1967), and confidence intervals in the Behrens-Fisher problem, c.f. Fisher (1956). It should also be noted that Fisher originated the idea of looking at confidence, conditional on "relevant" subsets, which is the basis for many of the betting examples.)

In this and other examples where $\{\alpha(x)\}$ loses in betting, one can ask the crucial question - Is there a better α that could be used? The question has no clear answer, because the purpose of α is not clearly defined. One possible justification for $\alpha(x) = \frac{1}{2}$ in the above example is that it is the unique limiting probability of $C(x)$ for sequences of what could be called increasingly vague prior distributions (c.f. Stone (1970)). (A more formal Bayesian justification along these lines would be a robust Bayesian justification, to the effect that the class of possible priors is so large that the range of possible posterior probabilities for $(-\infty,x)$ will include 1/2 for all x.) An alternative justification can be found by retreating to decision theory, attempting to quantify how well $\alpha(x)$ performs as an indicator of whether or not θ is in $C(x)$, and then seeing if there is any better α. For instance, using the quadratic scoring function of deFinetti (1962) (any proper scoring function is a possibility - see Good (1952), Savage (1971), Buehler (1971), and Lindley (1982) for other scoring functions) as an indicator of how well $\alpha(x)$ performs, would mean considering the loss function

$$L(x,\alpha(x),\theta) = (I_{C(x)}(\theta) - \alpha(x))^2. \tag{3.3}$$

(For the moment, we are considering $\{C(x)\}$ as given, and worrying only about the choice of α. Note that, for any "posterior" distribution on θ, the optimal choice of $\alpha(x)$ for (3.3) is the posterior probability of $C(x)$, so (3.3) is a natural measure of the accuracy of α.) One can then ask if there is a better α in terms of (3.3), employing usual decision-theoretic ideas. The answer in the case of Example 2 is - no. A standard limiting Bayes argument can be used to show that $\alpha(x) = \frac{1}{2}$ is admissible for this loss, and hence no improvement (for all θ or all π) is possible. (The same cannot necessarily be said, however, if choice of $C(x)$ is brought into the picture. For instance, a reasonable overall loss for $\{C(x), \alpha(x)\}$ is

$$L(C(x),\alpha(x),\theta) = c_1(I_{C(x)}(\theta)-\alpha(x))^2 + c_2(1-I_{C(x)}(\theta)) + c_3\mu(C(x)),$$

where c_i are constants and μ is a measure of the size of $C(x)$. It can be shown in Example 2 that $\{C^*(x),\alpha^*(x)\}$, with $\alpha^*(x) \equiv \frac{1}{2}$ and

$$C^*(x) = \begin{cases} \{x-1\} & \text{with probability } g(x) \\ \{x+1\} & \text{with probability } 1-g(x), \end{cases}$$

is a better procedure than the given $\{C(x),\alpha(x)\}$.)

Decision-theoretic inadmissibility, with respect to losses such as (3.3), can be related to incoherency, and seems to be a criterion somewhere between weak incoherency and incoherency (c.f. Robinson (1979a)). This supports the feeling that it may be a more valid criterion than the betting criterion. This is not to say that the betting scenarios are not important. Buehler, in discussion of Fraser (1977), makes the important point that, at the very least, betting scenarios show when quantities such as $\alpha(x)$ "behave differently from ordinary probabilities." And as Hill (1974) says

> "...the desire for coherence...is not primarily because he fears being made a sure loser by an intelligent opponent who chooses a judicious sequence of gambles...but rather because he feels that incoherence is symptomatic of something basically unsound in his attitudes."

Nevertheless, Objection 5 often prevents betting incoherency from having a conclusive impact, and so decision-theoretic inadmissibility (with respect to an agreed upon criterion) is more often convincing.

Decision-theoretic methods of evaluating "inferences" such as $q_x(\theta)$ (i.e., distributions for θ given x) have also been proposed (c.f. Gatsonis (1981) and Eaton (1982)). For the most part, however, there has been little attention directed to these matters.

5. VIOLATION OF THE LP: INADMISSIBILITY AND INCOHERENCY

A violation of the LP will occur when there are two experiments E_1 and E_2, with $x_1' \in \mathcal{X}_1$ and $x_2' \in \mathcal{X}_2$ satisfying (for some positive constant c)

$$f_1(x_1'|\theta) = cf_2(x_2'|\theta) \text{ for all } \theta, \tag{5.1}$$

and for which different actions or conclusions would be recommended were x_1' or x_2' observed. Using the notation of Section 3, it is thus felt that

$$Ev(E_1,x_1') \neq Ev(E_2,x_2'). \tag{5.2}$$

Consider, in this situation, the mixed experiment $E^{1/2}$, in which $J = 1$ or 2 with probability $\frac{1}{2}$ each is observed (independent of elements of the E_i, both probabilistically and structurally), and experiment E_J is then performed. The Weak Conditionality Principle states that

$$Ev(E^{1/2},(j,x_j)) = Ev(E_j,x_j),$$

which combined with (5.2) yields the conclusion

$$Ev(E^{1/2},(1,x_1')) \neq Ev(E^{1/2},(2,x_2')). \qquad (5.3)$$

Since x_1' and x_2' have proportional likelihood functions, behaving as in (5.3) violates sufficiency and cannot be Bayesian, and will be seen to entail inadmissibility and incoherency in a variety of situations. Note that the experiment $E^{1/2}$ preserves all structural features of E_1 and E_2, so the only possible objection to concluding (5.3) would be to the use of the WCP. It is our understanding, however, that Pivotal, Structural, and virtually all other approaches (except, of course "pure" frequentist theory) accept and extensively use the WCP.

