CHAPTER 5. IMPLEMENTATION OF THE LIKELIHOOD PRINCIPLE

5.1 INTRODUCTION

The LP strikes us as correct, and behaving in violation of it would be a source of considerable discomfort. Yet the LP does not tell one what to do (although insisting on methods based on the observed likelihood function certainly reduces the possibilities). It can indeed be argued that there is sometimes *no* sensible method of behavior which is completely consistent with the LP.

This raises a very important distinction which is often misunderstood in foundational matters. "Foundations" usually proceeds by formulating properties of desirable behavior, and then seeing what can be deduced from these properties. The quintessential example is that from (very reasonable) axioms of "consistent" or "rational" behavior, it can be deduced that any "consistent" analysis corresponds to some Bayesian analysis. This does *not* imply, however, than any particular form of consistent (Bayesian) analysis is necessarily satisfactory, since, as C.A.B. Smith said in Savage, et. al. (1962),

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

And there is no guarantee that a nonobnoxious consistent way of behaving exists. (See Berger (1984e) for further discussion.) Thus foundational arguments (including the LP) can logically be considered irrelevant from an operational perspective.

This is certainly overstating the case, somewhat, in that, at the very least, foundational arguments can be invaluable in giving direction to

our efforts. Thus the "consistency" theory strongly suggests that truth lies in a Bayesian direction, and the LP strongly suggests that truth lies in the direction of methods based on determination and utilization of the likelihood function (for the observed x). Luckily (or inevitably) these two directions are compatible.

To show that the LP is not irrelevant, we must argue that a sensible method of analysis exists which is compatible with it. This is simply too much to ask; it would involve demonstrating that such a methodology works well "across-the-board" in statistics. Instead, we will content ourselves to arguing for what, we feel, this methodology must be, namely robust Bayesian analysis. We start out, however, with a very brief description of non-Bayesian likelihood methods. Until Section 5.4, we will assume that the likelihood function $\ell_x(\theta)$ (for the observed x) exists.

5.2 NON-BAYESIAN LIKELIHOOD METHODS

It should first be mentioned that there are classically based likelihood methods such as maximum likelihood estimation and likelihood ratio testing. Although these are usually given evidential interpretations in frequentist terms, the concepts themselves are clearly of great importance in likelihood methods. The literature on these subjects is too vast to even attempt mentioning.

Since the LP states that all evidence about θ is contained in $\ell_x(\theta)$, one conceivable solution to the problem of what to do is simply to report $\ell_x(\theta)$, leaving its use and interpretation "to the user" (c.f., Fisher (1956a) and Box and Tiao (1973)). This is not necessarily unreasonable, as "eyeballing" a likelihood function often reveals most things of interest, at least when Θ is low dimensional. Many people probably could learn to usefully deal with likelihood functions as the basic elements of statistics (and indeed many now do). Even a Bayesian should encourage reporting of likelihood functions. Thus Good (1976) says

> "If a Bayesian is a subjectivist he will know that the initial probability density varies from person to person and so he will see the value of graphing the likelihood function for communication. A Doogian will consider that even his own initial probability density is not unique so he should approve even more".

Nevertheless, reporting of $\ell_x(\theta)$ can not be considered to be the end of the statistician's job; properly using $\ell_x(\theta)$ can be difficult and crucial. Also, the natural visual interpretation that will be ascribed to $\ell_x(\theta)$ by most users is that of a probability distribution for θ, an interpretation needing careful handling.

Most of the likelihood methods that have been proposed are dependent on the interpretation that $\ell_x(\theta_1)/\ell_x(\theta_2)$ measures the relative support of the data for θ_1 and θ_2. Extensive development of this idea can be found in Hacking (1965) and Edwards (1972). Other likelihood developments can be found in Fisher (1956a), Barnard, Jenkins and Winsten (1962), Birnbaum (1962a), Barnard (1967a), Sprott and Kalbfleisch (1969), Kalbfleisch and Sprott (1970), Andersen (1970, 1971, 1973), Kalbfleisch (1971, 1978), Barndorff-Nielsen (1971), Sprott (1973a, 1973b), Cox and Hinkley (1974), Cox (1975), Tjur (1978), Hinkley (1978, 1979, 1980, 1982), Grambsch (1980), Barnett (1982), and many of the references given in Chapter 2. "Plausibility Inference" (c.f. Barndorff-Nielsen (1976)) is also related. (Not all of these authors necessarily subscribe to the LP, of course.)

