CHAPTER 3. THE LIKELIHOOD PRINCIPLE AND GENERALIZATIONS

3.1 INTRODUCTION

The LP deals with situations in which X has a density $f_\theta(x)$ (with respect to some measure ν) for all $\theta \in \Theta$. Of crucial importance is the *likelihood function for* θ *given* x, given by

$$\ell_x(\theta) = f_\theta(x), \tag{3.1.1}$$

i.e., the density evaluated at the observed value X = x and considered as a function of θ. Often we will call $\ell_x(\theta)$ the *likelihood function for* θ or simply the *likelihood function*. The LP, which follows, is stated in a form suitable for easy initial understanding; certain implicit qualifications are discussed at the end of the section.

THE LIKELIHOOD PRINCIPLE. All the information about θ *obtainable from an experiment is contained in the likelihood function for* θ *given* x. *Two likelihood functions for* θ *(from the same or different experiments) contain the same information about* θ *if they are proportional to one another.*

It has been known since Fisher (1925, 1934) that the "random" likelihood function $\ell_X(\theta)$ is a minimal sufficient statistic for θ, and hence contains all information about θ from a classical viewpoint. The LP goes considerably farther, however, maintaining that only $\ell_x(\theta)$ for the actual observation X = x is relevant.

EXAMPLE 9. Suppose $Y_1, Y_2, \ldots$ are i.i.d. Bernoulli (θ) random variables. In experiment E_1, a fixed sample size of 12 observations is decided upon, and the sufficient statistic $X_1 = \sum_{i=1}^{12} Y_i$ turns out to be $x_1 = 9$. In experiment E_2, it is decided to take observations until a total of 3 zeroes has been observed, at which point the sufficient statistic $X_2 = \sum Y_i$ turns out to 9. The distribution of X_1 in E_1 is binomial with density

$$f_\theta^1(x_1) = \binom{12}{x_1}\theta^{x_1}(1-\theta)^{12-x_1},$$

which for $x_1 = 9$ yields the likelihood function

$$\ell_9^1(\theta) = \binom{12}{9}\theta^9(1-\theta)^3.$$

The distribution of X_2 in E_2 is negative binomial with density

$$f_\theta^2(x_2) = \binom{x_2+2}{x_2}\theta^{x_2}(1-\theta)^3,$$

which for $x_2 = 9$ yields the likelihood function

$$\ell_9^2(\theta) = \binom{11}{9}\theta^9(1-\theta)^3.$$

In this situation, the LP says that (i) for experiment E_i alone, the information about θ is contained solely in $\ell_9^i(\theta)$; and (ii) since $\ell_9^1(\theta)$ and $\ell_9^2(\theta)$ are proportional as functions of θ, the information about θ in experiments E_1 and E_2 is identical.

These conclusions are, of course, at odds with frequentist reasoning. The binomial and negative binomial distributions will tend to give different frequentist measures. For instance, a one-tailed significance test of H_0: $\theta = \frac{1}{2}$ will give significance levels of $\alpha = .0730$ and $\alpha = .0338$ in the

binomial and negative binomial cases, respectively, so, if significance at the $\alpha = .05$ level was sought, one would either reject or not reject depending on the model. (See Lindley and Phillips (1976) for further discussion.)

This example also evidences a consequence of the LP that will be discussed later, namely that the "stopping rule" is irrelevant when drawing inferences about θ. Here, it does not matter whether the stopping rule was to sample until the twelfth observation or until 3 zeroes were obtained; the data that 9 ones and 3 zeroes were obtained is all that should be relevant.

It is interesting that even certain Bayesians would, at least formally, also espouse violation of the LP in this example. For instance, the noninformative (generalized) priors for θ that are recommended by Jeffreys (1961) are $\pi_1(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$, in the binomial case, and $\pi_2(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-1}$, in the negative binomial case. These will lead to different posterior distributions and hence (typically) different inferences, even when the likelihood functions are proportional. (See Hill (1974a) for further discussion.)

EXAMPLE 10. Let $\mathcal{X} = \{1,2,3\}$ and $\Theta = \{0,1\}$, and consider experiments E_1 and E_2 which consist of observing X_1 and X_2 with the above $\mathcal{X}$ and the same θ, but with probability densities as follows:

	x_1		
	1	2	3
$f_0^1(x_1)$	.90	.05	.05
$f_1^1(x_1)$	.09	.055	.855

,

	x_2		
	1	2	3
$f_0^2(x_2)$	.26	.73	.01
$f_1^2(x_2)$	.026	.803	.171

.

If, now, $x_1 = 1$ is observed, the LP states that the information about θ should depend on the experiment only through $(f_0^1(1), f_1^1(1)) = (.9, .09)$. Furthermore, since this is proportional to $(.26, .026) = (f_0^2(1), f_1^2(1))$, it should be true that $x_2 = 1$ provides the same information about θ as does $x_1 = 1$. Another way of stating the LP for testing simple hypotheses, as here, is that the experimental information about θ is contained in the likelihood ratio for

the observed x. Note that the likelihood ratios for the two experiments are also the same when 2 is observed, and also when 3 is observed. Hence, no matter which experiment is performed, the *same* conclusion about θ should be reached for the given observation. This example clearly indicates the startling nature of the LP. Experiments E_1 and E_2 are very different from a frequentist perspective. For instance, the test which accepts $\theta = 0$ when the observation is 1 and decides $\theta = 1$ otherwise is a most powerful test with error probabilities (of Type I and Type II, respectively) .10 and .09 for E_1, and .74 and .026 for E_2. Thus the classical frequentist would report drastically different information from the two experiments. (And the conditional frequentist is also likely to report E_1 and E_2 differently; indeed, for E_2 it is hard to perform any sensible conditional frequentist analysis because of the three point $\mathcal{X}$ and the widely differing error probabilities.)

This example emphasizes a very important issue. It is clear that experiment E_1 is more likely to provide useful information about θ, as reflected by the overall better error probabilities. The LP in no sense contradicts this. The LP applies only to the information *about* θ that is available from knowledge of the experiment and the *observed* x. Even though E_1 has a much better chance of yielding good information, the LP states that the conclusion, once x is at hand, should be the same, regardless of whether x came from E_1 or E_2. The conflict of the LP with frequentist justifications seems inescapable. (See also Birnbaum (1977).)

Hill (1987a,b) discusses a number of important clarifications or qualifications of the LP. Several of these are discussed in depth later in the monograph, but it is perhaps pedagocically best to at least mention them here.

The first has to do with the role of θ. As presented up until now, θ represents only the unknown aspect of the probability distribution of X. For the bulk of the monograph we will confine attention to this case, it being the

most familiar statistical situation. Often, however, there are unknowns which are relevant to a statistical problem but which do not directly affect the distribution of X. One example is prediction, in which it is desired to predict an unknown random variable Z, after observing X. Other examples arise in design and sequential analysis problems, where as-yet-unobserved data can affect the decision to be made. Examples are given in Section 3.5.

In general, therefore, the LP should be formulated in such a way that θ consists of *all* unknown variables and parameters that are relevant to the statistical problem. (Any attempt to precisely define "relevant to the statistical problem" would involve both decision theory and model formulation, and lead us too far astray.) The major difficulty with working in such generality is that of defining what is then meant by a likelihood function for θ (cf. Bayarri, DeGroot, and Kadane (1987)). We have opted for discussing this general situation only in Section 3.5, though we believe that virtually all issues raised for the special case of θ being the model parameter also apply to appropriate formulations of the general situation. In any case, it is important to keep in mind the qualification that θ must contain all unknowns relevant to the problem for the LP to be valid in its simple form.

A second qualification for the LP is that it only applies *for a fully specified model* $\{f_\theta\}$. If there is uncertainty in the model, and if one desires to gain information about which model is correct, that uncertainty must be incorporated into the definition of θ.

A third qualification is that, in applying the LP to two different experiments, it is imperative that θ be the same unknown quantity in each. Thus, in Example 9, we assumed that θ represented the same success probability in either the binomial or negative binomial experiment. In applying the LP to two different experiments, we also require that the choice of an experiment be *noninformative* (e.g. implemented by a chance mechanism not involving θ);

this might be violated if the experimenter chooses among possible experiments on the basis of prior beliefs. Informative experimental choices may be handled by the methods discussed in Section 4.2.7.

Further elaboration and other qualifications will be introduced as we proceed. Understanding the limitations and the domain of applicability of the LP is almost as important as understanding its basis and implications.

3.2 HISTORY OF THE LIKELIHOOD PRINCIPLE

For a history of the concept of likelihood, see Edwards (1974). The name "likelihood" first appeared in Fisher (1921). Fisher made considerable use of likelihood and conditioning concepts (cf. Fisher (1925, 1934, 1956a)) and came close to espousing the LP in Fisher (1956a), but refrained from complete committment to the principle. Versions of the LP were developed and promoted by Barnard in a series of works (Barnard (1947a, 1947b, 1949)). Likelihood concepts were also employed by a number of other statisticians, cf. Bartlett (1936, 1953).

The LP received major notice in 1962, due to Barnard, Jenkins, and Winsten (1962) and Birnbaum (1962a). Both papers (and the Discussions of them) contained numerous compelling examples in favor of the LP, and also provided axiomatic developments of the LP from the simpler (and more believable) concepts of sufficiency and conditionality. Birnbaum's development is more convincing, and will be given in the next section. The work since then on the LP and its consequences is considerable, as can be seen from the references. Noteworthy general discussions can be found in Pratt (1965), Cox and Hinkley (1974), Dawid (1981), Barnett (1982), and especially Basu (1975).

In fairness, it should be mentioned that Barnard came to support only a limited version of the LP and Birnbaum ultimately came close to

rejecting it. The reasons will be discussed in Sections 3.6.4 and 4.1, respectively.

The above development is a brief history of the LP from a non-Bayesian perspective. The LP was always implicit in the Bayesian approach to statistics. This is because, if $\pi(\theta)$ is a prior density for θ, then the posterior density is

$$\pi(\theta|x) = \pi(\theta)\ell_x(\theta)/m(x)$$

(assuming $m(x) = E^{\pi}\ell_x(\theta) > 0$), which depends on the experiment only through $\ell_x(\theta)$ (presuming that selection of π is independent of E and x). Since all Bayesian inference follows from the posterior, the LP is an immediate consequence of the Bayesian paradigm. Thus Jeffreys (1961) says

> "Consequently the whole of the information contained in the observation that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give to the likelihood."

An important point here is that $\ell_x(\theta)$ is all that matters to a Bayesian, no matter what prior density π is used. It is tempting, therefore, to say that, if $\ell_x(\theta)$ contains all the sample information about θ regardless of the known prior, then $\ell_x(\theta)$ should contain all the sample information even when the prior is unknown.

The above relationship between the LP and Bayesian analysis should probably be qualified to some extent, in that it is possible to be a "frequentist Bayesian." One can believe that only frequentist measures of procedure performance have validity, and yet, because of various rationality 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. The posterior distribution would provide a convenient mathematical device for determining the best procedure, from this viewpoint, but overall

frequentist Bayes measures of performance, not posterior Bayes measures, would be the relevant measures of accuracy. The LP directly attacks this view, arguing that thinking "conditional Bayes," not "frequentist Bayes," is important.

As somewhat of an aside here, there are two other reasons why Bayesians should be very interested in the LP. The first is that, in complicated real problems, Bayesians will often spend much of their time simply looking at likelihood functions and doing maximum likelihood analyses, due to calculational complexities of a full Bayesian analysis. Emphasizing the importance of the observed likelihood function is thus to be encouraged. Finally, there is the very pragmatic reason that promoting the Bayesian position can often be most effectively done by first selling the LP, since the latter can be done without introducing the emotionally charged issue of prior distributions (see Berger (1984b)).

3.3 BIRNBAUM'S DEVELOPMENT - THE DISCRETE CASE

Birnbaum's (1962a) development of the LP from the intuitively simpler and more plausible concepts of sufficiency and conditionality is formally correct only in the case of experiments with discrete densities (see Section 3.4.1). Since the discrete case is also the easiest to understand intuitively, we restrict ourselves in this section to a discrete sample space $\mathcal{X}$. We carefully outline Birnbaum's argument, to allow easy dissection by those who find it hard to believe the conclusion. The mathematical style is kept fairly informal; rigor poses no problem because of the discreteness.