As a final comment before proceeding, note that the formal setup implicitly involves discrete $\mathcal{X}$. As mentioned earlier, however, versions of the LP can be developed for continuous $\mathcal{X}$, the only essential difference being that (5.1) should be replaced by

$$f_1(x_1|\theta) = c(x_1)\ f_2\ (g(x_1)|\theta) \text{ for all } \theta \text{ and } x_1 \in B, \qquad (5.4)$$

where $B \subset \mathcal{X}$ has $P_\theta^1(B) > 0$ for all θ, $c(x_1) > 0$ for $x_1 \in B$, and $g\colon \mathcal{X}_1 \to \mathcal{X}_2$ is one-to-one. (For a formulation of this without the assumption of densities, see Berger and Wolpert (1984).) All subsequent expressions should then be understood to hold with x_1' replaced by x_1 (in B) and x_2' replaced by $g(x_1)$. The set B is thus to be a set of positive measure for which x_1 and $x_2 = g(x_1)$ have proportional likelihood functions, and yet supposedly call for differing actions or conclusions. We will usually refrain from explicitely stating conditions or results in the continuous setting, but will nevertheless consider important continuous examples. We look first at the situation from a decision-theoretic viewpoint.

5.1 Decision - Theoretic Evaluation

Suppose $Ev(E^{1/2},(j,x_j))$ is decision-theoretic in nature, consisting of the action to be taken when (j,x_j) is observed, to be denoted $\delta((j,x_j)) = \delta_j(x_j)$, along with knowledge that the loss $L((j,x_j), \delta_j(x_j),\theta)$ is to be suffered. Note that this includes decision-theoretic inference, as mentioned in Section 4.2. We will only consider situations in which

$$L((1,x_1'),a,\theta) = L((2,x_2'), a,\theta) \text{ for all } a \text{ and } \theta, \qquad (5.5)$$

so that any Bayes rule would be the same whether $(1,x_1')$ or $(2,x_2')$ is observed. (Of course, losses are usually of the form $L(a,\theta)$, with no dependence on the data. The possibility of data dependence is allowed to deal with losses like (3.3).) Because of (5.3), we are assuming that the proposed procedure $\delta^0((j,x_j)) = \delta_j^0(x_j)$ satisfies

$$\delta_1^0(x_1') \neq \delta_2^0(x_2'), \qquad (5.6)$$

and are out to establish that this is inadmissible. We first look at the most clearcut situation, that of convex loss.

5.1.1 Convex Loss

Suppose $\mathcal{Q}$ is convex, and that L satisfies (5.5) and is strictly convex in a for all θ and the observations (j,x_j'). Then the procedure

$$\delta^*((j,x_j)) = \begin{cases} \frac{c}{(c+1)}\,\delta_1^0(x_1')+\frac{1}{(c+1)}\,\delta_2^0(x_2') & \text{for } x_j = x_1' \text{ or } x_2' \\ \delta_j^0(x_j) & \text{otherwise,} \end{cases}$$

(where c is from (5.1)) satisfies (using (5.5) and strict convexity)

$$L((j,x_j'),\delta^*((j,x_j')),\theta) < \frac{c}{(c+1)}\,L((1,x_1'),\delta_1^0(x_1'),\theta) + \frac{1}{(c+1)}\,L((2,x_2'),\delta_2^0(x_2'),\theta). \quad (5.7)$$

An easy calculation, using (5.1) and (5.5), then shows that (for the Experiment $E^{1/2}$)

$$R(\theta,\delta^0)-R(\theta,\delta^*) = \frac{(1+c)}{2c}\,f_1(x_1'|\theta)\,\Delta(\theta) > 0$$

where $\Delta(\theta)$ is the difference between the right and left hand sides of (5.7). The following lemma is immediate. (This is all, of course, a simple form of the Rao-Blackwell theorem.)

Lemma 1. In the above situation, δ^0 is

(a) strictly inadmissible, providing x_1' has positive probability for all θ;

(b) extended inadmissible if, in addition to (a), $f_1(x_1'|\theta)$ and L are continuous in θ, and Θ is compact.

(The continuous analog of this lemma is also very easy.)

Example 3. Suppose E_1 is binomial (n,θ) and E_2 is negative binomial (m,θ), where X_2 is the number of failures and m, the number of successes at which experimentation stops, is less than n. The densities are $f_1(x_1|\theta) = \binom{n}{x_1}\theta^{x_1}(1-\theta)^{n-x_1}$ and $f_2(x_2|\theta) = \binom{m+x_2-1}{x_2}\theta^m (1-\theta)^{x_2}$, which are proportional when $x_1 = m \equiv x_1'$ and $x_2 = n-m \equiv x_2'$. Thus, in the mixed experiment $E^{1/2}$, the LP would call for the same action to be taken if either $(1,x_1')$ or $(2,x_2')$ were observed.

If now the goal is to estimate θ under quadratic loss $L = (\theta-a)^2$ (or any other strictly convex loss), and one uses different estimates of θ for $(1,x_1')$ and $(2,x_2')$, then Lemma 1(a) applies, and the behavior is strictly inadmissible. (Neither Pivotal nor Structural analysis would necessarily say that different actions should be taken in this problem, but Akaike (1982), in criticizing the LP, seems to say that different analyses are called for).