We do not detail these developments for several reasons. First, the space requirement would simply be prohibitive. Second, many of the techniques proposed, while valuable, are either designed only for a narrow class of problems, and hence do not provide a basis for a general likelihood based theory, or attempt generality but fall prey to counterexamples. (See Birnbaum

(1962a), the discussion in Kalbfleisch and Sprott (1970), Plante (1971), Basu (1975), Hill (1973, 1975), and Levi (1980) for some such counterexamples.) Finally, and most importantly, we will argue in the next section that there are compelling reasons for utilizing $\ell_x(\theta)$ through Bayesian analysis, and hence that non-Bayesian likelihood techniques are inherently limited. Such techniques can offer substantial improvements over classical methods, however, and should be useful for those unwilling to accept a Bayesian approach. Also, many of the technical developments in these articles can be useful even to a Bayesian.

5.3 ARGUMENTS FOR BAYESIAN IMPLEMENTATION

Savage, in the discussion of Birnbaum (1962a), said

> "...I suspect that once the likelihood principle is widely recognized, people will not long stop at that halfway house but will go forward and accept the implications of personalistic probability for statistics."

It would be inappropriate here to present the full range of arguments for Bayesian analysis. Instead, we will concentrate on indicating how sensible use of the likelihood function seems possible only through Bayesian analysis.

5.3.1 General Considerations

First, believers in the LP should, it seems,be especially wary of what Good (1976) called 'adhockeries'. These are superficially reasonable methods of analysis which, however, have no firm foundational basis. Careful investigation of adhockeries always seems to reveal a flaw. Non-Bayesian use of likelihood functions virtually always proceeds by developing an adhoc method of dealing with involved situations. No adhoc method ever seems to be sufficient. Indeed, the rationality or consistency justification for Bayesian analysis gives a strong indication that no adhoc method will ever prove foolproof.

An example of the problems faced by non-Bayesians is that, discussed in Section 3.5, of dealing with informative nuisance parameters. Such nuisance parameters are part of the likelihood function, yet need to essentially be eliminated before progress can be made. The Bayesian approach provides a natural (though maybe difficult) way of doing this; determine a prior distribution and integrate out the nuisance parameter (after multiplying the likelihood function and the prior). Simple alternatives, such as maximizing over the nuisance parameter, are simply too crude to give general hope of success (see Lindley in the discussion of Birnbaum (1962a)), although fairly sophisticated methods (such as those in Hinde and Aitken (1984)) may often work reasonably well.

The only situations in which pure likelihood methods are completely convincing are simple ones (such as testing two simple hypotheses), where they in fact correspond to Bayes procedures. Thus Birnbaum (1962a) says (and supports with examples)

> "And, at least for such simple problems, one might say that (LP) implies (Bayes) in the very broad and qualitative sense that *use* of statistical evidence as characterized by the likelihood function alone entails that inference - or decision-making behavior - will be externally indistinguishable from (some case of) a Bayesian mode of inference."

The above arguments will not be very compelling to most non-Bayesians, so let us turn to the key issue - that $\ell_x(\theta)$ need make little sense unless interpreted through a Bayesian filter. If π is a prior (density for convenience) on Θ, then a Bayesian believing π is reasonable or plausible would view $\ell_x(\theta)$ through the posterior distribution

$$\pi(\theta|x) = \ell_x(\theta)\pi(\theta)/\int\pi(\theta)\ell_x(\theta)d\theta,$$

which essentially corresponds to viewing $\ell_x(\theta)$ as a probability density w.r.t. the (properly normalized version of) π. The prior π need not be proper, and indeed those wanting "objectivity" might desire to use a "noninformative" prior π as the basis of the normalizing measure. In any case, the key to the Bayesian approach is to treat $\ell_x(\theta)$ as an actual probability density - and it is reasonable to do so only when it is considered a density w.r.t. the presumed prior measure for θ.

A number of justifications for this view have been advanced. First is the quite persuasive argument that probability is the language of uncertainty, so the uncertainty about θ, reflected in $\ell_x(\theta)$, should be expressed probabilistically. Second, it usually is necessary to compare or relate one *subset* of Θ to another, and some method of averaging over $\ell_x(\theta)$ is then needed. Indeed, Basu (1975) presents reasonable arguments that $\ell_x(\theta)$ should be "additive" when Θ is discrete. (His argument, however, that in reality Θ is always discrete, is much less convincing than the corresponding argument that $\mathcal{X}$ is discrete; we measure X to only a certain accuracy, but θ could still be anything.)

Non-Bayesian averaging of $\ell_x(\theta)$ has the severe problem that reparameterization can change the answer. One can make a change of variables $\eta = \psi(\theta)$, where ψ is a 1-1 function, and the resulting likelihood function for η, namely $\ell_x(\psi^{-1}(\eta))$, could look completely different. Adhoc averaging methods will virtually always give different conclusions for the reparameterized likelihood function (as will many other intuitive likelihood techniques), a very disturbing prospect. Of course, the interpretation of $\ell_x(\theta)$ as a probability density w.r.t. the prior measure is immune to this problem, since a reparameterization simply introduces a Jacobian in the transformed prior.