3.3.1 Evidence, Conditionality, and Sufficiency

By an experiment E, we herein mean the triple $(X, \theta, \{f_\theta\})$, where the random variable X, taking values in $\mathcal{X}$ and having density $f_\theta(x)$ for some θ in Θ, is observed. (Because of the discreteness, the density can be assumed to exist, and we will take all subsets of $\mathcal{X}$ to be measurable.) For simplicity of notation, $\mathcal{X}$ and Θ will be suppressed in the description of E. Virtually all statistical methodologies require only the above information concerning an

experiment. (The "structural theory" of Fraser and the "pivotal theory" of Barnard deem additional information relating X, θ, and the randomness to be important, however. This issue will be discussed in Sections 3.6.4 and 3.7.)

The outcome of the experiment is the data $X = x$, and from E and x we are to infer or conclude something about θ (or about something related to θ). Following Birnbaum (1962a), we will call this inference, conclusion, or report the *evidence about* θ *arising from* E *and* x, and will denote this by Ev(E,x). We presuppose nothing about what this evidence is; it could (at this stage) be any standard measure of evidence, or something entirely new. (Since E is an argument, it could certainly be a frequentist measure.) Also, we do not preclude the possibility that Ev(E,x) depends on "other information," such as prior information about θ, or a loss function in a decision problem. The focus will be on the manner in which the "report" Ev(E,x) should depend on E and x. (Dawid (1977) prefers to talk about *methods* of inference based on E and x, and principles which these methods should satisfy. In a sense, by letting Ev(E,x) denote whatever conclusion one is going to report, we are also taking this view, while keeping Birnbaum's notation.) As one final point, Ev(E,x) could be a collection of "evidences" about θ, obviating the criticism that the LP is based on the assumption that a single measure of evidence exists.

The Conditionality Principle essentially says that, if an experiment is selected by some random mechanism independent of θ, then only the experiment actually performed is relevant. (The selection mechanism is ancillary, so this is a version of conditioning on an ancillary statistic.) The general conditionality principle is not needed here. Indeed we need only the following considerably weaker principle, named by Basu (1975).

WEAK CONDITIONALITY PRINCIPLE (WCP). *Suppose there are two experiments* $E_1 = (X_1, \theta, \{f_\theta^1\})$ *and* $E_2 = (X_2, \theta, \{f_\theta^2\})$, *where only the unknown parameter* θ *need be common to the two experiments. Consider the* <u>*mixed*</u> *experiment* E^*, *whereby* $J = 1$ *or* 2 *is observed, each having probability* $\frac{1}{2}$ *(independent of* θ, X_1, *or* X_2*), and experiment* E_J *is then performed. Formally,* $E^* = (X^*, \theta, \{f_\theta^*\})$,

where $X^* = (J, X_J)$ *and* $f^*_\theta((j, x_j)) = \frac{1}{2} f^j_\theta(x_j)$. *Then,*

$$Ev(E^*, (j, x_j)) = Ev(E_j, x_j),$$

i.e., *the evidence about* θ *from* E^* *is just the evidence from the experiment actually performed.*

The WCP is nothing but a formalization of Example 2, and hence is essentially due to Cox (1958). It is hard to disbelieve the WCP, yet, as mentioned after Example 2, even the WCP alone has serious consequences.

Turning finally to the familiar concept of sufficiency, we state the following weak version (named by Dawid (1977)).

WEAK SUFFICIENCY PRINCIPLE (WSP). Consider an experiment $E = (X, \theta, \{f_\theta\})$, *and suppose* $T(X)$ *is a sufficient statistic for* θ. *Then, if* $T(x_1) = T(x_2)$, $Ev(E, x_1) = Ev(E, x_2)$.

The LP will be seen to follow directly from the WCP and WSP. A variety of alternate principles also lead to the LP (cf. Basu (1975), Dawid (1977), Barndoff-Nielsen (1978), Berger (1984a), Bhave (1984), and Evans, Fraser, and Monette (1985c, 1986)). The WCP and WSP are the most familiar, however. Another prominent principle is "Mathematical Equivalence," given in Birnbaum (1972). This principle is a weak version of the sufficiency principle, stating that if, in a given experiment E, $f_\theta(x_1) = f_\theta(x_2)$ for all θ, then $Ev(E, x_1) = Ev(E, x_2)$. One could base the LP on mathematical equivalence, plus a minor generalization of the WCP. The weakening of sufficiency is carried to the ultimate in Evans, Fraser, and Monette (1986), which derives the LP solely from a generalized version of the conditionality principle.

3.3.2 Axiomatic Development

The formal statement of the LP is as follows.

FORMAL LIKELIHOOD PRINCIPLE. Consider two experiments $E_1 = (X_1, \theta, \{f^1_\theta\})$ *and* $E_2 = (X_2, \theta, \{f^2_\theta\})$, *where* θ *is the same quantity in each experiment. Suppose that for the particular realizations* x^*_1 *and* x^*_2 *from* E_1 *and* E_2, *respectively,*

$$\ell_{x_1^*}(\theta) = c\ell_{x_2^*}(\theta)$$

for some constant c (i.e., $f_\theta^1(x_1^*) = cf_\theta^2(x_2^*)$ *for all* θ). *Then*

$$Ev(E_1,x_1^*) = Ev(E_2,x_2^*).$$

LIKELIHOOD PRINCIPLE COROLLARY. If $E = (X, \theta, \{f_\theta\})$ *is an experiment, then* $Ev(E,x)$ *should depend on* E *and* x *only through* $\ell_x(\theta)$.

THEOREM 1 (Birnbaum (1962a)). *The Formal Likelihood Principle follows from the* WCP *and the* SP. *The converse is also true.*

Proof. If E_1 and E_2 are the two experiments about θ, consider the mixed experiment E* as defined in the WCP. From the WCP we know that

$$Ev(E^*,(j,x_j)) = Ev(E_j,x_j). \tag{3.3.1}$$

Next, thinking solely of E* with random outcome (J,X_J), consider the statistic

$$T(J,X_J) = \begin{cases} (1,x_1^*) & \text{if } J = 2,\ X_2 = x_2^* \\ (J,X_J) & \text{otherwise.} \end{cases}$$

(Thus the two outcomes $(1,x_1^*)$ and $(2,x_2^*)$ result in the same value of T.) T is a sufficient statistic for θ. This is clear, since

$$P_\theta(X^* = (j,x_j)|T = t \neq (1,x_1^*)) = \begin{cases} 1 & \text{if } (j,x_j) = t \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\begin{aligned} P_\theta(X^* = (1,x_1^*)|T = (1,x_1^*)) &= 1\text{-}P_\theta(X^* = (2,x_2^*)|T = (1,x_1^*)) \\ &= \frac{\frac{1}{2}\cdot f_\theta^1(x_1^*)}{\frac{1}{2}\cdot f_\theta^1(x_1^*) + \frac{1}{2}f_\theta^2(x_2^*)} \\ &= c/(1+c), \end{aligned}$$

all of which are independent of θ. The WSP thus implies that

$$Ev(E^*,(1,x_1^*)) = Ev(E^*,(2,x_2^*)). \tag{3.3.2}$$

Combining (3.3.1) and (3.3.2) establishes the result.

To prove that the LP implies the WCP, observe that, for E*,

$$\ell_{(j,x_j)}(\theta) = \tfrac{1}{2} f_\theta^j(x_j).$$

This is clearly proportional to $f_\theta^j(x_j)$, the likelihood function in E_j when x_j is observed, so the LP implies that

$$Ev(E^*,(j,x_j)) = Ev(E_j,x_j).$$

To prove that the LP implies the WSP, it suffices to note that, if $T(x_1) = T(x_2)$ in an experiment for which T is sufficient, then x_1 and x_2 have proportional likelihood functions. ||

Proof of the LP Corollary. For given $x^* \in \mathcal{X}$, define

$$Y = \begin{cases} 1 & \text{if } X = x^* \\ 0 & \text{if } X \neq x^*, \end{cases}$$

and note that Y has distribution given by

$$f_\theta^Y(1) = f_\theta(x^*) = 1 - f_\theta^Y(0). \tag{3.3.3}$$

For the experiment E* of observing Y, it follows from the LP that

$$Ev(E,x^*) = Ev(E^*,1).$$

But E*, and hence Ev(E*,1), depend only on $f_\theta(x^*) = \ell_{x^*}(\theta)$ (using (3.3.3)). ||

The above results are worth dwelling upon for a moment. The LP is extremely radical from the viewpoint of classical statistics, as will be seen in Chapter 4. Yet to reject the LP, one must *logically* reject either the WCP or the WSP. But the WSP is, itself, a cornerstone of classical statistics, and there is nothing in statistics as "obvious" as the WCP (or Example 2).

3.4 GENERALIZATIONS BEYOND THE DISCRETE CASE

Basu (1975) and others have argued that the sample space $\mathcal{X}$ in any physically realizable experiment must be finite, due to our inability to measure with infinite precision. This suggests that the Likelihood Principle for discrete experiments (as in Section 3.3) is all that one needs. We are

philosophically in agreement with this.

On the other hand, continuous and other more general probability distributions are enormously useful in simplifying statistical computations and in providing numerical approximations which are often quite accurate. It is possible for the likelihood function for a continuous model to differ strikingly from that of the discrete model it is intended to approximate, so it is not obvious that the validity of the LP in discrete problems extends to its validity in the approximating continuous problems. In any case, extension of the LP to more general situations can only strengthen its case. Such an extension is our task in the present section.

As in Section 3.3, an experiment $E = (X, \theta, \{P_\theta\})$ will be understood to involve the observation of the random variable X, having probability distribution P_θ on $\mathcal{X}$, $\theta \in \Theta$. (It will not be necessary to assume the existence of a density.) There is, unavoidably, measure-theoretic mathematics in this section, but the section can be skipped, if desired, without any essential loss of continuity.

The sample space $\mathcal{X}$ will be assumed to be a locally-compact Hausdorff space whose topology admits a countable base (LCCB space, for short), and the P_θ will be assumed to be Borel measures. Of course, X often arises as an $\mathcal{X}$-valued random variable on a probability space $(\Omega, \mathcal{F}, \{\mu_\theta\})$ equipped with a family of probability measures indexed by $\theta \in \Theta$. Such underlying structure will not be relevant in our analysis, however.

3.4.1 Difficulties in the Nondiscrete Case

In an experiment $E = (X, \theta, \{P_\theta\})$ for which there is an $x \in \mathcal{X}$ satisfying $P_\theta(\{x\}) = 0$ for every $\theta \in \Theta$, it is difficult to assign any particular meaning to "Ev(E,x)". For example, Basu (1975) and Joshi (1976) have observed that a naive application of Birnbaum's (1962a) sufficiency principle would suggest for such an x that Ev(E,x) = Ev(E,y) for *every* $y \in \mathcal{X}$, since the map $T: \mathcal{X} \to \mathcal{X}$ which takes x onto y and leaves all other points (including y) fixed is sufficient for θ. This is particularly disturbing for continous

distributions, since then $P_\theta(\{x\})=0$ for every $x \in \mathcal{X}$ and every $\theta \in \Theta$; Birnbaum's sufficiency principle then suggests that all possible observations lend precisely the same evidence (and therefore none) about θ.

The unique specification of a likelihood function causes similar problems. If there is no single σ-finite measure ν on $\mathcal{X}$ whose null sets coincide with those Borel sets N for which $P_\theta(N) = 0$ for all $\theta \in \Theta$, then no likelihood function exists. This is the usual state of affairs in nonparametric problems (recall that Θ could be an arbitrary index set) and can even arise in simple parametric examples; for example, $P_\theta(A) = \frac{1}{2}\int_A dx + \frac{1}{2} I_A(\theta)$, $\Theta = \mathcal{X} = [0,1]$, describes an experiment in which $X = \theta$ with probability $\frac{1}{2}$ and is otherwise uniformly distributed over the unit interval; no σ-finite measure ν dominates $\{P_\theta\}$, and no likelihood function exists. (Incidentally, this seems to be a source of confusion in certain "counterexamples" to the LP such as the second example in Section 2.5 of Birnbaum (1969).)