Example 1 (continued). Example 1 is an example where the same probability distribution can arise from different structural models. One component of Structural analysis is that of construction of "structural distributions" for any θ given the data, which can presumably be used, as are posterior or fiducial distributions, to make inferences or probability statements about θ. The structural densities, based on S, for $\theta = (\sigma_1,\tau,\phi)$ are given for the two models (2.2) and (2.3), respectively, by (see Fraser (1968))

$$\pi_1(\theta|s) = K_1(s)f(s|\sigma_1,\tau,\phi)\sigma_1^{-2}\phi^{-1}, \tag{5.8}$$

$$\pi_2(\theta|s) = K_2(s)f(s|\sigma_1,\tau,\phi)(\tau^2+\phi^2)^{-1}\phi^{-1}. \tag{5.9}$$

(These correspond to the posterior distributions with respect to the right invariant Haar measures on the lower and upper triangular group decompositions of Σ.)

We now consider the mixed experiment $E^{1/2}$, where E_1 and E_2 are the experiments of observing S (or really $X^1,\ldots,X^n$) from the models (2.2) and (2.3), respectively. Following structuralist theory (maintaining compatibility with the WCP), gives as the structural distribution for $E^{1/2}$

$$\pi(\theta|(j,s)) = \pi_j(\theta|s). \tag{5.10}$$

Note that the LP applies with $B = \mathcal{X}_1$ (see (5.4)), g chosen to be the identity map, and $c(x_1) = 1$. Thus, if different actions are to be taken for (1,s) and (2,s), as could well be called for if (5.10) is used, then strict inadmissibility can result. We consider two examples.

<u>Case 1</u>. Suppose it is desired to estimate Σ (which is equivalent to θ) under the strictly convex loss

$$L(\delta,\Sigma) = \mathrm{tr}(\delta\Sigma^{-1}) - \log\det(\delta\Sigma^{-1})-2. \tag{5.11}$$

(The loss $L(\delta,\Sigma) = \mathrm{tr}(\delta\Sigma^{-1}-I)^2$ would work similarly. The losses and following results are all well known, and are discussed in James and Stein (1961), Eaton (1970), Selliah (1964), and Takemura (1982).) It can be shown that the optimal estimator for this loss and the structural distribution (5.10) (treated as a posterior) is $\delta^0((j,s)) = \delta_j^0(s)$, where

$$\delta_1^0(s)=s_L\begin{pmatrix}\frac{1}{n+1} & 0\\ 0 & \frac{1}{n-1}\end{pmatrix}s_L^t,\ \delta_2^0(s)=s_U\begin{pmatrix}\frac{1}{n-1} & 0\\ 0 & \frac{1}{n+1}\end{pmatrix}s_U^t, \qquad (5.12)$$

where $s = s_L s_L^t = s_U s_U^t$, s_L and s_U being lower and upper triangular, respectively. Also, $\delta_1^0(S)$ and $\delta_2^0(S)$ will differ with probability one, so the inequality (5.7) will hold with probability one. Thus δ^0 is strictly inadmissible. (For indications of how much improvement over δ^0 is possible, see Takemura (1982).)

Case 2. Suppose someone wants to know the "confidence" to be attached to a set $C \subset \mathcal{X}$, based on observation of (j,s) from $E^{1/2}$. Presumably the structuralist would assign confidence

$$\alpha((j,s)) = \int_C \pi_j(\theta|s)d\theta. \qquad (5.13)$$

If now the "success" of such an inference is measured by an "inference loss" such as (3.3), which does satisfy (5.5) and is strictly convex in α, then strict inadmissibility results if C is such that $\alpha((1,S))$ and $\alpha((2,S))$ differ with nonzero probability (and there are many such C). Clearly, many other variations on this theme are possible.

5.1.2 Complete Class Theorems

Convex losses are, of course, rather special, and it would be nice to have general theorems concerning strict inadmissibility for other situations.

We review below some general theorems from statistical decision theory which can be of use in establishing strict inadmissibility. We use the common terminology that an <u>essentially complete</u> class of decision rules $\mathcal{C}$ is a class such that, if $\delta^0 \notin \mathcal{C}$, then there exists a $\delta^* \in \mathcal{C}$ such that

$$R(\theta,\delta^*) \le R(\theta,\delta^0) \quad \text{for all } \theta, \tag{5.14}$$

while a <u>complete class</u> $\mathcal{C}$ is a class such that, in addition, $R(\theta,\delta^*) < R(\theta,\delta^0)$ for some θ. We are informal about technical conditions in the following.

<u>Theorem 3</u>. The class of all decision rules based on a sufficient statistic forms an essentially complete class (c.f. Ferguson (1967) or Berger (1980)).

Our interest in this result is, of course, that if (5.1) (or (5.4)) hold, then

$$T(J,X_J) = \begin{cases} (J,X_J) & \text{if } X_J \neq x_1' \text{ or } x_2' \text{ (or } x_1 \text{ or } g^{-1}(x_2) \notin B \\ & \text{for the continuous case)} \\ X_1 & \text{otherwise} \end{cases}$$

is a sufficient statistic, and hence we can find a procedure based only on T (and hence satisfying the LP) which is as good as a procedure δ^0 satisfying (5.6). Unfortunately, we desire to show more: namely, that δ^0 is strictly inadmissible. There are usually two problems in doing this. First, it is necessary to show that the inequality in (5.14) is strict for some θ (i.e., that procedures based on a sufficient statistic form a complete class); and second, that the inequality can be extended to hold for all θ, or at least for all θ in the support of any possible prior π. Considering the last problem first, the following lemma is an easy consequence of complex analysis.

Lemma 2. Suppose Θ is a subset of $\mathbb{R}^m$, that (5.14) holds with strict inequality for some θ, and that both risk functions are analytic functions in each coordinate (as is frequently the case when dealing with exponential families and otherwise). Then $R(\theta,\delta^*) < R(\theta,\delta^0)$, except possibly for $\theta \in \Theta^*$, some set of discrete points having no limit point.