In some situations, it is clearly imperative to determine and introduce π. One such situation is that of Section 4.5.2, in which the likelihood function itself conveys almost no information unless θ is severely restricted through π (i.e., a suitable model for the population is introduced). Indeed the nonparametric situation discussed in Section 3.6.1 is the general

prototype for this situation, in that the likelihood function is very difficult to use unless θ is substantially restricted a priori, corresponding to proposing a model (or class of models) for the distribution of X. The "generalized inverse" problems discussed in Jaynes (1981) also have this same flavor.

The failure of the likelihood function to provide clearly interpretable information, when Θ is huge, is sometimes deemed a criticism of the LP. Instead, we view it as an indication that prior information must be used in such situations. (See also the discussion in Section 4.5.)

5.3.2 The Fraser-Monette-Ng, Stone, and Stein Examples

Next, we turn to three important examples which have been viewed as counterexamples to the LP, but instead are viewed by us as indications that a Bayesian (rather than intuitive) interpretation of the likelihood function is needed. The first is an example from Fraser, Monette, and Ng (1984). (See also Evans, Fraser, and Monette (1986) and the discussion therein for additional development.)

EXAMPLE 34. Suppose $\mathcal{X} = \Theta = \{1,2,\ldots\}$, and

$$(5.3.1) \qquad f_\theta(x) = \frac{1}{3} \text{ for } x = \begin{cases} \theta/2, 2\theta, 2\theta+1 & \text{when } \theta \text{ is even} \\ (\theta-1)/2, 2\theta, 2\theta+1 & \text{when } \theta \neq 1 \text{ is odd} \\ 1,2,3 & \text{when } \theta = 1. \end{cases}$$

The likelihood function is easily seen to be

$$\ell_x(\theta) = \frac{1}{3} \text{ for } \theta = \begin{cases} x/2, 2x, 2x+1 & \text{when } x \text{ is even} \\ (x-1)/2, 2x, 2x+1 & \text{when } x \neq 1 \text{ is odd} \\ 1,2,3 & \text{when } x = 1. \end{cases}$$

Thus, for any x, the data intuitively gives equal support to the three possible θ compatible with that observation. On solely likelihood based grounds, therefore, any of the three θ would be a suitable estimate. Consider, therefore, three possible estimators, δ_1, δ_2, and δ_3, corresponding to using the first, middle, and last possible θ, respectively: thus

$$\delta_1(x) = \begin{cases} x/2 & \text{when } x \text{ is even} \\ (x-1)/2 & \text{when } x \neq 1 \text{ is odd} \\ 1 & \text{when } x = 1, \end{cases}$$

$$\delta_2(x) = 2x, \text{ and } \delta_3(x) = 2x+1.$$

Now

$$P_\theta(\delta_2(X) = \theta) = P_\theta(X = \theta/2) = \begin{cases} 1/3 & \text{when } \theta \text{ is even} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$P_\theta(\delta_3(X) = \theta) = P_\theta(X = (\theta-1)/2) = \begin{cases} 1/3 & \text{when } \theta \neq 1 \text{ is odd} \\ 0 & \text{otherwise,} \end{cases}$$

while, amazingly,

$$P_\theta(\delta_1(X) = \theta) = \begin{cases} P_\theta(\{1,2,3\}) = 1 & \text{when } \theta = 1 \\ P_\theta(\{2\theta,2\theta+1\}) = 2/3 & \text{otherwise.} \end{cases} \tag{5.3.2}$$

Even more surprising is that the confidence set $C_1(x) = \{2x,2x+1\}$ seems twice as good from a "pure likelihood" viewpoint as $C_2(x) = \{\delta_1(x)\}$, and yet

$$P_\theta(C_1(X) \text{ contains } \theta) = \begin{cases} 0 & \text{when } \theta = 1 \\ 1/3 & \text{otherwise} \end{cases}$$

while

$$P_\theta(C_2(X) \text{ contains } \theta) = \begin{cases} 1 & \text{when } \theta = 1 \\ 2/3 & \text{otherwise.} \end{cases}$$

Of course, the measures here are frequentist measures, but the decision-theoretic or coherency evaluation arguments of Section 3.7 can be applied to indicate substantial inferiority of δ_2, δ_3, or C_1 in repeated use.