Even in problems where there is a measure ν with the indicated properties, the Radon-Nikodym derivatives

$$\ell_x(\theta) = f_\theta(x) = P_\theta(dx)/\nu(dx)$$

are determined only up to sets of ν-measure zero; these functions of θ could be specified in an entirely arbitrary manner for all x in any set $N \subset \mathcal{X}$ with $\nu(N) = 0$. One way to salvage a likelihood principle in the face of such ambiguity is to specify a particular version of $P_\theta(dx)/\nu(dx)$ for each θ; for example, in case a (ν-almost everywhere) continuous density exists we could set Ω_x = {open neighborhoods of $x \in \mathcal{X}$} and put

$$\ell_x(\theta) = \inf_{V \in \Omega_x} \sup_{\substack{U \in \Omega_x \\ U \subset V}} (P_\theta(U)/\nu(U))$$

for x in the support of ν, $\ell_x(\theta) = 0$ otherwise.

By restricting our attention to (ν-almost everywhere) continuous densities, continuous sufficient statistics, etc. we could develop versions of the conditionality, sufficiency, and likelihood principles very similar to those in the discrete setting.

Instead we will develop versions of these principles applicable for all experiments, including those with discontinuous density functions and even those for which no likelihood function exists. The price we pay for such generality is that our conclusions will all be weakened by the qualification "for all $x \in \mathcal{X}$ outside a fixed set N with $P_\theta(N) = 0$ for all θ", which we shall abbreviate "for $\{P_\theta\}$ a.e. x". It is important to note that N will be unknown to the statistician, and hence the only assurance that the actual observation x is not in N is the faith that events of probability zero do not happen. This is, of course, a statement in the classical frequentist framework, but establishing a version of the LP within this framework should, at least, be convincing to frequentists.

3.4.2. Evidence, Conditionality, and Sufficiency

As before, denote by Ev(E,x) the (undefined) evidential content of an observation x in an experiment $E = (X, \theta, \{P_\theta\})$. The following are the appropriate generalizations of the WCP and sufficiency principle for non-discrete experiments.

WEAK CONDITIONALITY PRINCIPLE. Consider the mixture, E^*, *of two experiments* $E_1 = (X_1, \theta, \{P_\theta^1\})$ *and* $E_2 = (X_2, \theta, \{P_\theta^2\})$, *defined as* $E^* = (X^*, \theta, \{P_\theta^*\})$, *where* $X^* = (J, X_J)$, $J = 1$ *or* 2 *(as* E_J *is performed) with probability* $\frac{1}{2}$ *each (independent of* θ*), and*

$$P_\theta^*(A) = \frac{1}{2} P_\theta^1(\{x_1: \ (1,x_1) \in A\}) + \frac{1}{2} P_\theta^2(\{x_2: \ (2,x_2) \in A\}).$$

Then,

$$Ev(E^*,(j,x_j)) = Ev(E_j,x_j) \text{ for } \{P_\theta^*\} \text{ - a.e. } (j,x_j).$$

If the sample spaces in E_1 and E_2 are countable, we could delete "impossible" outcomes (i.e., x_i for which $P_\theta^i(x_i) = 0$ for all $\theta \in \Theta$) and dispense with the "$\{P_\theta^*\}$ - a.e." qualification above, thus recovering the discrete WCP.

A formal definition of sufficiency is as follows. Let $E = (X, \theta, \{P_\theta\})$ be an experiment and $T: \mathcal{X} \to \mathcal{T}$ a measurable map from $\mathcal{X}$ to

another LCCB space $\mathcal{T}$. The statistic T determines a family $\{P_\theta^T\}$ of Borel measures on $\mathcal{T}$ by

$$P_\theta^T(A) = P_\theta(T^{-1}(A)),$$

and hence an experiment $E^T = (T, \mathcal{T}, \{P_\theta^T\})$. Unless T is 1-1 we expect (in general) that E^T will tell us less about θ than E, since different outcomes $x \in \mathcal{X}$ with possibly different evidential import can be mapped onto the same $T(x) \in \mathcal{T}$. The exceptional case is that in which T is sufficient.

DEFINITION. *For the experiment* E^T, *suppose there exists a family* $\{g_t: t \in \mathcal{T}\}$ *of Borel probability measures on* $\mathcal{X}$ *satisfying*

$$P_\theta(A) = \int_{\mathcal{T}} g_t(A) P_\theta^T(dt) = \int_{\mathcal{X}} g_{T(x)}(A) P_\theta(dx)$$

for all Borel sets $A \subset \mathcal{X}$. *Then* T *is called "sufficient" (or sometimes "sufficient for* θ*").*

Note that g_t is not permitted to depend upon θ; otherwise $g_t = P_\theta$ would always work. Any one-to-one measurable mapping T is sufficient; just let g_t be a point mass at $T^{-1}(t) \in \mathcal{X}$.

The Sufficiency Principle makes precise the notion that T(x) in $\mathcal{T}$ tells as much about θ as x in E;

SUFFICIENCY PRINCIPLE (SP). If $T: \mathcal{X} \to \mathcal{T}$ *is sufficient, then*

$$Ev(E,x) = Ev(E^T, T(x)) \quad \text{for } \{P_\theta\} \text{ - a.e. } x \in \mathcal{X}.$$

Again we may delete the impossible outcomes when $\mathcal{X}$ is countable to remove the "$\{P_\theta\}$ - a.e." qualification and conclude that Ev(E,x) = Ev(E,y) whenever a sufficient statistic T satisfies T(x) = T(y), and so recover the discrete WSP of Section 3.3.1.

3.4.3. The Relative Likelihood Principle

Let $E_1 = (X_1, \theta, \{P_\theta^1\})$ and $E_2 = (X_2, \theta, \{P_\theta^2\})$ be two experiments and suppose (for motivational purposes) that each admits a likelihood function,

i.e. a σ-finite measure ν_i on the sample space $\mathcal{X}_i$ and a family $\{f_\theta^i(\cdot)\}$ of integrable functions satisfying

$$P_\theta^i(A) = \int_A f_\theta^i(x)\nu_i(dx), \quad A \subset \mathcal{X}_i.$$

The Likelihood Principle (were it to hold here) would assert that

$$Ev(E_1,x_1) = Ev(E_2,x_2)$$

whenever $f_\theta^1(x_1) = cf_\theta^2(x_2)$ for all $\theta \in \Theta$ and some constant $c = c(x_1,x_2)$ not depending on θ, i.e. whenever the *relative likelihood* $c = f_\theta^1(x_1)/f_\theta^2(x_2)$ does not depend on θ. Our freedom to specify $f_\theta^i(x_i)$ arbitrarily whenever $\nu_i(\{x_i\}) = 0$ makes it clear that this principle needs reformulation before it is suitable for experiments with uncountable sample spaces. (However, at points x_1 and x_2 which are atoms of ν_1 and ν_2, respectively, the LP is reasonable, and can be shown to follow from the WCP and SP as in Section 3.3.)

To develop a suitable general principle, we generalize the concept that the relative likelihood of x_1 and x_2 is independent of θ. Basically, if a mapping exists between two subsets of $\mathcal{X}_1$ and $\mathcal{X}_2$ for which the Radon-Nikodym derivative of the induced measure with respect to the existing measure (on, say, $\mathcal{X}_1$) is independent of θ, then we can establish an equivalence of evidence between the corresponding observations in the subsets. The reasons for generalizing the LP in this direction are: (i) It can be stated in great generality, without requiring models or densities; (ii) It will be shown to follow from the WCP and SP, as did the LP; and (iii) It, in turn, can be shown to imply (in substantial generality) the Stopping Rule Principle and Censoring Principle, besides having directly important implications of its own. The major limitation of the RLP (compared to the LP) is that it does not provide any such convenient summarization of evidence as the likelihood function (which need not exist in the general case).

RELATIVE LIKELIHOOD PRINCIPLE (RLP). Let φ: $U_1 \to U_2$ *be a Borel bimeasurable one-to-one mapping from* $U_1 \subset \mathcal{X}_1$ *onto* $U_2 \subset \mathcal{X}_2$, *and suppose there exists a strictly positive function* c *on* U_1 *such that for all* $\theta \in \Theta$,

$$(3.4.1) \qquad P_\theta^2(A) = \int_{\varphi^{-1}(A)} [1/c(x_1)]P_\theta^1(dx_1), \quad A \subset U_2.$$

Then $Ev(E_1,x_1) = Ev(E_2, \varphi(x_1))$ *for* $\{P_\theta^1\}$ - *a.e.* $x_1 \in U_1$.

Note that the RLP does *not* say anything for particular x_1. Indeed, if x_1 has zero probability for all θ, then φ could be defined arbitrarily at x_1 and still satisfy (3.4.1). Thus the RLP can only be interpreted in a pre-experimental sense: if φ satisfies (3.4.1), evidentiary equivalence holds with probability one on U_1. Where φ or U_1 come from is irrelevant. The following theorem shows that the RLP is indeed a generalization of the LP.

THEOREM 2. *For two experiments* $E_1 = (X_1, \theta, \{P_\theta^1\})$ *and* $E_2 = (X_2, \theta, \{P_\theta^2\})$ *with countable sample spaces devoid of outcomes impossible under all* θ, *the* LP *and the* RLP *are equivalent.*

Proof. Without loss of generality, we take the dominating measures ν_1 and ν_2 to be counting measure on $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, so the likelihood functions are $f_\theta^i(x_i) = P_\theta^i(\{x_i\})$. First, assume the validity of the LP, and let

$$P_\theta^2(A) = \int_{\varphi^{-1}(A)} [1/c(x)]P_\theta^1(dx)$$

for some φ: $U_1 \to U_2$ and all $A \subset U_2$. Fix any $x_1 \in U_1$ and set $x_2 = \varphi(x_1)$, $A = \{x_2\}$. Then $f_\theta^2(x_2) = [1/c(x_1)]f_\theta^1(x_1)$ for all θ, so the LP asserts that $Ev(E_1,x_1) = Ev(E_2, \varphi(x_1))$.

Conversely, assume the RLP holds, and suppose that $f_\theta^1(x_1) = cf_\theta^2(x_2)$ for some $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, $c > 0$, and all $\theta \in \Theta$. Put $U_1 = \{x_1\}$, $U_2 = \{x_2\}$, and define φ: $U_1 \to U_2$ by $\varphi(x_1) = x_2$. (Note that we are free to choose U_1, U_2, and φ in any fashion compatible with the conditions in the RLP, but evidentiary equivalence need not hold on any null set.) Regard c as the constant value of a strictly positive function on U_1. Then the RLP asserts that

$$Ev(E_1,x_1) = Ev(E_2, \varphi(x_1)) \quad \text{for } \{P_\theta^1\} \text{ - a.e. } x_1 \in U_1,$$

i.e. that $Ev(E_1,x_1) = Ev(E_2,x_2)$ (by hypothesis $\mathcal{X}_1$ contains no point at which

$f_\theta^1(x_1)$ vanishes for all θ, so the "$\{P_\theta^1\}$ - a.e." qualification is unnecessary). ||

THEOREM 3. *The* WCP *and the* SP *together imply the* RLP.

Proof. Let E_1 and E_2 be two experiments, φ a bimeasurable mapping from a Borel set $U_1 \subset \mathcal{X}_1$ onto $U_2 \subset \mathcal{X}_2$, and c: $U_1 \to (0,\infty)$ a measurable function satisfying

$$P_\theta^2(A) = \int_{\varphi^{-1}(A)} [1/c(x)]P_\theta^1(dx)$$

for all Borel $A \subset U_2$, all $\theta \in \Theta$. Let E^* be the mixture of E_1 and E_2, and define a mapping T: $\mathcal{X}^* \to \mathcal{X}^*$ by

$$T(i,x_i) = \begin{cases} (2,\varphi(x_1)) & \text{if } i = 1 \text{ and } x_1 \in U_1 \\ (i,x_i) & \text{else.} \end{cases}$$

This determines a new experiment $E^{*T} = (T, \mathcal{X}^*, \{P_\theta^T\})$, where $P_\theta^T(A) = P_\theta^*(T^{-1}(A))$.