When the conditions of the lemma are satisfied, δ^0 is clearly strictly inadmissible for all nonatomic countably additive priors π. Usually it is possible to do even more: by slightly altering δ^* (using, say, a local averaging process), one can often get strict risk inequality for all θ.

The other concern mentioned above, that equality could hold in (5.14) for all θ, is more of a problem. We discuss below a few of the ways in which this could be attacked.

Lemma 3. If $R(\theta,\delta^*) = R(\theta,\delta)$ implies that $\delta^* = \delta$ (with probability 1 for all θ), then procedures based on a sufficient statistic form a complete class (and hence inequality will hold in (5.14) for at least one θ).

The condition in Lemma 3 can be verified for a number of situations (besides the obvious one of a convex loss). For instance, if the nonrandomized rules form a complete class (as in finite action problems with nonatomic densities - see, e.g., Dvoretsky, Wald, and Wolfowitz (1951), and in location parameter problems - see, e.g. Farrell (1964)), then this is easily seen to be satisfied. (If $R(\theta,\delta^*) = R(\theta,\delta)$, the randomized rule which chooses between δ^* and δ with probability $\frac{1}{2}$ has the same risk, but by assumption can be improved upon.) Another possible situation in which this can be verified is when $R(\theta,\delta)$ can be expressed as $E_\theta\, \psi(\delta(X))$ (thus $\psi(\delta(X))$ is an "unbiased estimator of risk" - c.f. Stein (1981)), ψ is a one-to-one operator, and $\{P_\theta\}$ is a complete family of distributions. For other situations in which the condition of Lemma 3 can be satisfied, see Brown, Cohen, and Strawderman (1980).

Lemma 4. The Bayes rules form a complete class if Θ is compact, $R(\theta,\delta)$ is continuous in θ for all δ, $L(\theta,\cdot)$ is lower semicontinuous, and $\mathcal{a}$ is a complete separable metric space and is compact or has a suitable compactification (c.f. Wald (1950) or Brown (1976)).

The Bayes rules also form a complete class in certain testing situations with compact null hypothesis (c.f. Brown, Cohen, and Strawderman (1980)). The use of such a result is that it can sometimes be verified directly that δ^0 cannot be a Bayes rule. More generally, a complete class can sometimes be shown to consist of appropriate limits of Bayes rules, and it may be possible to show that δ^0 cannot be a limit of Bayes rules.

A number of examples of strict inadmissibility in, say, Example 1 could be developed using the results in this subsection. For instance, testing or finite action problems could be formulated, in which choosing the action according to the structural distribution (5.10) results in (effectively) a randomized and hence inadmissible rule, which could furthermore be shown to be strictly inadmissible via analyticity and monotonicity arguments. The point of this section, however, was more to convey the feeling that violation of the LP is in general, likely to result in some form of inadmissibility.

5.2 Betting Evaluations

First, it is easy to show that violation of the LP leads to, at least, weak incoherence.

Theorem 4. Consider the mixed experiment $E^{1/2}$ in the discrete case, and suppose a fixed set $C \subset \Theta$ is assigned "confidence" $\alpha((j,x_j'))$ when (j,x_j') is observed, where $\alpha((1,x_1')) \equiv \alpha_1 \neq \alpha_2 \equiv \alpha((2,x_2'))$. Consider the following strategy: define

$$L = \begin{cases} 1 & \text{if } \alpha_1 < \alpha_2 \\ 2 & \text{if } \alpha_2 < \alpha_1 \end{cases},$$

place no bet unless X_J equals x_1' or x_2', and then bet $c_J\alpha_J$ that $\theta \in C$ if $J = L$ and $c_J(1-\alpha_J)$ that $\theta \notin C$ if $J \neq L$, where $c_1 = 1$ and $c_2 = c$ (from (5.1)). The expected gain for this betting strategy (assuming odds corresponding to α_J are being given if (J,x_J') is observed) is $\frac{1}{2}f_1(x_1'|\theta)|\alpha_1-\alpha_2|$. Hence the probabilities assigned to C are weakly incoherent if $f_1(x_1'|\theta) > 0$ for all θ.

Proof. A straightforward calculation, using (5.1).||

Comment 1. In the continuous case, the theorem also holds, the only changes needed being the replacement of α_j by $\alpha_j(x_j)$ (defined as the "confidence" in C if (j,x_j) is observed), betting only if $j = 1$ and $x_1 \in B$ or $j = 2$ and $x_2 \in g(B)$ (see (5.4)), letting $L = L(x^*) = 1$ if $\alpha_1(x_1) < \alpha_2(g(x_1))$ and letting $L = 2$ otherwise (where x^* denotes occurrence of x_1 or $x_2 = g(x_1)$), and replacing c_j by $c_j(x_j)$, defined as 1 for $j = 1$ and as $c(g^{-1}(x_2))$ (see (5.4)) for $j = 2$. The expected gain is then

$$\int_B \frac{1}{2}\left|\alpha_1(x_1) - \alpha_2(g(x_1))\right| f_1(x_1|\theta)d\theta.$$

Comment 2. The above results only prove weak incoherence, although under suitable extra conditions they imply incoherence. To prove incoherence in general, results such as those in Heath and Sudderth (1978) and Lane and Sudderth (1983) can often be employed. Under minor technical conditions, these papers show that the quoted posterior (or structural distribution)

$\pi(\theta|(j,x_j))$ for $E^{1/2}$ is incoherent unless it is the posterior for a (in general, finitely additive) prior on Θ. This seems unlikely to be the case if the LP has been violated. Situations like Example 1 require the finitely additive theorems, and it is not clear whether (5.10) is - or is not - the posterior for a finitely additive prior (we would guess not), but Example 3 requires only the countably additive theorems of Lane and Sudderth (1983) (which apply essentially whenever $\mathcal{X}$ or Θ is compact).