Let us now consider what happens when $\ell_x(\theta)$ is passed through a Bayesian filter. A Bayesian has a prior density π for θ, and his posterior density will be

$$\pi(\theta|x) = \frac{\ell_x(\theta)\pi(\theta)}{m(x)} = \frac{\pi(\theta)I_{\{\delta_1(x),\delta_2(x),\delta_3(x)\}}(\theta)}{\pi(\delta_1(x))+\pi(\delta_2(x))+\pi(\delta_3(x))}.$$

Thus, indeed, the data conveys nothing to the Bayesian except that θ is $\delta_1(x)$, $\delta_2(x)$, or $\delta_3(x)$. Is the Bayesian indifferent between $\delta_1(x)$, $\delta_2(x)$, and $\delta_3(x)$, however? He is only if $\pi(\delta_1(x)) = \pi(\delta_2(x)) = \pi(\delta_3(x))$, which cannot hold for all x (when π is a proper density). Indeed it will typically be the case, at least for densities which are monotonically decreasing for large θ, that $\pi(\delta_1(x)) > \pi(\delta_2(x)) + \pi(\delta_3(x))$. Thus a Bayesian would never always use δ_2, δ_3, or C_1, and would, in fact, tend to use δ_1. The Bayesian thus avoids the danger inherent in pure likelihood reasoning.

As a final comment on this example, note that a (sophisticated) noninformative prior Bayesian obtains a reasonable (objective) answer to this problem. Although one might naively give θ a constant (improper) prior density, resulting in the ill-advised $\pi(\theta|x) = \ell_x(\theta)$, it is clear from (5.3.1) that θ is approximately a scale parameter. This would lead a noninformative prior Bayesian to use the Jeffrey's (1961) prior density for a scale parameter, namely $\pi(\theta) = \theta^{-1}$. With this noninformative prior, not only is $\delta_1(x)$ again the clear choice for θ, but the posterior probability that $\theta = \delta_1(x)$ is approximately 2/3 for large x. (This, incidentally, provides a conditional justification for the frequentist report in (5.3.2).)

Thus, either a proper prior Bayesian or a careful noninformative prior Bayesian will easily arrive at a sensible likelihood-based conclusion in this example. We have seen no pure likelihood methods which can make the same claim.

EXAMPLE 35. Stone (1976) (see also Hill (1981) for a discussion similar to the following) considers a very interesting example in which a drunken soldier, starting at an intersection O in a city (which has square blocks), staggers around on a random path trailing a taut strong. Eventually the soldier stops at an intersection (after walking at least one block) and buries a treasure. Let θ denote the path of the string from O to the treasure. Letting N, S, E,

and W stand for a path segment one block long in the indicated direction, θ can be expressed as a sequence of such letters, say

$$\theta = N\ N\ E\ S\ W\ S\ W\ W.$$

(Note that NS, SN, EW, and WE cannot appear, as the taut string would just be rewound. In expressions below, however, we allow such combinations to appear for notational convenience, although they are to be understood to cancel.)

After burying the treasure, the soldier randomly chooses one of the four possible directions and walks one block in that direction (still keeping the string taut). Let X denote this augmented path, so that X is one of the paths $\{\theta N, \theta S, \theta E, \theta W\}$, with probability $\frac{1}{4}$ each. We observe X, and are to find the treasure.

Note first that, for given $X = x$, the only possible values of θ are $\{xN, xS, xE, xW\}$, and since the probability that $X = x$ when each of these θ obtains is $\frac{1}{4}$, we have the likelihood function

$$\ell_x(\theta) = \frac{1}{4}$$

for each of the four possible θ.

Stone uses this example to indicate a problem with use of the "noninformative" prior $\pi(\theta) = 1$ for all possible paths θ, since an easy calculation then shows that the posterior probability of each of the four possible θ, given x, is $\frac{1}{4}$. This supposedly conflicts with the intuition that, given θ, X is three times as likely to extend the path as to backtrack (there are 3 directions to extend the path and only one to backtrack), so that x "most likely" arose from the *one* θ (among the four possibilities) for which x is an extension. Fraser, in the discussions of Stone (1976) and Hill (1981), indicates that this strikes him as a conclusive counterexample to the LP itself, the "likelihoods" of $\frac{1}{4}$ seeming absurd from a frequentist (conditional on θ) viewpoint.

To us, this example again serves to indicate that the likelihood function can really be utilized only through Bayesian analysis. For instance, forget for a moment the amusing structure of Stone's example, and just consider

the statistical problem involving θ and X. Suppose θ was in actuality generated according to the (prior) distribution

$$\pi(\theta) = \begin{cases} 1/(2\cdot 3^n - 1) & \text{if length of } \theta \leq n \\ 0 & \text{if length of } \theta > n, \end{cases}$$

and that a path x of length $< n$ is observed. Then a Bayesian analysis is certainly correct, and the posterior probability of each of the possible θ given x is indeed $\frac{1}{4}$. Returning to the example of the soldier, this makes clear that if θ is felt to have essentially a uniform prior in the neighborhood of x, then the analysis decried by Stone and Fraser is correct.