First we show that T is sufficient. For each $t = (i,x_i) \in \mathcal{X}^*$ define a measure g_t on $\mathcal{X}^*$ by

$$g_t(A) = \begin{cases} \varepsilon_{x_i}(A_i) = \varepsilon_t(A) & \text{if } i = 1 \text{ or } x_i \notin U_2 \\ (c\varepsilon_{x_1}(A_1) + \varepsilon_{x_2}(A_2))/(1+c) & \text{if } i = 2,\ x_2 \in U_2, \text{ and } x_1 = \varphi^{-1}(x_2). \end{cases}$$

Here $c = c(x)$ and $\varepsilon_{x_1}, \varepsilon_{x_2}, \varepsilon_t$ denote the unit point masses at $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, $t \in \mathcal{X}^*$ respectively; A_i denotes $\{x_i \in \mathcal{X}_i\colon\ (i,x_i) \in A\}$. It is straightforward to verify that

$$P_\theta^*(A) = \int g_t(A) P_\theta^T(dt)$$

for each Borel $A \subset \mathcal{X}^*$, so T is sufficient.

By the SP we can conclude that

$$Ev(E^*,(1,x_1)) = Ev(E^{*T},(2,\varphi(x_1))) \quad \text{and}$$

$$Ev(E^*,(2,x_2)) = Ev(E^{*T},(2,x_2))$$

for $\{P_\theta^1\}$ - a.e. $x_1 \in \mathcal{X}_1$ and $\{P_\theta^2\}$ - a.e. $x_2 \in \mathcal{X}_2$. In particular, for

$(\{P_\theta^1\}$ - a.e.) $x_1 \in U_1$ and $X_2 = \varphi(x_1)$ we have

$$Ev(E^*,(1,x_1)) = Ev(E^{*T}, (2,x_2)) = Ev(E^*,(2,x_2)).$$

By the WCP we have

$$Ev(E^*,(1,x_1)) = Ev(E_1,x_1) \text{ and } Ev(E^*,(2,x_2)) = Ev(E_2,x_2),$$

so we can conclude that

$$Ev(E_1,x_1) = Ev(E_2, \varphi(x_1))$$

for $\{P_\theta^1\}$ - a.e. $x_1 \in U_1$. ||

The RLP will be used in Chapter 4 to establish general versions of important consequences of the LP. Theorem 3 demonstrates that rejection of these consequences (and several are quite unpalatable from the frequentist viewpoint) implies rejection of the WCP or the SP.

3.5 PREDICTION, DESIGN, NUISANCE PARAMETERS, AND THE LP

3.5.1 Introduction

The LP as stated above has the very important qualification that it does not apply if θ does not include all unknown quantities germane to the experiment or problem. For instance, in design or prediction problems the unknown future observation is obviously relevant, and yet is not necessarily a part of θ - the parameter defining the distribution of the observable X. A related difficulty is that, often, only a part of θ is really of interest, the remainder being a "nuisance" parameter. These issues are explored in this section.

We begin by expanding the definition of θ to include unobserved and nuisance variables. Define

$$\theta = (z;\omega) = (y,w;\xi,\eta),$$

where $z = (y,w)$ is the value of an unobserved variable Z, with y being of interest and w being a nuisance variable, and where $\omega = (\xi,\eta)$ is the parameter

that determines the distributions of both X and Z, with ξ being of interest and η being a nuisance parameter. (We will purposefully remain vague on the definition of "nuisance variable" and "nuisance parameter"; formal definitions could be attempted along decision-theoretic lines, but would take us too far afield.) To indicate that evidence about ξ and y is desired from E we will write

$$Ev_{y,\xi}\ (E,x)$$

for the evidence about ξ and y from the observation of x in an experiment E.

Two difficulties arise in attempting to apply the LP in this more general context. The first is that this generalized θ is no longer just the parameter defining the distribution of X. Thus the definition in (3.1.1) of $\ell_X(\theta)$ as the density of X given θ may no longer be a suitable definition. Indeed, if Z is conditionally independent of X given ω, then (by the definition of conditional independence) it can be shown that (3.1.1) becomes

$$\ell_X\ (\theta) \equiv f_{Z,\omega}\ (x) = f_{\omega}\ (x),$$

which does not even involve z. The second difficulty is that the nuisance parameter, η, *will* appear in this likelihood function even though it is *not* of interest.

To resolve these difficulties and indicate the role of the LP, we will discuss alternative definitions of the likelihood function which bring out the role of important unobserved variables and suppress the role of nuisance parameters, and we will indicate under what circumstances these forms of the likelihood function may be substituted for the simple (3.1.1).

3.5.2 Unobserved Variables: Prediction and Design

The following example shows that a naive application of the LP can be misleading if future observations are of interest.

EXAMPLE 11. We have available a sequence of observations $X_i = (U_i, V_i)$ $(i = 1,2,\ldots)$ where

$$P(V_{i+1} = 1|V_i = 1) = 1/2, \qquad P(V_{i+1} = 0|V_i = 1) = 1/2$$

$$P(V_{i+1} = 1|V_i = 0) = 0, \qquad P(V_{i+1} = 0|V_i = 0) = 1.$$

(Define $V_0 = 1$). When $V_i = 1$, U_{i+1} will be independent of the previous U_i with a $\eta(\xi,1)$ distribution. When $V_i = 0$, on the other hand, U_{i+1} will be zero. (This would correspond to a situation in which a measuring instrument is used to obtain the important observation U_i, while V_i tells whether the equipment will work the next time ($V_i = 1$) or has irreparably broken ($V_i = 0$)).

Imagine that $x_1,\ldots,x_n$ have been observed, and that $v_i = 1$ for $i = 1,\ldots,n-1$. The likelihood function for ξ is then given by

$$\ell_x(\xi) = \prod_{i=1}^{n} f_\xi(u_i) \propto \text{a } \eta(\bar{u}_n, n^{-1}) \text{ density.}$$

The LP thus says that the evidence about ξ is contained in $\ell_x(\xi)$, and if we are stopping the experiment nothing else is needed. However, in deciding whether or not to take another observation, it is obvious that knowledge of v_n is crucial. If $v_n = 1$ it may be desirable to take another observation, but if $v_n = 0$ it would be a waste of time (since the measuring instrument is broken). This example is related to a limitation of sufficiency (cf. Bahadur (1954)).

The apparent failure of the LP in Example 11 is really the failure to include all unknowns in the specification of θ; only ξ is included. For this problem the next observation, X_{n+1} (and perhaps further observations), are also important unknowns. And the likelihood function for this future observation and ξ *does* depend on v_n. Examples such as this have often been touted as counterexamples to the LP. There are at least two possible replies.

The first possible response is to simply exclude problems involving such unobserved Z from consideration. This was essentially the tack we took earlier in the monograph, motivated by a desire for simplicity of exposition. This response is clearly not very satisfying.

A second possible response is to redefine the likelihood function so as to incorporate Z. In the first edition of this monograph it was essentially suggested that one *define* the likelihood function for $\theta = (z,\omega) = (y, w; \xi, \eta)$ to be

$$\ell_x(\theta) = f_{(\xi,\eta)}(x,y,w); \tag{3.5.1}$$

this is, of course, just the joint density of (X,Z) given the parameter $\omega = (\xi,\eta)$, but here it is to be considered a function of the unknown $\theta = (z,\omega)$ when the observed value X = x is inserted. Such redefinition of $\ell_x(\theta)$ indeed works, in the sense that the LP will still then apply and be derivable from appropriate versions of the Conditionality Principle and Sufficiency Principle. We have not carefully investigated this, however. (It should be emphasized that (3.5.1) is not the density of X, given θ, so that this likelihood function is quite different from (3.1.1). For Bayesians, the distinction is whether to include the unobserved variable Z as part of the model parameter or as part of the observation; we will argue in the next section that it makes no difference.)

While (3.5.1) can be used to establish the LP in this more general context, it has certain practical limitations as a definition of likelihood. The most serious limitation is that it must be utilized very cautiously. Common techniques such as *maximum likelihood* can often be disastrous if applied directly to this $\ell_x(\theta)$. For examples, see Bayarri, DeGroot and Kadane (1987); henceforth, BDK.

A related objection to (3.5.1) is that its definition is, in a sense, quite arbitrary. Extensive discussion of this point can also be found in BDK, with many examples. It is a point with which we essentially agree but, following Berliner (1987), view as tangential to the LP. The LP leaps into action *after* X, Z, ω, and $f_\omega(x,z)$ have been defined, and X = x observed. The process of getting to this point is inherently vague and rather arbitrary; but that doesn't alter the fact that, having reached this point and assuming that the model is correct, all information about $\theta = (z,\omega)$ is contained in (3.5.1) for the given data.

While (3.5.1) is thus formally satisfactory for use in the LP, the practical difficulties surrounding its use and definition suggest looking for an alternative "likelihood function." A very appealing possibility is presented in Butler (1987), discussion of which we defer to the next section. Among the many other references discussing likelihood for unobserved variables (typically in prediction) are Geisser (1971), Kalbfleisch (1971), Lauritzen (1974), Aitchison and Dunsmore (1975), Hinkley (1979), and Butler (1986).

Design problems deserve special emphasis. Before the experiment is conducted, X itself is the unobserved variable, and should hence be identified with Z in the above formulation. (In sequential or multistage experiments, at each step or stage the previously taken observations are x, while the future observations are Z.) The LP does not forbid averaging over *unobserved* variables, and so does not formally contraindicate use of many classical design criteria. For instance, the LP does *not* say that it is wrong to choose the sample size in a testing problem by consideration of type I and type II error probabilities. (Of course, after the data have been taken, the LP would argue against use of these pre-experimental error probabilities as measures of evidence for or against the hypotheses.)

While not disallowing the use of classical design criteria, the LP can have a substantial practical effect on design; a proponent of the LP (i.e. a *conditionalist*) would want to design an experiment so as to have a high probability of obtaining accurate conditional (post-experimental) conclusions, rather than mere pre-experimental frequentist assurances of accuracy. The difference in viewpoint can be significant in that the conditionalist can be more flexible in his approach to design, often simply sampling data until enough (conditional) evidence has been accumulated. By the Stopping Rule Principle (discussed in Section 4.2 and shown to be a consequence of the LP) it is quite valid for the conditionalist to employ such stopping rules of convenience. A frequentist analysis, on the other hand, requires that the probabilities of stopping for each possible reason be known at the outset, and that all these stopping probabilities be incorporated in the analysis.

Similarly the LP gives little guidance in assessing the *overall* performance of a decision procedure δ. Such an assessment might be desired in quality control and other situations where a particular procedure will be used repeatedly. Thus suppose one faces a sequence of problems $X_i \sim P_{\theta_i}$, on each of which a certain procedure δ will be used. Evaluation of the procedure δ will typically involve some type of average over the sample space because future observations X_i are unknown; as with design problems, however, this in no way contradicts the LP. (The LP does, of course, say that it is wrong to report such procedure performance assessments as the evidence about a particular θ_i upon observing a particular x_i). See Section 4.1 for further discussion.

3.5.3 Nuisance Variables and Parameters

When $\theta = (\xi,\eta)$ with η a nuisance variable, the LP says that all evidence about θ is contained in the likelihood function $\ell_x(\theta)$; it seems reasonable to interpret this broadly enough to infer that $\ell_x(\theta)$ should also contain all evidence about the part ξ of θ. This can be made formal through the *NUISANCE VARIABLE LIKELIHOOD PRINCIPLE. Since evidence about* θ *depends on* E *and* x *only through* $\ell_x(\theta)$, $Ev_\xi(E,x)$ *also depends on* E *and* x *only through* $\ell_x(\theta)$. *More generally when* $\theta = (y,w;\xi,\eta)$, *where* y *and* ξ *are the important unobserved variables and unknown parameters while* w *and* η *are nuisance variables and parameters,* $Ev_{y,\xi}(E,x)$ *depends on* E *and* x *only through* $\ell_x(\theta)$ *as defined in (3.5.1).*

With this amendment, the LP says that $Ev_\xi(E,x)$ (or more generally $Ev_{y,\xi}(E,x)$) involves E and x only through $\ell_x(\theta) = \ell_x(\xi,\eta)$ (or more generally $\ell_x(\theta) = \ell_x(y,w;\xi,\eta)$), but does not say what to do about η (or (w,η)); the LP does not say how to interpret $\ell_x(\theta)$ so as to isolate the evidence about y and ξ. While this formally falls in the domain of "utilization of the likelihood function," a topic that we are avoiding, a brief discussion of certain methods of dealing with such nuisance quantities is desirable.