5.3 The Stone Example

As a final interesting example of these ideas, consider the well known example of Stone (1976) (essentially done earlier by Piesakoff (1950) in a less entertaining fashion), in which a soldier leaves a bar at 0, walking one block in each of a succession of randomly selected directions N, S, E, W, the only restriction being that he never immediately backtracks. The soldier trails a taut string, and at some point stops and buries a treasure. Let θ denote the path (a succession of the symbols N, S, E, and W) to this point. (Thus Θ is effectively the free group on two generators.) The soldier then picks a random direction (by spinning a lady) walks one block in that direction (drawing up the string if necessary) and passes out. We observe his complete path X the next day, and have one guess as to where the treasure is. Letting ω denote a random variable that is N, S, E, or W with probability $\frac{1}{4}$ each, the above can be modeled "structurally" as $X = \theta\omega$, with the convention that, if the last two symbols in a string are opposite directions, they cancel.

Although, given x, the likelihood function for θ assigns value $\frac{1}{4}$ to each of xN, xS, xE, and xW, and it would seem that this is the natural structural distribution for θ (obtained by taking θ to be $x\omega^{-1}$ with the probabilities associated with ω), Fraser (in the discussion of Stone (1976)) argues that the correct structural distribution is

$$\pi_1(\theta|x) = \begin{cases} \frac{3}{4} & \text{if } x \text{ is an extension of } \theta \\ \frac{1}{12} & \text{if } \theta \text{ is an extension of } x \\ 0 & \text{otherwise.} \end{cases}$$

In arguing against the LP, Fraser, in the discussion of Hill (1981), constructs an alternative model for X and θ as follows. Let x_0 denote a particular fixed path, suppose θ is as above, and let X be determined according to the probabilities $P(X=0|\theta=x_0) = 1$; $P(X=x_0|\theta=x_0z) = \frac{1}{4}$ and $P(X=0|\theta=x_0z) = \frac{3}{4}$ for z = N, S, E, and W; and $P(X=\theta|\theta) = \frac{1}{4}$ and $P(X=0|\theta) = \frac{3}{4}$ for other θ. (The soldier trails an elastic string, and after burying the treasure at the end of $\theta \neq x_0$ he passes out and has a 75% chance of being snapped back to 0; the end of x_0, however, is very slippery, so if the soldier buries the treasure there and passes out he will be snapped back to 0 for sure. There also happens to be a "lady" who walks the streets within one block of the end of x_0 (her place of business), and if the soldier passes out at x_0N, x_0S, x_0W, or x_0E and doesn't get snapped back to 0, the lady will take him back to x_0.)

For this model, if the observation is $X = x_0$, the likelihood function again assigns value $\frac{1}{4}$ to each of x_0N, x_0S, x_0E, and x_0W, and due to additional "symmetry" in the model, Fraser feels that this is an accurate representation of the probabilities of each θ: thus it appears that the recommended structural distribution (when x_0 is observed) is

$$\pi_2(\theta|x_0) = \frac{1}{4} \quad \text{for } \theta = x_0N, x_0S, x_0E, \text{ and } x_0W.$$

Consider now the mixed experiment $E^{1/4}$. The choice $\lambda = \frac{1}{4}$ is more convenient than $\lambda = \frac{1}{2}$. (For instance, suppose that on $\frac{1}{4}$ of the nights the soldier leaves the bar at 0 with a lady and follow's Stone's scenario, and the other $\frac{3}{4}$ of the nights he leaves alone and follows the alternate scenario. Upon finding the soldier in the morning we know which scenario eventualized, because if he is unaccompanied when he passes out he bumps his head as he falls.) By the WCP, the structural distribution for θ in $E^{1/4}$ should satisfy

$$\pi(\theta|(j,x_0)) = \pi_j(\theta|x_0). \qquad (5.15)$$

Needless to say, if someone were to repeatedly act in accordance with (5.15), he would be operating in an inferior fashion. From a betting viewpoint this is easy to establish. Let us instead, however, consider a decision-theoretic scenario.

A forgetful professor follows a permanently placed string to his office each day. The path happens to be x_0. Every so often, he notices a string running parallel to his and, if the strings run together for the entire journey, the professor is intrigued and mentions the curiosity to a soldier he sometimes finds at the end of x_0. The soldier, in such cases, tells the professor the situation and offers to let him look in one direction for the treasure (which is worth one unit). The soldier requires, however, that the professor pay him for this privilege, the cost being .7 if the soldier had not bumped his head and .2 if he had bumped his head (in which case he is confused and sells out cheaply). In the first case (J=1) the professor following (5.15) is 75% "sure" that the treasure can be found by backtracking one block, and in the second case (J=2) feels that any direction is equally likely. He, hence, accepts the offer of the soldier in either case. For definiteness, let us suppose that, when J=2, the professor randomly chooses one of the three directions that extend the path x_0.