The difficulty here is that it was never described "why" the soldier stopped at a given intersection and buried the treasure, i.e., how θ was generated. We, in fact, would doubt that θ was locally uniform at any x. Far more reasonable would be to assume that the soldier stops after a path of length n, with some probability p_n, and that all paths of length n (there are $N_n = 4\cdot 3^{n-1}$ of them when $n \geq 1$) have equal probability of occurring. Then, if θ is a path of length n,

$$\pi(\theta) = p_n/N_n.$$

For a given x of length $m \geq 1$, three of the possible θ are of length m+1 while one is of length m-1. The posterior probabilities of these are

$$\frac{(p_{m+1}/N_{m+1})\frac{1}{4}}{3(p_{m+1}/N_{m+1})\frac{1}{4} + (p_{m-1}/N_{m-1})\frac{1}{4}} = \frac{p_{m+1}}{3p_{m+1} + 9p_{m-1}}$$

(for the θ of length m+1) and

$$\frac{(p_{m-1}/N_{m-1})\frac{1}{4}}{3(p_{m+1}/N_{m+1})\frac{1}{4} + (p_{m-1}/N_{m-1})\frac{1}{4}} = \frac{9p_{m-1}}{3p_{m+1} + 9p_{m-1}}$$

(for the θ of length m-1). If $p_{m+1} \cong p_{m-1}$, then these probabilities are $\frac{1}{12}$ and $\frac{3}{4}$, respectively, indicating that it certainly is sensible to presume that the treasure is buried at that θ for which x is an extension of the path. (This analysis is very similar to that of Dickey in the discussion of Stone (1976) and to the analysis in Hill (1981) using a finitely additive prior on n).

The above considerations can be reinforced by considering a second model proposed by Fraser in the discussion of Hill (1981). Fraser's model is that the observation X is generated from θ according to the following scheme, where x_0 denotes a *given* particular path and 0 the origin:

$$f_\theta(x_0) = \frac{1}{4} \text{ and } f_\theta(0) = \frac{3}{4} \text{ when } \theta \in \{x_0N, x_0E, x_0S, x_0W\},$$

$$f_\theta(0) = 1 \text{ when } \theta = x_0,$$

$$f_\theta(\theta) = \frac{1}{4} \text{ and } f_\theta(0) = \frac{3}{4} \text{ for the remaining } \theta.$$

(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 good samaritan who walks the streets within one block of a shelter at x_0, 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 good samaritan will take him back to x_0.) Suppose now that the observation from this model just happens to be x_0, so that the likelihood function for θ is the same as that obtained from Stone's model for the observation x_0. The LP says that the conclusions in each case should be the same, and we concur. Since θ is still the path generated by the drunken soldier, the prior defined by $\pi(\theta) = p_n/N_n$ is still appropriate, and the resulting Bayesian analysis sensible. (Alternatively, if θ had been generated in such a way that the prior was felt to be locally uniform near x_0 - note that any proper prior could only be locally uniform near some of the possible observations - the Bayesian analysis with $\pi(\theta) = K$ would be appropriate.)

This Bayesian reasoning is in conflict with frequentist reasoning, which states in the situation of Stone that, conditional on θ, X is three times as likely to extend the path as to backtrack, while in the situation of Fraser there is no reason to think this. Such reasoning seems to be the basis of the claim by Fraser (in the discussion of Hill (1981)) that the situation provides a counterexample to the LP. To us it instead provides yet another

counterexample to frequentist reasoning. If there is doubt as to this, imagine that θ really was generated according to one of the priors considered here (and a compelling case can be made that the drunken soldier actually does generate θ according to $\pi(\theta) = p_n/N_n$), in which case there seems little doubt that the two models give the same answer. (See also Berger (1984a).)

This example again shows the possible error in attempting to base an analysis solely on $\ell_x(\theta)$, and shows how the Bayesian perspective resolves the difficulties. One interesting feature of this example is that the natural noninformative prior density is constant, and results in the ill-advised $\pi(\theta|x) = \frac{1}{4}$, for the four possible θ. The difficulty with the noninformative prior approach here is that the parameter space can be viewed as the free group on two generators and, as shown by Peisakoff (1950), this group is too large for group-based statistical analyses to work. (Peisakoff discusses the problem from the viewpoint of invariance theory, but invariance theory has a very close relationship with noniformative prior Bayesian theory - c.f. Berger (1980).) Bondar and Milnes (1981) provide extensive discussion concerning when such groups are "too large."