The first key observation is a formalization of the suggestion in Butler (1987) for dealing with nuisance variables or parameters that have *known* distributions:

MARGINALIZATION PRINCIPLE: If the distribution of an unobserved nuisance variable or parameter is given, form a marginal likelihood function *from the joint density of* X *and the nuisance variable or parameter by simply integrating out the nuisance variable or parameter in this joint density.*

The first step in this marginalization process can always be done; w can be immediately eliminated (if present) because $\ell_X(\theta) = f_{(\xi,\eta)}(x,y,w)$ specifies its distribution. Thus $\ell_X(\theta)$ can be reduced to

$$\ell_X^*(y,\xi,\eta) = \int f_{(\xi,\eta)}(x,y,w)\, dw.$$

A further marginalization step can be taken when the distribution of η (or part of η) is given. Thus if $\eta = (\eta^1,\eta^2)$, and it is given that η^2 has density $\pi(\eta^2|\xi,\eta^1)$, the likelihood function can be further marginalized to

$$\ell_X^*(y,\xi,\eta^1) = \int f_{(\xi,\eta)}(x,y,w)\, \pi(\eta^2|\xi,\eta^1)\, dw\, d\eta^2. \tag{3.5.2}$$

EXAMPLE 11.1. Consider the random effects problem where

$$X_{ij} = \eta_i + \varepsilon_{ij}, \qquad i = 1,\ldots,I, \qquad j = 1,\ldots,J,$$

the ε_{ij} being i.i.d. $\mathcal{N}(0,\sigma^2)$ and the η_i being i.i.d. $\mathcal{N}(\mu,\tau^2)$; here σ^2, μ, and τ^2 are unknown. Suppose that interest centers on the "hyperparameters" $\xi = (\mu,\tau^2)$. Then the parameters $\eta^1 = \sigma^2$ and $\eta^2 = (\eta_1, \eta_2,\ldots, \eta_I)$ are nuisance parameters, and the distribution of η^2 is given. Indeed $\pi(\eta^2|\xi)$ is $\mathcal{N}_I(\mu\underset{\sim}{1},\tau^2\underset{\sim}{I})$, where $\underset{\sim}{1} = (1,\ldots,1)^t$ and $\underset{\sim}{I}$ is the identity matrix. A standard calculation (cf. Berger (1985)) then yields for (3.5.2) (note that (y,w) is not present here)

$$\ell_x^*(\xi,\eta^1) = \ell_x^*(\mu,\tau^2,\sigma^2)$$

$$\propto \frac{\exp\{-\sum_{i=1}^{I}(\bar{x}_i-\mu)^2/[2(\tau^2+\frac{\sigma^2}{J})]\}\ \exp\{-s^2/(2\sigma^2)\}}{(\tau^2+\sigma^2/J)^{I/2}\ \sigma^{I(J-1)}},$$

where $\bar{x}_i = \sum_{j=1}^{J} x_{ij}/J$ and $s^2 = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$.

The suggestion to use (3.5.2) as the likelihood was made in Butler (1987) to answer the criticisms in Bayarri, DeGroot, and Kadane (1987) concerning the arbitrariness and difficulty in use of the likelihood defined in (3.5.1); use of (3.5.2) seems to be quite successful in this regard. We support using (3.5.2) as the "practical" definition of likelihood, noting that it is fully consistent with our preferred (see Chapter 4) Bayesian approach to utilization of $\ell_x(\theta)$. Most non-Bayesians would also probably approve of (3.5.2) as the definition of likelihood; failure to do so leaves one open to the serious criticisms in Bayarri, DeGroot, and Kadane (1987). It is also probably true that a version of the LP based on (3.5.2) could be shown to follow (with certain qualifications - cf. the comments at the end of the section) from versions of the Conditionality Principle and Sufficiency Principle. We have not looked into the matter, however.

Use of (3.5.2) does not completely solve the nuisance parameter problem, of course, because $\ell_x^*(y,\xi,\eta^1)$ still depends on the nuisance parameter η^1. There is, unfortunately, no "consensus" approach to elimination of η^1. In the remainder of the section, a brief introduction to some of the proposed methods for elimination of η^1 will be given.

The Bayesian approach to the problem is conceptually straightforward. One simply determines $\pi(\eta^1|\xi)$, the conditional prior density of η^1 given ξ, and calculates the *reduced likelihood function*

$$\ell_x^B(y,\xi) = \int \ell_x^*(y,\xi,\eta^1)\ \ \pi(\eta^1|\xi)\ d\eta^1. \tag{3.5.3}$$

The product of this and the marginal prior density, $\pi(\xi)$, will be proportional

to the posterior distribution of (y,ξ) given x, so that $\ell_x^B(y,\xi)$ clearly suffices for the Bayesian. A strong case can be made that even the non-Bayesian conditionalist should operate by using (3.5.3), with $\pi(\eta^1|\xi)$ chosen to be some "noninformative" prior density for η^1 given ξ. Presentation of this case would, unfortunately, take us too far afield.

The most common non-Bayesian approach to elimination of η^1 is through maximization: i.e., consideration of

$$\tilde{\ell}_x(y,\xi) = \sup_{\eta^1} \ell_x^*(y,\xi,\eta^1).$$

The dangers in use of $\tilde{\ell}_x$ have been well-documented and have resulted in a search for alternative methods (see Section 5.2 for references).

Alternative non-Bayesian methods typically approach the problem of eliminating η^1 through ideas of partial or conditional likelihood. The idea of partial likelihood (cf. Kalbfleisch (1974), Sprott (1975), Cox (1975), Dawid (1975, 1980), Barndorff-Nielsen (1978, 1980), Hinkley (1980), and Kay (1985)) is to factor the likelihood as (ignoring, for simplicity, future observations $Z = (y,w)$ and the possibility that part of η has a known distribution)

$$\ell_x(\theta) = \ell_x^1(\xi)\ \ell_x^2(\xi,\eta), \tag{3.5.4}$$

and then to work with $\ell_x^1(\xi)$ exclusively. This is successful when ℓ_x^2 does not contain much information about ξ, or when the information is very hard to extract because of high variation due to η. It is particularly attractive in the special case (to which we return in Chapter 4) in which ℓ_x^2 contains *no* information about ξ, i.e. in which

$$\ell_x(\theta) = \ell_x^1(\xi)\ \ell_x^2(\eta). \tag{3.5.5}$$

This arises when an ancillary statistic T exists for ξ, ancillary in the strong sense that

$$f_\theta(x) = g_\xi(x|T)\ h_\eta(T);$$

(3.5.5) is then immediate. (Other, broader, definitions of ancillarity also appear in the literature, but lead to expressions as in (3.5.4) rather than (3.5.5). Also, attempts have been made to find approximate decompositions of the form (3.5.5); cf. Hinde and Aitkin (1986).)

EXAMPLE 12. Suppose E consists of observing

$$X = ((Y_1,Z_1),\ldots,(Y_n,Z_n)),$$

where the (Y_i,Z_i) are i.i.d. pairs having a common bivariate normal distribution with unknown mean (μ_Y,μ_Z) and covariance matrix

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} .$$

Of interest is the regression of Y on Z; thus interest centers on $\xi = (\alpha,\beta,\tau^2)$, where

$$\alpha = \mu_Y - \beta\mu_Z,\ \beta = \frac{\sigma_{12}}{\sigma_{22}},\ \tau^2 = \sigma_{11}(1 - \frac{(\sigma_{12})^2}{\sigma_{11}\sigma_{22}}),$$

since $E(Y_i|Z_i) = \alpha+\beta Z_i$ and τ^2 is the conditional variance of Y_i given Z_i.

Letting $\theta = (\mu_Y, \mu_Z, \sigma_{11}, \sigma_{12}, \sigma_{22})$, $\eta = (\sigma_{22}, \mu_Z)$, and $T = (Z_1,\ldots,Z_n)$, a standard calculation gives

$$f_\theta(x) = \frac{k_1}{\tau^n} \exp\{-\frac{1}{2\tau^2} \sum_{i=1}^{n} [y_i-(\alpha+\beta z_i)]^2\} \frac{k_2}{\sigma_{22}^{n/2}} \exp\{-\frac{1}{2\sigma_{22}} \sum_{i=1}^{n} (z_i-\mu_Z)^2\}$$

$$\equiv g_\xi(x|T)h_\eta(T).$$

Thus (3.5.5) is satisfied (and, indeed, T is ancillary for ξ).

It seems natural, when (3.5.5) holds, to state that all evidence about ξ available from E and x is summarized in $\ell_x^1(\xi)$. Thus, in Example 12, it seems natural to base the regression analysis on $g_\xi(x|T)$, the conditional distribution of the Y_i given the observed z_i. This is, indeed, virtually always done in regression; the z_i are treated as nonrandom, i.e., are

conditioned upon.

Basing the analysis only upon $\ell_x^1(\xi)$ is not always justified. If knowledge of η would communicate information about ξ, then $\ell_x^2(\eta)$ cannot, theoretically, be ignored. (For practical reasons, however, one might frequently ignore such information - see Section 4.5.4) The most natural way to rigorously state this is in terms of Bayesian analysis: if ξ and η are apriori independent, then $\ell_x^2(\eta)$ contains no information about ξ. This is clear, since then (3.5.3) becomes (ignoring y)

$$\ell_x^B(\xi) = \int \ell_x^1(\xi)\, \ell_x^2(\eta)\, \pi_2(\eta)\, d\eta \propto \ell_x^1(\xi).$$

The standard conditioning on the z_i in Example 12 is thus rigorously justifiable only when μ_Z and σ_{22} are felt to be apriori independent of α, β, and τ^2, a reasonable assumption in many situations.

Although Bayesian reasoning provides the intuitive basis for stating that a nuisance parameter carries no information about ξ, we will sidestep the issue and simply give an operational definition compatible with the LP.

DEFINITION. *Suppose* E *is such that* (3.5.5) *is satisfied. Let* E^η *be the "thought" experiment in which, in addition to* X, η *is observed. Then* η *is a* <u>*noninformative nuisance parameter*</u> *if* $Ev_\xi(E^\eta,(x,\eta))$ *is independent of* η.

NONINFORMATIVE NUISANCE PARAMETER PRINCIPLE (NNPP). *If* E *is as in* (3.5.5) *and* η *is a noninformative nuisance parameter, then*

$$EV_\xi(E,x) = Ev_\xi(E^\eta,(x,\eta)).$$

The NNPP states the "obvious," that if one were to reach the identical conclusion for *every* η, were η known, then that same conclusion should be reached even if η is unknown. This principle will be used in the discussion of random stopping rules and random censoring in Chapter 4.

As a final qualification, it should be noted that each of these methods for suppressing the role of nuisance parameters is *only* applicable when a decision or action is to be taken on the basis of evidence already recorded, and no further taking of evidence is contemplated. For example, the Likelihood Principle does not imply that the Bayesian's reduced likelihood function, $\ell_x^B(y,\xi)$, summarizes all evidence from an experiment E about a parameter of interest ξ and an unobserved variable of interest y, if that evidence must later be combined with other evidence from further trials also governed by the same nuisance parameter η. Future observations may offer new evidence about the *joint* distribution of ξ and η; by integrating away (or by maximizing away) the nuisance parameter η we would lose the chance to use that new evidence to transform present evidence about η into evidence about ξ. Thus, in Example 11.1, it would not suffice to carry along only $\ell_x^*(\mu,\tau^2,\sigma^2)$ if additional replications x_{ij} (for i = 1,...,I) were to be obtained at a later time. Even if future observations will not be taken, a Bayesian could not report $\ell_x^B(\xi)$ as a complete summary of the evidence to another Bayesian who might use a different conditional prior $\pi(\eta^1|\xi)$; despite the nuisance, the entire likelihood function $\ell_x^*(\xi,\eta^1)$ must be reported in order to convey all information.