Let $m(x_0)$ denote the probability with which the above event happens, and let p_N, p_S, p_E, and p_W denote the probabilities with which the soldier buries the treasure at x_0N, x_0S, x_0E, and x_0W, given that $X = x_0$. One of the directions is the "backtrack" direction, say N (recall x_0 is fixed), so the professor's long run gain will be

$$m(x_0)\{\tfrac{1}{4}[p_N-.7]+\tfrac{3}{4}[\tfrac{1}{3}(p_S-.2)+\tfrac{1}{3}(p_E-.2)+\tfrac{1}{3}(p_W-.2)]\} = m(x_0)(-.075).$$

Thus he will lose money if $m(x_0) > 0$. Of course, all the above probability calculations are based on assuming the existence of a true (countably additive) probability distribution π describing the soldier's paths θ (but not on knowledge of this distribution).

It is of interest to see how a Bayesian would approach the problem. He would think about the generation of θ , perhaps deciding that the soldier has a probability p_n of having a path θ of length n, and assigning equal probability to all paths of length n, there being $N_n = 4 \cdot 3^{n-1}$ such paths. Thus the prior probability of a particular path of length n would be $\pi(\theta) = p_n/N_n$. Now if $X = x_0$ is observed, x_0 being of length m, the posterior probability that θ is x_0 "backtracked" can be calculated to be

$$9 \cdot p_{m-1}/(3\ p_{m+1} + 9\ p_{m-1}),$$

while the posterior probability that θ is any particular "extension" of x_0 is

$$p_{m+1}/(3\ p_{m+1} + 9\ p_{m-1}).$$

If p_{m+1} is thought to be approximately equal to p_{m-1}, then the posterior distribution is essentially π_1. The Bayesian, of course, feels this is reasonable for <u>either</u> of the two models when x_0 is observed (since the likelihood functions are the same), and will play the soldier's game but will always backtrack. (This Bayesian analysis is essentially that in Dickey's discussion of Stone (1976).)

6. CONCLUDING REMARKS

1. What is being criticized about Pivotal and Structural analysis is, at most, a very small part of the two theories. The major part of both theories is the reduction of the original data and model to simpler entities which preserve all available information. The reductions are especially valuable in the very common situations where the structural model is known with considerable confidence, but the distribution of the error component, ω, is uncertain. All theories of inference, including Bayesian, can take advantage of the simplifications that result from such "necessary" reductions.

   Pivotal and Structural analysis come into possible conflict with the LP only at the terminal stage of analysis. We are actually somewhat unclear as to when terminal Pivotal analysis conflicts with the LP; we have seen statements to the effect that it can, but not explicit examples. (This explains the emphasis on examples from the structural theory in this paper.)

   Structural theory can conflict with the LP at the terminal stage of analysis in two ways: first, when significance testing or frequentist confidence intervals are developed, and second, when structural distributions are created. We have concentrated on the problems arising with conflicting structural distributions, but again must emphasize that terminal analysis with structural distributions is only a very small part of Structural analysis. (Indeed, in Fraser (1979), very little emphasis is placed on structural distributions.) Of course, the more classical frequentist type of terminal analysis can also conflict with the LP, but probably not too seriously. Indeed the degree to which Pivotal and Structural analysis violate the LP is, on the whole very small, and not really worth making an issue of, except that the theories purport to establish the lack of validity of the LP.

Incidentally, the "disproof" of the LP (by Structural analysis at least) seems to simply be the fact that the recommended terminal analysis, especially that based on structural distributions, can conflict with the LP. To us, however, the justification for terminal analysis with structural distributions is not on very firm footing (compared with the reduction analyses of structural theory), since it involves, at some point, a fiducial type inversion. Besides the examples in this paper, which point out questionable properties of such terminal analysis, there are the examples with non-amenable groups. Indeed, if in the situation of Example 1 the model is $X^t = A\omega^t$, where A is known only to be a nonsingular matrix and of interest is $\Sigma = AA^t$, then the structural analysis (c.f. Fraser (1973)) gives as a structural distribution the posterior distribution with respect to the (right and left) Haar measure on the full linear group, which is non-amenable. Use of this distribution is well known to result in extended inadmissibility and incoherence in a wide number of situations (such as in the estimation problems considered in Example 1 (continued)). Likewise, in the Stone example (where Θ is again non-amenable), the "natural" structural distribution for θ seems to be the one giving probability $\frac{1}{4}$ to each of the paths compatible with x, and is fairly clearly bad, and while Fraser gives a justification for the alternate structural distribution π_1, he does not clearly expose the error in the derivation of the "natural" structural distribution.

Of course, all "objective" statistical theories have trouble dealing with problems involving non-amenable groups (c.f. Bondar (1981)), and Structural analysis often fares better than most: it frequently bases the analysis on amenable subgroups of a non-amenable group, such as in the earlier version of Example 1. (Structural analysis also justifies use of the right invariant Haar measure as a noninformative prior, a justification also obtained from classical invariance theory, but lacking in many purely Bayesian theories.)

2. While this paper cconcentrated on Pivotal and Structural theories because of their clearly voiced opposition to the full LP, there are, of course many other theories of inference which also violate the LP and are susceptible to the same inadmissibility and incoherency criticisms. Among these are fiducial theory (c.f. Wilkinson (1977)), Plausibility inference (c.f. Barndorff-Nielsen (1976)), and many noninformative prior Bayesian theories in which the noninformative prior depends on E, so that one may, in a situation like Example 3, end up using two different noninformative priors to process proportional likelihood functions. Examples of inadmissibility and incoherency are easy to construct for any such situation.

   Of course, one "escape" available to any theory conflicting with the LP is to reject the Weak Conditionality Principle, for then, when faced with the mixed experiment E^{λ}, one could change the component experiment analyses. Rejecting what to many is the only "obvious" principle in statistics is not a very appealing escape, however.