EXAMPLE 36. Stein (1962) constructed the following example to show the difficulty in casually applying the LP. An unknown quantity $\theta > 0$ can be measured by $X \sim \eta(\theta,\sigma^2)$ (σ^2 known) or by Y having density

$$f(y|\theta) = cy^{-1} \exp\{-\frac{d^2}{2}(1 - \frac{\theta}{y})^2\}I_{(0,b\theta)}(y), \tag{5.3.3}$$

where c is the appropriate normalizing constant, b is enormous (say, $10^{10^{1000}}$), and d is large (say, 50). The likelihood functions $\ell_x(\theta)$ and $\ell_y(\theta)$ for the respective experiments would be (ignoring multiplicative constants and recalling that $\theta > 0$)

$$\ell_x(\theta) = \exp\{-\frac{1}{2\sigma^2}(\theta-x)^2\}I_{(0,\infty)}(\theta),$$

$$\ell_y(\theta) = \exp\{-\frac{d^2}{2y^2}(\theta-y)^2\}I_{(y/b,\infty)}(\theta).$$

Suppose now that the observations are such that $x = y = \sigma d$. Then the only difference between $\ell_x(\theta)$ and $\ell_y(\theta)$ is the difference between the factors $I_{(0,\infty)}(\theta)$ and $I_{(y/b,\infty)}(\theta)$, which can be shown to be negligible because b is so huge. Thus the LP says that the given observations, x and y, provide (essentially) the same information about θ. We agree with this entirely.

Next, Stein observes that the usual, say 95%, frequentist or objective conditional confidence interval for θ when x is observed is

$$(5.3.4) \qquad (x-(1.96)\sigma,\ x+(1.96)\sigma)$$

(note that $x/\sigma = d = 50$, so the restriction to $\theta > 0$ is essentially irrelevant), and hence that application of the LP implies that the interval

$$(5.3.5) \qquad (y-(1.96)[y/d],\ y+(1.96)[y/d])$$

should be used if y is observed. Again we agree, providing the interval in (5.3.4) is inappropriate.

Considering now the interval in (5.3.5) as a frequentist interval (to be used for all y), a calculation shows (see Berger (1980)) that

$$(5.3.6) \qquad P_\theta\left(Y - \frac{(1.96)Y}{d} < \theta < Y + \frac{(1.96)Y}{d}\right) < 10^{-100}.$$

This, to a frequentist, casts extreme doubt on the premise that the interval in (5.3.5) contains θ, and seems to indicate a failure of the LP.

To a Bayesian, there is no real problem with this example. The use of (5.3.5) was predicated on the validity of (5.3.4), which in turn follows only if the prior is approximately locally uniform within several standard deviations σ of the actual observation x (and is well behaved outside this region). In reality, this will never be the case for all x and σ; any *proper* prior will give substantially different results as x and particularly σ vary. Indeed, note that it was assumed that $x = y = \sigma d$ in the above conditional analysis, and since it can be shown that Y is almost certain to be enormous (on the order of b in size), it follows that we must imagine that x and σ are also enormous. The use of (5.3.4), when x and σ are enormous, will rarely be conditionally sound. It is this use of (5.3.4), not the use of the LP, which is in error. And if a very small y just happens to occur, then and only then

is use of (5.3.4), and hence (5.3.5), indicated.

Two observations concerning a Bayesian analysis of this problem are in order. The first is that clever Bayesian reasoning is not required to show the inadequacy of the interval in (5.3.5) for all but very small y. Indeed, for virtually any prior distribution, the interval in (5.3.5) will have posterior probability near zero. The second observation is that a standard noninformative prior Bayesian analysis *does* work well here. For the density in (5.3.3), θ can easily be seen to be a scale parameter, and again the standard noninformative prior density would be $\pi(\theta) = \theta^{-1}$. A Bayesian analysis with this improper prior gives very sensible answers and shows the interval in (5.3.5) to be seriously inadequate for all but very small y.

It is worthwhile to summarize the three main points that are illustrated in the above examples.

1. Intuitive utilization of likelihood functions can be misleading. In Examples 34 and 35, for instance, the usual interpretation of a likelihood function as a measure of the comparative support of the data for the various θ, while formally correct, can lead to an erroneous conclusion if prior information is not considered.

Example 36 also demonstrates that intuitive approaches which work well in a certain situation should not be carelessly transferred to different situations with a similar likelihood function. It is true that, when prior information is vague in the normal mean situation, the "confidence" interval

(5.3.7) $$(x-(1.96)\sigma,\ x+(1.96)\sigma)$$

is a reasonable conditional procedure. Naively transferring this to the Y situation fails, however, because (5.3.7) is reasonable only when σ is small enough for the prior information to indeed be vague, and the Y problem involves observations which will usually correspond to huge σ. This "error" is noted and extensively discussed in Basu (1975).