3.6 CRITICISMS OF BIRNBAUM'S AXIOMATIC DEVELOPMENT

Birnbaum's axiomatic development of the LP has been subjected to considerable scrutiny. Errors in Birnbaum's arguments did exist, as was mentioned in Section 3.4.1 (see also Birnbaum (1972), Basu (1975), Joshi (1976), and Godambe (1979)), but these errors were correctable and did not affect the basic truth of the arguments. Also easily handled are certain criticisms of the LP arising from its misapplication or misinterpretation. Several such misapplications and misinterpretations have already been mentioned; for completeness we restate them here.

(i) The LP applies only when θ includes all unknowns relevant to the problem. For design, prediction, sequential analysis, meta-analysis, and in many scenarios, the important unknowns often include more than just θ, the unknown parameter of the probability model. But the LP can be reformulated to include such unknowns; see Section 3.5.

(ii) Sometimes a frequentist measure of the performance of a procedure - such as a sampling inspection plan or a diagnostic test - is specified, by contract or law, to be of primary interest. Then, of course, the LP (when stated for θ alone) does not apply.

(iii) There can be ambiguities in the definition of the likelihood function. The problem can usually be resolved, however, by the approaches discussed in Sections 3.4 and 3.5.

(iv) There can be situations in which the *choice* of experiment conveys information about θ. For instance, one might judge that the experimenter never would have chosen the given experiment unless he suspected that, say, θ was small. The LP will still then apply, in the sense that the *experimental* evidence is still contained in $\ell_x(\theta)$; it is just that one will then have additional evidence provided by the choice of experiment. (In a sense, the choice of experiment should be treated as additional data.)

(v) There are periodically attempts to prove the LP wrong by arguing, in a given example, that a particular likelihood-based method (e.g., maximum likelihood estimation) gives a bad result. But the LP prescribes no particular method for utilization of $\ell_x(\theta)$. This issue is extensively discussed in Chapter 5.

(vi) The LP does not apply to the information conveyed about *different* parameters from different experiments. It may be tempting to say that, if E_1 is binomial (n,θ_1) and E_2 is binomial (n,θ_2) and 10 successes (or ones) are observed in each of the experiments, then since the likelihood functions

for the two situations are the same (as functions), one should reach the same conclusions about θ_1 and θ_2. But the LP does *not* say this; it applies only when θ_1 and θ_2 are the same parameter, i.e., are physically or conceptually the same quantity.

There have been a number of criticisms directed at the explicit and implicit principles used in Birnbaum's development of the LP. We address these criticisms in this and the following sections.

3.6.1 The Model Assumption

The most frequently expressed criticism of the LP is that it is supposedly very dependent on assuming a particular parametric model with a density for X; since models are almost never known exactly, it is felt that the LP is only rarely applicable. It is, of course, easy to criticize almost any statistical theory for being model dependent, but let us examine the issue seriously anyway.

The first point to note is that, even if there are various possible models under consideration, the LP still says that the information in the data, for any possible model, is contained in the likelihood function for that model. The evidence conveyed by the data certainly changes as different models are considered, but the likelihood functions should still be considered the vehicles of this evidence.

To be more formal about this, we need only recall that θ need not be restricted to being a typical parameter, and indeed can represent various models. The situation of discrete X is easiest to see: thus, if $\mathcal{X} = \{x_1, x_2, \ldots\}$, we could simply let $\theta = (\theta_1, \theta_2, \ldots)$ denote a point on the infinite dimensional simplex

$$\Theta = \{\theta: \ 0 \leq \theta_i \leq 1 \text{ and } \sum \theta_i = 1\},$$

and define

$$P_\theta(x_i) = \theta_i.$$

Then $\{P_\theta\}$ is the class of all probability distributions on $\mathcal{X}$, and the LP applies to this completely nonparametric setup, as well as to any situation

where a restricted class of models (corresponding to some subset of Θ) is considered. Of course, we will usually only be interested in some function $\psi(\theta)$, but if all the evidence about θ is contained in the likelihood function, then the same should be true of $\psi(\theta)$. The argument in Section 3.4 in favor of considering only discrete situations (in foundations) thus indicates that the LP always applies.

Even in continuous situations, there is no need to tie the LP to restrictive parametric models. For instance, consider the following example.

EXAMPLE 13. Suppose $X_1,\ldots,X_n$ are i.i.d. observations from some distribution, known to have a density (with respect to a given measure ν), but otherwise unknown. Let Θ be the set of all such densities, so that the density of $X = (X_1,\ldots,X_n)$ is

$$f_\theta(x) = \prod_{i=1}^{n} \theta(x_i).$$

For instance, this would be the situation if the x_i were known to have a distribution with a continuous density with respect to Lebesgue measure on a Euclidean space. Thus a likelihood function does exist in such nonparametric situations, and the LP (more properly the heuristic LP discussed in Section 3.4.1) would apply. "Robustness" problems typically fall into the setting where a subset of Θ (say, all densities close to some prescribed parametric family of densities) is under consideration. Again, the LP will usually apply.

It can be argued, of course, that one may be dealing with a general non-dominated family $\{P_\theta\}$ or, alternatively, that the LP does not really apply to the nondiscrete case, but there is still the RLP to contend with. Again, θ could just be used to index the distribution, so the RLP will essentially always be applicable, yet it is inconsistent with frequentist reasoning and will be seen to yield strong conclusions such as the Stopping Rule and Censoring Principles. In conclusion, therefore, although the LP is usually stated in terms of a particular parametric model with densities, it (or its generalizations) are essentially always applicable. (Implementing the

LP can, of course, be much more difficult in nonparametric situations, as will be discussed in Chapter 5.)

3.6.2 The Evidence Assumption

A less common criticism of Birnbaum's development is the questioning of the existence or meaning of Ev(E,x). As noted in Section 3.3.1, however, this can have essentially any interpretation (initially) and need not consist of any single measure, so it is hard to see the force of this objection.

3.6.3 The Weak Conditionality Principle

A possible point of criticism is the Weak Conditionality Principle. Indeed, a committed frequentist might well reject this principal, saying it is based on the erroneous belief that one can obtain evidence (in the intuitive sense) about a particular θ from a particular experiment (c.f., Neyman (1957, 1977)). Instead, the argument goes, one can only state the performance of a procedure that will be used repeatedly, and this should (or at least could) involve averaging over both E_1 and E_2. In a sense, this position is logically viable. Its scientific desirability is very questionable, however, as Example 2 in Section 2.1 illustrates. This issue will be discussed further in Section 4.1.

Durbin (1970) raises the point that if the Weak Conditionality Principle is allowed to apply only to conditioning variables which depend solely on a minimal sufficient statistic, then the LP does not follow. (This is because, in the proof of Theorem 1, the conditioning statistic, J, is not part of the minimal sufficient statistic when the two likelihood functions are proportional. Sufficiency says "discard J," after which it is clearly impossible to condition on J.) No plausible reason has been advanced for so restricting the Weak Conditionality Principle, however, and the idea seems unreasonable as a reexamination of Example 2 shows.

EXAMPLE 2 (continued). Let x_C denote the outcome of the California experiment, and suppose that there was some possible outcome x_N of the New York experiment

for which $\ell_{x_C}(\theta)$ would have been proportional to $\ell_{x_N}(\theta)$. Then, in the mixed experiment E*, the outcomes x_C and x_N would be identified by a minimal sufficient statistic, precluding application of the restricted WCP. If, however, there was no x_N, then conditioning on the California experiment would be allowed. Thus, by Durbin's argument, whether or not one chooses to condition on the actually performed California experiment with observation x_C would depend on the existence, or lack thereof, of an observation x_N, in the unperformed New York experiment, having a likelihood function proportional to that of x_C. Such dependence of conditioning on the incidental structure of an *unperformed* experiment would be rather bizarre.

Other rejoinders to Durbin's criticism can be found in Birnbaum (1970) and Savage (1970). Savage invokes a "continuity" argument, showing that following Durbin's restricted WCP can involve drawing substantially different conclusions when a problem is changed in an insignificant way (such as slightly perturbing the likelihood function of x_N above).

3.6.4. The Sufficiency Principle

Surprisingly, the most common and serious axiomatic criticisms of the LP are those directed at the Sufficiency Principle. This may seem strange, sufficiency being such a central part of classical statistics, but issues can be raised.

The first issue is a valid limitation of the SP: if one faces a decision in which the consequences (or loss) depend on x, and not just on the action taken and unknown θ, then the SP need not be valid. Such situations are relatively rare, however, and could be handled with a reformulation of the LP to the effect that Ev(E,x) should depend on $\ell_x(\theta)$ *and* x.

A second issue, raised by Kalbfleisch (1974, 1975), is that the LP does not follow from the WCP and SP if sufficiency is not allowed to apply to simple mixture experiments. The problems with such a restriction of sufficiency are that (i) It seems artificial, there being no intuitive reason to restrict sufficiency to certain types of experiments; (ii) It is difficult and perhaps

impossible to clearly distinguish between mixture and non-mixture experiments (cf. the discussion in Kalbfleisch (1975)); (iii) Mixture experiments can often be shown to be equivalent to non-mixture experiments (cf. Birnbaum (1962a)), making the distinction seem unreasonable; and (iv) In almost any situation, behavior in violation of sufficiency can be shown to be inferior (see Section 3.7). Evans, Fraser, and Monette (1986) contains further discussion.

The most serious criticism of the SP comes from ideas of Barnard (cf. Barnard, Jenkins, and Winsten (1962), Barnard (1980, 1981), Barnard and Godambe (1982), and the discussions in Birnbaum (1962a), Basu (1975), and Wilkinson (1977)) and Fraser (cf. Fraser (1963, 1968, 1972, and 1979)). They question the "sufficiency" of representing the experimental structure solely in terms of probability distributions on the sample space indexed by the unknown θ; Dawid (1977) called this the Distribution Principle (DP). The criticism of the DP (and hence the SP) is that there may be important information lost concerning the relationship between X, θ, and the "randomness" in the problem. (An important observation is that, while relevant to the LP, this criticism is not relevant to certain of the most controversial relatives of the LP, such as the Stopping Rule Principle; cf. Dawid (1986).)

This criticism turns out to be quite difficult to answer, striking at the core of virtually all approaches to statistics. One response is to attempt an axiomatic development of the LP which incorporates "structural" information. Such a development can be found in Berger (1984a), but is something of a failure, containing a suspect axiom from the above viewpoint. Also in Berger (1984a), therefore, the issue is addressed from the viewpoint of coherency and admissibility; it is shown that incorporating "structural" information in violation of sufficiency results in inferior behavior. These arguments are familiar, but because of the importance of the issue and the bearing these arguments have on *any* proposed violation of the LP, they are reviewed in Section 3.7. (Evans, Fraser, and Monette (1986) also contains relevant discussion.) Incidentally, the need to resort to coherency and admissibility bears out I. J. Good's discussion of Birnbaum (1962a), that

derivation of the LP via the WCP and SP is mainly a *sociological* contribution to statistics, since Bayesian coherency axiomatics would give the LP directly. While agreeing, we feel that the sociological contribution is very substantial; many people will (for whatever reasons) accept the WCP and SP, yet resist the LP.

In the remainder of this section, we briefly outline the objection to the SP that is raised in the theories of Pivotal Inference (cf. Barnard (1980, 1982) and Barnard and Sprott (1983)) and Structural Inference (cf. Fraser (1968, 1972, 1979)). The key idea is that 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 $\mathcal{Q}$. 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 (cf. 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 14. 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

(3.6.1) $$X = h(\theta,\omega) = (\sigma_1\omega_1,\ \tau\omega_1+\phi\omega_2);$$

(ii) ω is the same but

(3.6.2) $$X = h^*(\theta,\omega) = (\tau'\omega_1+\phi'\omega_2,\ \sigma_2\omega_1),$$

where $\sigma_2 = \sqrt{\tau^2+\phi^2}$, $\tau' = \sigma_1\tau/\sigma_2$, and $\phi' = \sigma_1\phi/\sigma_2$. In Pivotal Inference, one would write (3.6.1) and (3.6.2) as

(3.6.1)' $$\omega = (\omega_1,\omega_2) = (X_1/\sigma_1,\ (X_2-\tau X_1/\sigma_1)/\phi),$$

(3.6.2)' $$\omega = (\omega_1,\omega_2) = (X_2/\sigma_2,\ (X_1-\tau' X_2/\sigma_2)/\phi'),$$

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 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, Σ) distribution.