3. Those who remain unconvinced by the inadmissibility and incoherency arguments, and stand fast in their objection to the LP because it ignores

the structural model, must still take notice of certain related principles, such as the Stopping Rule principle of Barnard (1949) and the Censoring principal of Pratt (in Birnbaum (1962)). (See also Pratt (1965), Basu (1975), and Berger and Wolpert (1984) for very general versions.) These principals have a crucial impact on statistics and can be derived using only the Weak Conditionality Principle and a structural version of sufficiency which preserves the structural information.

4. The LP is non-operational, in the sense that it does not say how the likelihood function is to be used. Cogent arguments can be given (c.f. Basu (1975) and Berger and Wolpert (1984)) that only Bayesian utilization of the likelihood function really makes sense. However, Bayesian analysis has to be concerned with sensitivity to the prior input, and, for a variety of practical and theoretical reasons, a sensible analysis from this viewpoint may formally violate the LP (usually, by having some aspects of the prior depend on E or the data). A theoretical justification for possible violation of the LP could be given in terms of Good's Type II Rationality (c.f. Good (1976)), but the practical necessities are fairly clear. In the same way, practical considerations may lead to a mild degree of inadmissibility or incoherency. (For extensive discussion of these issues and references see Berger and Wolpert (1984) and Berger (1983).) The LP (and admissibility and coherence) should be considered ideals, to which one should strive to adhere, rather than absolute prescriptions.

ACKNOWLEDGMENTS

I would like to thank Leon Gleser, Ker-Chau Li, and Herman Rubin for valuable discussions on these matters. This research was supported by the National Science Foundation under Grant MCS-8101670

REFERENCES

AKAIKE, H. (1982). On the fallacy of the likelihood principle. Statistics and Prob. Letters, 1, 75-78.

BARNARD, G. A. (1949). Statistical inference (with Discussion). J. Roy. Statist. Soc. B, 11, 115-139.

BARNARD, G. A. (1980). Pivotal inference and the Bayesian controversy (with Discussion). In Bayesian Statistics (J.M. Bernardo, M.H. DeGroot, D. V. Lindley, and A.F.M. Smith, eds.). Valencia: University Press.

BARNARD, G. (1982). A coherent view of statistical inference. To be published in the proceedings of the Symposium on Statistical Inference and Applications. University ofWaterloo, August 1981.

BARNARD, G. A. and GODAMBE, V. P. (1982). Allan Birnbaum (memorial article). Ann. Statist., 10, 1033-1039.

BARNARD, G. A., JENKINS, G. M., and WINSTEN, C. B. (1962). Likelihood inference and time series. J. Roy. Statist. Soc. A, 125, 321-372.

BARNARD, G. A. and SPROTT, D. A. (1983). The generalised problem of the Nile: robust confidence sets for parametric functions. Ann. Statist., 11, 104-113.

BARNDORFF-NIELSEN, O. (1976). Plausibility inference (with Discussion). J. Roy. Statist. Soc. B, 38, 103-131.

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

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

BERGER, J. (1983). The robust Bayesian viewpoint. In Robustness in Bayesian Statistics (J. Kadane, ed.). Amsterdam: North-Holland Publishing Co.

BERGER, J. and WOLPERT, R. (1984). The Likelihood Principle: A Review and Generalizations. To appear in the Monograph Series of the Institute of Mathematical Statistics.

BIRNBAUM, A. (1962). On the foundations of statistical inference (with Discussion). J. Amer. Statist. Assoc., 57, 269-306.

BIRNBAUM, A. (1969). Concepts of statistical evidence. In Philosophy, Science, and Method: Essays in honor of Ernest Nagel (S. Morgenbesser, P. Suppes, and M. White, eds.). New York: St. Martin's Press.

BIRNBAUM, A. (1970). On Durbin's modified principle of conditionality. J. Amer. Statist. Assoc., 65, 402-403.

BIRNBAUM, A. (1972). More on concepts of statistical evidence. J. Amer. Statist. Assoc., 67, 858-861.

BONDAR, J. V. (1977). On a conditional confidence principle. Ann. Statist., 5, 881-891.

BONDAR, J. V. and MILNES, P. (1981). Amenability: a survey for statistical applications of Hunt-Stein and related conditions on groups. Z. Wahrscheinlich. verw. Gebiete, 57, 103-128.

BROWN, L. D. (1967). The conditional level of Student's t-Test. Ann. Math. Statist., 38, 1068-1071.

BROWN, L. D. (1973). Estimation with incompletely specified loss functions. J. Amer. Statist. Assoc., 70, 417-427.

BROWN, L. D. (1976). Notes on Statistical Decision Theory. (Unpublished lecture notes, Ithaca.)

BROWN, L. D., COHEN, A., and STRAWDERMAN, W. E. (1980). Complete classes for sequential tests of hypothesis. Ann. Statist., 8, 377-398.

BUEHLER, R. J. (1959). Some validity criteria for statistical inference. Ann. Math. Statist., 30, 845-863.

BUEHLER, R. J. (1971). Measuring information and uncertainty. In Foundations of Statistical Inference (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston Publishing Co.

BUEHLER, R. J. (1976). Coherent preferences. Ann. Statist., 4, 1051-1064.

BUEHLER, R. J. and FEDDERSON, A. P. (1963). Note on a conditional property of Student's t. Ann. Math. Statist., 34, 1098-1100.

BUNKE, H. (1975). Statistical inference: Fiducial and structural vs. likelihood. Math Operationsforsch. U. Statist., 6, 667-676.

CORNFIELD, J. (1969). The Bayesian outlook and its application (with Discussion). Biometrics, 25, 617-657.