2. While not directly related to our central thesis, these examples indicate the care needed in the use of improper "noninformative" priors. When prior

opinions are indeed reflected by a locally noninformative prior (in the region of Θ for which the likelihood function is significant), the use of noninformative priors is reasonable as an approximation. (See also Box and Tiao (1973), Dickey (1976), and Berger (1984e).) It appears, however, especially from Example 35, that automatic use of noninformative priors can lead one astray. This is not to say that use of noninformative priors is to be avoided; indeed we feel that they are invaluable in obtaining relatively simple, good, and "objective" statistical procedures.

3. These examples can be turned around and used as indictments of frequency reasoning. Frequency reasoning in each example would correspond (at best) to Bayesian analysis with respect to a certain, very special, prior. Quite different answers were seen to obtain if other prior beliefs were held. This, of course, is another general justification for the Bayesian position: a "good" frequentist procedure is usually a Bayes procedure with respect to some prior, and if the corresponding prior does not seem reasonable, use of the procedure is suspect.

5.4 ROBUST BAYESIAN ANALYSIS

We seem to have been inexolorably led to Bayesian analysis. Our interpretation of the situation at this point is that we can best interpret the information from the data, namely $\ell_x(\theta)$, as a probability density on Θ w.r.t. some prior measure, π, reflecting our prior beliefs (or lack thereof) concerning θ. Thus one need only elicit his prior distribution, π_0, and perform a Bayesian analysis.

Unfortunately, elicitation of π_0 is not easy, and indeed cannot be done with complete accuracy in a finite amount of time. (We are thinking of π_0 as the prior which would be the result of infinitely long reflection on the problem.) It is not clear that writing down a quick guess at π_0 and performing a Bayesian analysis with this guess is better than other non-Bayesian methods of analysis. The fear is that the guess for π_0 might contain features which would be deemed to be in error upon further reflection, and that

these features might have such an overwhelmingly detrimental effect on the analysis that a classical analysis which ignores prior information might be preferable.

The obvious method of alleviating such fears is to do robust Bayesian analysis (see Berger 1984e, 1985, and 1987 for general surveys and references), wherein one considers, instead of a single guess for π_0, a class Γ of plausible prior distributions felt certain to contain π_0. From Γ (and the likelihood function) one obtains a class of possible posterior distributions to work with. (Note that, in this robust Bayesian sense, Ev(E,x) really is a set of "evidences".) If the conclusion or action to be taken is essentially the same for all such posteriors, then the problem is solved. Indeed, in a sense this is the *only* situation in which there can be said to be an unequivocal answer to a problem. (This holds true also when Γ is a class of priors of various individuals who must come to a joint conclusion.)

It may happen, however, that the conclusion or action to be taken is quite different for various posteriors in the class. When this is the case there are four options: (i) Attempt further prior elicitation (resulting in a narrowing of Γ); (ii) Obtain more data; (iii) Conclude that there is no answer; and (iv) Choose among the possible answers according to some criteria not involving further prior elicitation. Solutions (i) and (ii) are certainly to be attempted, if possible, but limited time or resources may preclude such solutions (an example of Good's Type II rationality). Note that solution (i) may be somewhat simpler than it seems at first sight, since the observed data may effectively rule out a large portion of Γ, meaning that further prior elicitation can be concentrated on specific aspects of the problem. Solution (iii) is certainly reasonable, and is in some sense the only truly honest conclusion if (i) or (ii) cannot be pursued. But in many situations it is necessary to proceed anyway and obtain an "intelligent" guess at the answer.

This brings us to solution (iv), i.e., the use of alternate criteria. There are many possibilities here, with the following five being the most important:

1. Put a prior distribution on Γ itself, and carry out a formal Bayesian analysis. (Note that this would be simply a formal prior distribution of some sort, since the prior elicitation process has supposedly ceased.)
2. Use minimax type criteria on posterior measures (e.g., posterior expected losses) for $\pi \in \Gamma$.
3. Use frequentist measures to select a "good" procedure compatible with $\pi \in \Gamma$.
4. Use some measure of "information" to select a prior in Γ, such as a "maximum entropy" prior (cf. Jaynes (1982)) or a "reference" prior (cf. Bernardo (1979)).
5. Use Type II maximum likelihood methods (cf. Good (1965)), essentially choosing the prior $\pi \in \Gamma$ which maximizes the marginal or predictive density $m(x|\pi) = E^{\pi}[f_{\theta}(x)]$ for the given data x (such a prior being the "most plausible" prior in Γ in light of the data). This is a standard adhoc Bayesian and empirical Bayesian technique.

Discussion and other references for these methods can be found in Berger (1984e, 1985) and Berger and Berliner (1986). Of interest here is that two of these methods, namely methods 3 and 4, can violate the LP. (Method 4 can violate the LP because the selected prior will typically depend on all of E, not just the observed likelihood function.) We will not enter into a discussion of the relative merits of the five methods, but do note that there seem to be statistical problems that are most amenable to solution by each of the methods. For instance, there are many high dimensional and nonparametric problems where it is hard to find *any* reasonable prior distribution, much less do a robust Bayesian analysis, and yet relatively simple frequentist procedures exist which can be meaningful to a conditionalist in the sense of Example 16 in Section 4.1.3. Consider the following example, which we learned from Brad Efron.