In the above type of situation, which we will call a P-S (for Pivotal-Structural) situation, an experiment is specified by $E = (X, \theta, h, \omega, Q)$. As in Example 14, one could have a single probability-modeled experiment, $E = (X, \theta, \{P_\theta\})$, arising from more than one P-S experiment. In such situations there is a definite loss of structure in reduction to a probability model. The question that will be addressed in the next section is whether this structure contains any useful information. Of course, the point is moot unless P-S theory actually recommends differing actions or conclusions for differing P-S models which have the same probability model. An example where this is the case for Pivotal theory can be found in the discussion by Barnard in Berger (1984a). A possible example for Structural theory is Example 14.

EXAMPLE 14 (continued). A part of Structural Inference is the construction of "structural distributions" for θ. These can presumably be used, in the same

manner as 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 (3.6.1) and (3.6.2), respectively, by (see Fraser (1968))

$$\pi_1(\theta|s) = K_1(s) f_\theta(s) \sigma_1^2 \phi^{-1}, \tag{3.6.3}$$

and

$$\pi_2(\theta|s) = K_2(s) f_\theta(s) (\tau^2+\phi^2)^{-1} \phi^{-1}. \tag{3.6.4}$$

(These happen to correspond to the posterior distributions with respect to the right invariant Haar measures on the lower and upper triangular group decompositions of $\sharp$.) Examples will be given in the next section which show that use of these differing structural distributions can lead to differing conclusions.

3.7 VIOLATION OF THE LIKELIHOOD PRINCIPLE: INADMISSIBILITY AND INCOHERENCY

3.7.1 Introduction

The alternative to justification of the LP from "first principles" is to show that behavior in violation of the LP is inferior. The only convincing method of demonstrating such inferiority is to show that such behavior can be improved upon in repeated use. We thus turn to measures of long run performance of statistical procedures or methods. We will not argue that measures of long run performance have an important practical role in statistics (as frequentists would argue), but we will argue that they have the important theoretical role of providing a test for proposed methodologies: it cannot be right (philosophically) to recommend repeated use of a method if the method has "bad" long run properties. Both of the main approaches to long run evaluation, decision theory and betting coherency, will be discussed. We will further 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.

A violation of the LP will occur (in the discrete case) 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)

$$f_\theta^1(x_1') = c\, f_\theta^2(x_2') \quad \text{for all } \theta, \tag{3.7.1}$$

and for which

$$Ev(E_1,x_1') \neq Ev(E_2,x_2'). \tag{3.7.2}$$

Consider now the mixed experiment E*, in which J = 1 or 2, with probability $\frac{1}{2}$ each, is observed (independent of all elements of the E_i), and experiment E_J is then performend. According to the WCP,

$$Ev(E^*,(j,x_j)) = Ev(E_j,x_j),$$

which combined with (3.7.2) yields the conclusion

$$Ev(E^*,(1,x_1')) \neq Ev(E^*,(2,x_2')). \tag{3.7.3}$$

It will be behavior according to (3.7.3) that is shown to be inferior in repeated use.

In the nondiscrete case, we can consider violation of the RLP (see Section 3.4.1). Thus suppose that, in the situation of the RLP, there exists a set $A \subset U_1$, with $P_\theta^1(A) > 0$ for all θ, and such that

$$Ev(E_1,x_1) \neq Ev(E_2, \varphi(x_1)). \tag{3.7.4}$$

Again considering the mixed experiment E* and applying the WCP, one obtains that, for $x_1 \in A$,

$$Ev(E^*,(1,x_1)) \neq Ev(E^*,(2, \varphi(x_1))), \tag{3.7.5}$$

behavior which will be shown to also have bad long run properties.

The experiment E* will preserve all "structural" features of E_1 and E_2, so the only objection that could be raised concerning the above line of reasoning is the use of the WCP. Although some frequentists will reject the WCP (and are then exempt from the conclusions of this section) most will find such rejection difficult. Virtually all other theories accept the WCP, and are hence subject to evaluation through E*. Among the theories which seem to accept the WCP, and yet sometimes advocate violation of the LP, are (the already discussed) Pivotal Inference and Structural Inference, Fiducial

Inference, Plausibility Inference (see Barndorff-Nielsen (1976)), and certain noninformative prior Bayesian theories (see Example 9 in Section 3.1). It should be noted that it is actually rather rare for these theories to conflict with the LP. Indeed the conflict would not be worth making an issue of, were it not for the purported refutations of the LP that seem to arise from these theories. The "refutations" are always of the form - "following theory A conflicts with the LP, so the LP must be wrong." We will argue (via long run evaluation) that the reverse is true.

3.7.2 Decision Theoretic Evaluation

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* $L(a,\theta)$. 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,

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

Following Hill (1974b) and Berger (1984a), 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* $\underset{\sim}{\theta} = (\theta_1,\theta_2,\ldots) \in \mathcal{C}$ (*a class of relevant sequences*), *and for each* θ_i *the experiment* E *will be independently performed yielding an observation* X_i (*from the distribution* P_{θ_i}). *Player* j *will use* $\delta^j(x_i)$, *paying to the other player his "loss"* $L(\delta^j(x_i),\theta_i)$. *After* n *plays, Player 2 will have won*

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

If, for any $\underset{\sim}{\theta} \in \mathcal{C}$,

$$P_{\underset{\sim}{\theta}}(\liminf_{n\to\infty} \frac{1}{n} S_n > 0) = 1, \tag{3.7.6}$$

then δ^2 *will be called* C-better *than* δ^1.

Although there are a number of reasonable choices for $\mathcal{C}$ in the Evaluation Game, a particularly attractive choice is

$$\mathcal{C}_C = \{\underset{\sim}{\theta}: \text{ there exists a compact set } K \subset \Theta \text{ for which } \theta_i \in K \text{ for every } i\}.$$

This choice is attractive because reality is bounded, but the bound is often unknown (and, hence, we entertain unbounded models). With such a $\mathcal{C}$, the Evaluation Game seems to be a 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, it would certainly seem unwise to call δ^1 fundamentally sound. The following theorem is useful in dealing with $\mathcal{C}_C$.

THEOREM 4. *Suppose* $R(\theta,\delta^2) < R(\theta,\delta^1)$ *for all* θ, *that* $[R(\theta,\delta^1)-R(\theta,\delta^2)]$ *is continuous in* θ, *and that the random variables*

$$Z_i = [L(\delta^1(X_i),\theta_i)-L(\delta^2(X_i),\theta_i)]$$

have uniformly bounded variances (which is trivially satisfied if L *is bounded). Then* δ^2 *is* $\mathcal{C}_C$*-better than* δ^1 *in the Evaluation Game.*

Proof. Define

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

By the strong law of large numbers,

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

so that, for any $\underset{\sim}{\theta}$,

$$P_{\underset{\sim}{\theta}}(\liminf_{n\to\infty} \frac{1}{n} S_n > 0) = P_{\underset{\sim}{\theta}}(\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi(\theta_i) > 0). \tag{3.7.7}$$

But since the θ_i lie in some compact set and $\psi(\theta)$ is continuous and

positive,

$$\inf_{i<\infty} \psi(\theta_i) > 0.$$

The conclusion is immediate from (3.7.7). ||

The condition "$R(\theta,\delta^2) < R(\theta,\delta^1)$ for all θ" in Theorem 4 implies that δ^1 is *inadmissible* in a frequentist decision-theoretic sense. This is really the key condition in the failure of δ^1 in the Evaluation Game. Indeed we can, in a loose sense, equate such failure with inadmissibility. The exact relationship depends on the choice of C in the Evaluation Game, so we will sometimes use the term "inadmissibility" to encompass the whole idea.

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, and inferior performance 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 (or other) 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 (1974b):

> "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 an example of a reasonable "inference" loss, imagine that a *given* "confidence" set C is to be used, and that the desired inference is a measure, $\delta(x)$, of the "chance" or "confidence" with which we wish to assert that C contains θ. No matter what interpretation is attached to $\delta(x)$, it seems reasonable to measure its performance via a loss function which reflects whether or not δ does a good job of indicating the presence of θ in C. One such loss function is

$$L(\delta(x),\theta) = (I_C(\theta)-\delta(x))^2, \tag{3.7.8}$$

essentially the quadratic scoring function of deFinetti (1962). (Any other proper scoring function would also be reasonable - c.f. Lindley (1982).) Note that for any "posterior" distribution, $\pi(\theta|x)$, for θ, the optimal choice of $\delta(x)$ in (3.7.8) is

$$\delta^{\pi}(x) = E^{\pi(\theta|x)}I_C(\theta) = P^{\pi(\theta|x)}(\theta \in C), \tag{3.7.9}$$

i.e., the posterior probability of C. Thus, to test the inferences provided by Structural Inference in Example 14, it seems reasonable to use the structural distributions provided by (3.6.3) and (3.6.4) to determine $\delta^{\pi_1}(s)$ and $\delta^{\pi_2}(s)$ via (3.7.9), and then test the implied procedure in the Evaluation Game for the mixed experiment E* (see Section 3.7.1). We will

return to this example later.

The simplest situation, in which violation of the LP (or RLP) results in failing the Evaluation Game for E*, is when L is strictly convex in "a" for all θ. (For some other situations, see Berger (1984a).) Consider first the discrete case in Section 3.7.1. A violation of the LP (see (3.7.3)) would imply use of a δ^1 in E* for which

$$\delta^1((1,x_1')) \neq \delta^1((2,x_2')). \tag{3.7.10}$$

Consider, however, the procedure

$$\delta^2((j,x_j)) = \begin{cases} \frac{c}{(c+1)}\,\delta^1((1,x_1'))+\frac{1}{(c+1)}\,\delta^1((2,x_2')) & \text{for } x_j = x_1' \text{ or } x_2' \\ \delta^1((j,x_j)) & \text{otherwise,} \end{cases} \tag{3.7.11}$$

where c is from (3.7.1). Using the strict convexity of L, one obtains that

$$L(\delta^2((j,x_j')),\theta) < \frac{c}{(c+1)}\,L(\delta^1((1,x_1')),\theta) + \frac{1}{(c+1)}\,L(\delta^1((2,x_2')),\theta). \tag{3.7.12}$$

An easy calculation, using (3.7.1), then shows that

$$R(\theta,\delta^1)-R(\theta,\delta^2) = \frac{(1+c)}{2c}\,f_\theta^1(x_1')\Delta(\theta), \tag{3.7.13}$$

where $\Delta(\theta)$ is the difference between the right and left hand sides of (3.7.12). Under the additional easily satisfiable conditions of Theorem 4, it is immediate that δ^1 fails the Evaluation Game for all $\underset{\sim}{\theta} \in \mathcal{C}_C$. (This is all, of course, a form of the Rao-Blackwell Theorem.)

EXAMPLE 9 (continued - see Section 3.1). Suppose it is desired to estimate θ under the loss $L = (\theta-a)^2$ (or any other strictly convex loss), and that δ_1 would be recommended for E_1 and δ_2 for E_2, where $\delta_1(9) \neq \delta_2(9)$; thus a violation of the LP will have occurred. (Neither Pivotal nor Structural inference would necessarily recommend different actions here, but the Jeffreys noninformative prior Bayes theory and also Akaike (1982) would seem to so recommend.) The situation meshes exactly with the discrete setting discussed

above, and so if one (following the WCP) used

$$\delta^1((j,x_j)) = \delta_j(x_j) \tag{3.7.14}$$

for the mixed experiment E*, (3.7.13) would hold. It follows from Theorem 4 that δ^1 fails the Evaluation Game for $\underset{\sim}{\theta} \in C_{C^*}$. Note that δ^1 would not fail the Evaluation Game for any $\underset{\sim}{\theta}$ which converged to zero or one. The failure of δ^1 for any $\underset{\sim}{\theta} \in C_{C^*}$, or even more generally for any $\underset{\sim}{\theta}$ which lies within a compact subset of Θ some positive fraction of the time, strikes us, however, as strong enough evidence to rule out using δ^1.