COX, D. R. (1958). Some problems connected with statistical inference. Ann. Math. Statist., 29, 356-372.

DAWID, A. P. (1977). Conformity of inference patterns. In Recent Developments in Statistics (J. R. Barra, et. al., eds.). Amsterdam: North-Holland Publishing Co.

DAWID, A. P. and STONE, M. (1982). The functional-model basis of fiducial inference. Ann. Statist., 10, 1054-1073.

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. J. Roy. Statist. Soc. B, 35, 189-233.

DEFINETTI, B. (1937). Foresight: its logical laws, its subjective sources. In Studies in Subjective Probability (H. E. Kyburg and H. E. Smokler, eds., 1962). New York: Wiley.

DEFINETTI, B. (1962). Does it make sense to speak of 'Good probability appraisers'? In The Scientist Speculates (I. J. Good, ed.). New York: Basic Books.

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

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

DURBIN, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality, and likelihood. J. Amer. Statist. Assoc., 65, 395-398.

EATON, M. L. (1970). Some problems in covariance estimation. Tech. Report No. 49, Stanford University.

EATON, M. L. (1982). A method for evaluating improper prior distributions. In Statistical Decision Theory and Related Topics III (S. S. Gupta and J. Berger, eds.). New York: Academic Press.

FARRELL, R. H. (1964). Estimators of a location parameter in the absolutely continuous case. Ann. Math. Statist., 35, 949-998.

FERGUSON, T. S. (1967). Mathematical Statistics: A Decision-Theoretic Approach. New York: Academic Press.

FISHER, R. A. (1956). Statistical Methods and Scientific Inference. Edinburgh: Oliver and Boyd.

FRASER, D. A. S. (1963). On the sufficiency and likelihood principles. J. Amer. Statist. Assoc., 58, 641-647.

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

FRASER, D.A.S. (1972). Bayes, likelihood, or structural. Ann. Math. Statist., 43, 777-790.

FRASER, D.A.S. (1973). Inference and redundant parameters. In Multivariate Analysis-III (P. R. Krishnaiah, ed.). New York: Academic Press.

FRASER, D.A.S. (1977). Confidence, posterior probability, and the Buehler example. Ann. Statist., 5, 892-898.

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

FREEDMAN, D.A. and PURVES, R. A. (1969). Bayes methods for bookies. Ann. Math. Statist., 40, 1177-1186.

GATSONIS, C.A. (1981). Estimation of the posterior density of a location parameter. Ph.D. Thesis, Cornell University, Ithaca.

GODAMBE, V. P. (1979). On Birnbaum's mathematically equivalent experiments. J. Roy. Statist. Soc. B, 41, 107-110.

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

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

HEATH, D. and SUDDERTH, W. (1978). On finitely additive priors, coherence, and extended admissibility. Ann. Statist., 6, 333-345.

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.). Amsterdam: North-Holland Publishing Co.

HILL, B. (1981). On some statistical paradoxes and non-conglomerability. In Bayesian Statistics (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.). Valencia: University Press.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In Fourth Berkeley Symposium Math. Statist. and Prob. Berkeley: University of California Press.

KALBFLEISCH, J. D. (1974). Sufficiency and conditionality. In Proceedings of the Conference on Foundational Questions in Statistical Inference (O. Barndorff-Nielsen, P. Blaesild, and G. Schou, eds.). Department of Theoretical Statistics, University of Aarhus.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with Discussion). J. Amer. Statist. Assoc., 72, 789-827.

LANE, D. A. and SUDDERTH, W. D. (1983). Coherent and continuous inference. Ann. Statist., 11, 114-120.

LECAM, L. (1955). An extension of Wald's theory of statistical decision functions. Ann. Math. Statist., 26, 69-81.

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

LINDLEY, D. V. (1982). Scoring rules and the inevitability of probability. Int. Statist. Rev., 50, 1-26.

MAULDON, J. G. (1955). Pivotal quantities for Wishart's and related distributions and a paradox in fiducial theory. J. Roy. Statist. Soc. B, 17, 79-85.

NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. Roy. Soc. A, 231, 289-337.

PEISAKOFF, M. (1950). Transformation Parameters. Ph.D. Thesis, Princeton University.

PIERCE, D. A. (1973). On some difficulties with a frequency theory of inference. Ann. Statist., 1, 241-250.

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

RAMSEY, F. P. (1926). Truth and Probability in the Foundations of Mathematics (1960). Patterson, New Jersey: Littlefield, Adams, and Co.

ROBINSON, G. K. (1979a). Conditional properties of statistical procedures. Ann. Statist., 7, 742-755.

ROBINSON, G. K. (1979b). Conditional properties of statistical procedures for location and scale parameters. Ann. Statist., 7, 756-771.

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

SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. J. Amer. Statist. Assoc., 66, 783-801.

SELLIAH, J. (1964). Estimation and testing problems in a Wishart distribution. Ph.D. thesis, Dept. of Statistics, Stanford University.

STEIN, C. (1961). Estimation of many parameters. Inst. Math. Statist. Wald Lectures, Unpublished.

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

STONE, M. (1970). Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. Ann. Math. Statist., 41, 1349-1353.

STONE, M. (1976). Strong inconsistency from uniform priors (with Discussion). J. Amer. Statist. Assoc., 71, 114-125.

TAKEMURA, A. (1982). An orthogonally minimax estimator of the covariance matrix of a multivariate normal population. Technical Report, Stanford University.

WALD, A. (1950). Statistical Decision Functions. New York: Wiley.

WALLACE, D. L. (1959). Conditional confidence level properties. Ann. Math. Statist., 30, 864-876.

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