EXAMPLE 37. The experiment, E, consists of observing $X_1,\ldots,X_{15}$, which are i.i.d. observations from a completely unknown continuous density f on R^1. (Here we identify θ with the unknown f, so Θ is the set of all continuous densities on R^1.) Of interest is ξ, the median of the unknown density. A simple binomial calculation shows that a 96.5% frequentist confidence interval for ξ is given by $[X_{(4)}, X_{(11)}]$, where the $X_{(i)}$ are the order statistics. Due to the extreme difficulty of constructing reasonable prior distributions on Θ, a Bayesian might well choose to simply use $[X_{(4)},X_{(11)}]$, with the interpretation provided by Example 16.

Thus, because of difficulties in performing a robust Bayesian analysis, a conditionalist might formally violate the LP. Of course, this could be viewed as merely a temporary condition due to the lack of development of Bayesian theory; certainly greater effort has been expended by statisticians on development of non-Bayesian theory. Also, the need to compromise should not be viewed as providing legitimacy to the compromises, but should instead be viewed as a forced stab in the dark. Thus Savage, in Savage et. al. (1962), states

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

A second reason for possible violation of the LP, as discussed in Section 4.1.3, is that many users of statistics will be unable to perform careful robust Bayesian analyses. For these users we must provide simple Bayesian procedures with "built in" robustness. In part, this robustness

should be measured in a frequency sense, since the procedures will be used repeatedly (i.e., for different X). Of course, good conditional performance of these procedures should still be of paramount concern. Note, in particular, that the noninformative prior or "objective" Bayesian procedures are usually very good procedures from this perspective of use by nonspecialists, and may formally violate the LP through dependence of the noninformative prior on E (and not just the observed $\ell_x(\theta)$). In a similar vein, Hajék (1967, 1971) argues that asymptotic theory (which can provide useful simple procedures for nonspecialists in complicated situations) can sometimes be more difficult if one is restricted to basing it only on the given likelihood function, and not on E as a whole.

As a final comment concerning Bayesian analysis, it should be mentioned that choice of a prior (or class Γ) will often have to wait until after the data is at hand and $\ell_x(\theta)$ is available. Thus in Barnard, Jenkins, and Winsten (1962) it is stated (where they refer to "weights" instead of a "prior")

> "The advantage of looking first at the likelihood function and then considering the weights, lies in the fact that the likelihood function will often be so near zero over much of the range of θ that the weights in these regions can be quickly dismissed from consideration."

This "choosing the prior after seeing the data" strikes many as unsavory, but it is absolutely essential when Θ is high dimensional or otherwise complicated. It is less disturbing when viewed from the robust Bayesian viewpoint, where a conclusion is deemed clearcut only when any reasonable prior passed over $\ell_x(\theta)$ gives essentially the same answer. See Berger(1984e) for further discussion.

5.5 CONCLUSIONS

At first sight, we seem to have come to the conclusion that the LP is not always applicable, in that the only satisfactory method of analysis based on the LP seems to be robust Bayesian analysis, which because of technical difficulties may sometimes require use of techniques that formally violate the LP. We emphatically believe, however, that the LP is always valid, in the sense that the experimental evidence concerning θ *is* contained in $\ell_x(\theta)$. Because of limited time and resources, however, interpreting or making use of this evidence *may* involve use of measures violating the LP. Of course, whenever such a measure is used one should make sure that it has not led to a recognizably erroneous conditional conclusion.

Until now (in this section) we have assumed the existence of $\ell_x(\theta)$. As mentioned in Sections 3.4 and 3.6.1, this assumption is (in a sense) always valid, since the sample space is always finite in reality and then $\ell_x(\theta)$ always exists, even when the model is uncertain or unknown. Practical considerations often call for the use of continuous approximations, however, for which the likelihood function may be ill-defined or not exist. (Of course, as mentioned in Section 3.6.1, even in many continuous nonparametric situations the likelihood function can be considered to exist.) In any case, the RLP always applies, and a good case can also be made that robust Bayesian analysis is the only reasonable method of analysis consistent with it. More frequent compromises may, however, be needed in these more difficult situations.

Even for those who find themselves unable to accept Bayesian methods, the LP should not be ignored and the conditional viewpoint should be kept in mind. If a classical procedure is being used, a quick check to make sure that it is saying something which is at least sensible conditionally seems only prudent. Statistics looks very bad when it recommends a conclusion that clearly contradicts common sense.