The non-discrete version of the above argument for convex loss would assume (see the discussion around (3.7.4)) that, in violation of the RLP for E*,

$$\delta^1((1,x_1)) \neq \delta^1((2, \varphi(x_1)), \quad \text{for } x_1 \in A. \tag{3.7.15}$$

The analog of (3.7.11) is now

$$\delta^2((j,x_j)) = E[\delta^1((J,X_J))|T(j,x_j)], \tag{3.7.16}$$

the conditional expectation of δ^1 given T, where T is the sufficient statistic (in E*)

$$T((j,x_j)) = \begin{cases} (2, \varphi(x_1)) & \text{if } j = 1 \text{ and } x_1 \in U_1 \\ (j,x_j) & \text{otherwise.} \end{cases}$$

The appropriate versions of (3.7.12) and (3.7.13) can easily be established and under reasonable conditions, failure of δ^1 in the Evaluation Game follows.

EXAMPLE 14 (continued). Suppose it is desired to estimate $\sharp$ (which is equivalent to θ) under the strictly convex loss

$$L(\delta,\sharp) = \mathrm{tr}(\delta\sharp^{-1})-\log\det(\delta\sharp^{-1})-2. \tag{3.7.17}$$

(The loss $L(\delta,\sharp) = \mathrm{tr}(\delta\sharp^{-1}-I)^2$ would work similarly - see James and Stein (1961) and Selliah (1964).) If one treats $\pi_1(\theta|s)$ and $\pi_2(\theta|s)$ in (3.6.3) and (3.6.4) as posteriors and calculates the optimal estimators with respect to

(3.7.17), one obtains

$$(3.7.18)\quad \delta_1(s) = s_L \begin{pmatrix} (n+1)^{-1} & 0 \\ 0 & (n-1)^{-1} \end{pmatrix} s_L^t, \quad \delta_2(s) = s_U \begin{pmatrix} (n-1)^{-1} & 0 \\ 0 & (n+1)^{-1} \end{pmatrix} s_U^t,$$

where $s = s_L s_L^t = s_U s_U^t$, s_L and s_U being lower and upper triangular, respectively. If these estimators would be used in E_1 and E_2, the WCP would lead to using the estimator $\delta^1((j,s)) = \delta_j(s)$ in the mixed experiment E*.

To establish failure of δ^1 in the Evaluation Game, let $A = \{s: \delta_1(s) \neq \delta_2(s)\}$ and note that A has probability one for all θ. This situation satisfies the conditions of the RLP with U_1 and U_2 being the entire sample space, $c(\cdot) \equiv 1$, and φ being the identity map (since the probability space is identical for E_1 and E_2), and also satisfies (3.7.15). The estimator δ^2 in (3.7.16) is simply

$$\delta^2((j,s)) = \tfrac{1}{2}\,\delta^1((1,s)) + \tfrac{1}{2}\,\delta^2((2,s))$$

$$= \tfrac{1}{2}\,\delta_1(s) + \tfrac{1}{2}\,\delta_2(s),$$

and, from the strict convexity of the loss, it follows easily that (for E*)

$$R(\theta,\delta^2) < R(\theta,\delta^1) \quad \text{for all } \theta.$$

Furthermore, the conditions of Theorem 4 can easily be verified in this situation, and so the conclusion of the theorem applies: δ^2 is better than δ^1 in the Evaluation Game for all bounded sequences $\underset{\sim}{\theta}$.

Of course, this same analysis would hold for *any* estimators that differ for E_1 and E_2, not just for δ_1 and δ_2 in (3.7.18). Thus violating the RLP by using different estimators in the two cases seems definitely contra-indicated.

The same kind of conclusion follows in the "inference" situation of giving the "confidence" to be attached to a set C, using a loss such as (3.7.8). If $\pi_1(\theta|s)$ and $\pi_2(\theta|s)$ are used as posteriors to produce probabilities that θ is in C (via (3.7.9)) and these probabilities differ (as will usually be the

case), an analysis virtually identical to that above shows that the violation of the RLP results in an inference for E* which fails the Evaluation Game for all bounded $\underset{\sim}{\theta}$. Again, one could object to evaluating inferences via (3.7.8), but use of any reasonable measure of the performance of inferences would lead to the same conclusion.

3.7.3 Betting Evaluation

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 $\{C(x), \delta(x)\}$ with the interpretation that θ is felt to be in $C(x)$ with probability $\delta(x)$. 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 $\delta(x)$ 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 $\delta(x)$ is thought to be the probability that θ is in $C(x)$, the proposer of $\{C(x), \delta(x)\}$ should be equally willing to accept either the bet that θ is in $C(x)$, at odds of $(1-\delta(x))$ to $\delta(x)$, or the bet that θ is not in $C(x)$, at odds of $\delta(x)$ to $(1-\delta(x))$. An evaluations game, as in Section 3.7.2, is then proposed, in which the master of ceremonies again generates θ_i and X_i, Player 1 stands ready to accept bets on $\{C(x), \delta(x)\}$, and Player 2 bets $s(x)$ at odds determined by $\delta(x)$. 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

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

and of interest is again the limiting behavior of $\frac{1}{n}\sum_{i=1}^{n} W_i$. If, for some $\varepsilon > 0$,

$$(3.7.19) \qquad P_{\underset{\sim}{\theta}}\left(\lim_{n\to\infty} \inf \frac{1}{n} \sum_{i=1}^{n} W_i > \varepsilon\right) = 1$$

for all sequences $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots)$, then $\{C(x), \delta(x)\}$ will be called *incoherent*, or alternatively $s(x)$ will be said to be a *super relevant* betting strategy. If it is merely the case that (3.7.19) holds for any $\underset{\sim}{\theta} \in \mathcal{C}_C$ with $\varepsilon = 0$, then $\{C(x), \delta(x)\}$ will be called *weakly incoherent* or $s(x)$ will be said to be *weakly relevant*. (These concepts can be found in different, but closely related, forms 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), Levi (1980), and Lane and Sudderth (1983).)

If $\{C(x), \delta(x)\}$ 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 $\delta(x)$. 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 $\underset{\sim}{\theta} \in \mathcal{C}_C$ seems quite serious.

Objection 3. Of course, frequentists who quote a confidence level δ for $\{C(x)\}$ remove themselves from the game, since they do not claim that δ 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 $\delta(x)$ for some x. It is after all $\{\delta(x)\}$ 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 $\{\delta(x)\}$ 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 $\{\delta(x)\}$ 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 15. Suppose $\mathcal{X}$ and Θ are the integers, and that $P_\theta(X=\theta+1) = P_\theta(X=\theta-1) = \frac{1}{2}$. We are to evaluate the confidence we attach to the sets $C(x) = \{x+1\}$ (the point $(x+1)$), and a natural choice is $\delta(x) = \frac{1}{2}$ (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 fixed θ, is $\frac{1}{2}[g(\theta+1)-g(\theta-1)] > 0$, from which it is easy to check that $\delta(x) = \frac{1}{2}$ is weakly incoherent. (A continuous version of this example, mentioned in Robinson (1979a), has $X \sim \eta(\theta,1)$, $\Theta = \mathbb{R}^1$, $C(x) = (-\infty, x)$, and $\delta(x) = \frac{1}{2}$.)

In this and other examples where $\{\delta(x)\}$ loses in betting, one can ask the crucial question - Is there a better δ that could be used? The question has no clear answer, because the purpose of δ is not clearly defined. One possible justification for $\delta(x) = \frac{1}{2}$ 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. (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 C(x) will include 1/2 for all x.) An alternative justification can be found by retreating to decision theory, and attempting to quantify how well $\delta(x)$ performs using a loss such as (3.7.8). One can then ask if there is a better δ in terms, say, of the decision-theoretic Evaluation Game for bounded $\underset{\sim}{\theta}$. The answer in the case of Example 15 is - no! A standard limiting Bayes argument can be used to show that $\delta(x) = \frac{1}{2}$ is decision - theoretically admissible for this loss, from which it follows that, for any other δ^*, a bounded (indeed constant) sequence $\underset{\sim}{\theta}$ can be found such that δ is better than δ^* in the Evaluation Game.

The Evaluation Game (or decision-theoretic inadmissibility) with respect to losses such as (3.7.8) 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 $\delta(x)$ "behave differently from ordinary probabilities." And as Hill (1974b) 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."

To show that violation of the LP (or RLP) leads to some form of incoherence, it is again necessary to consider the setup in Section 3.7.1.

Taking the discrete case first, suppose a fixed set $C \subset \Theta$ is assigned "confidence" α_1 in E_1 when x_1' is observed, but "confidence" $\alpha_2 (\neq \alpha_1)$ in E_2 when x_2' is observed. If the WCP is followed for the mixed experiment E*, the confidence function δ employed satisfies

$$\delta((1,x_1')) = \alpha_1 \neq \alpha_2 = \delta((2,x_2')),$$

the appropriate version of (3.7.3). Consider now the betting strategy (see the beginning of the section for interpretation)

$$s((j,x_j)) = \begin{cases} 0 & \text{if } x_j \neq x_1' \text{ or } x_2' \\ c_j\alpha_j & \text{if } j = k \text{ and } x_j = x_1' \text{ or } x_2' \\ -c_j(1-\alpha_j) & \text{if } j \neq k \text{ and } x_j = x_1' \text{ or } x_2', \end{cases}$$

where $c_1 = 1$, $c_2 = c$ (from (3.7.1)), and $k = 1$ or 2 as $\alpha_1 < \alpha_2$ or $\alpha_1 > \alpha_2$, respectively. If this strategy is used with odds corresponding to α_j when (j,x_j') is observed, the expected gain can be easily calculated to be

$$\frac{1}{2} f_\theta^1(x_1')|\alpha_1 - \alpha_2|.$$

If $f_{\theta_i}^1(x_1')$ is bounded away from zero for all bounded sequences $\underset{\sim}{\theta}$, it follows easily that δ is weakly incoherent.

In the nondiscrete case, one replaces α_j above by $\alpha_j(x_j)$ (the "confidence" in C if x_j is observed in E_j), and assumes that, for some $A \subset U_1$ with $P_\theta^1(A) > 0$ for all θ,

$$\alpha_1(x_1) \neq \alpha_2(\varphi(x_1)) \qquad \text{for } x_1 \in A.$$

The corresponding confidence function in the mixed experiment E* is $\delta((j,x_j)) = \alpha_j(x_j)$, which again violates the RLP. Consider, now, the betting strategy

$$s((j,x_j)) = \begin{cases} 0 & \text{if } (j,x_j) \notin A^* \\ c_j(x_j)\alpha_j(x_j) & \text{if } j=k((j,x_j)) \text{ and } (j,x_j) \in A^* \\ -c_j(x_j)(1-\alpha_j(x_j)) & \text{if } j\neq k((j,x_j)) \text{ and } (j,x_j) \in A^*, \end{cases}$$

where $c_1(x_1) \equiv 1$, $c_2(x_2) = c(\varphi^{-1}(x_2))$ (see (3.4.1)),

$$A^* = \{(1,x_1): \ x_1 \in A\} \cup \{(2,\varphi(x_1)): \ x_1 \in A\},$$

and

$$k((j,x_j)) = \begin{cases} 1 & \text{if } j=1 \text{ and } \alpha_1(x_1) < \alpha_2(\varphi(x_1)) \quad \textit{or} \\ & \quad j=2 \text{ and } \alpha_1(\varphi^{-1}(x_2)) < \alpha_2(x_2) \\ 2 & \text{otherwise.} \end{cases}$$

The expected gain for this betting strategy can easily be calculated to be

$$\int_A \frac{1}{2} \, |\alpha_1(x_1)-\alpha_2(\varphi(x_1))| P_\theta^1(dx_1).$$

Weak incoherency will again follow under reasonable conditions.

For general theorems on coherence, consult Heath and Sudderth (1978) and Lane and Sudderth (1983) and the references therein. These theorems indicate that, unless δ for E^* is compatible with some posterior distribution, incoherency will result. A coherent δ will not violate the LP (or RLP), and so incoherence of violation of the LP is quite general. Again, however, this may not be as convincing as the decision-theoretic refutation of violation of the LP which was discussed in Section 3.7.2.