Institute of Mathematical Statistics

LECTURE NOTES–MONOGRAPH SERIES
Shanti S. Gupta, Series Editor

Volume 6

The Likelihood Principle

(Second Edition)

James O. Berger
Purdue University

Robert L. Wolpert
Duke University

Institute of Mathematical Statistics
Hayward, California

Institute of Mathematical Statistics

Lecture Notes–Monograph Series

Series Editor, Shanti S. Gupta, Purdue University

The production of the IMS Lecture Notes–Monograph Series is managed by the IMS Business Office: Nicholas P. Jewell, IMS Treasurer, and Jose L. Gonzalez, IMS Business Manager.

Library of Congress Catalog Card Number: 88-81456

International Standard Book Number 0-940600-13-7

Printed in the United States of America

To my parents, Orvis and Thelma

James Berger

To my wife, Ruta

Robert Wolpert

PREFACE

This monograph began with research designed to provide a generalization of the Likelihood Principle (LP) to quite arbitrary statistical situations. The purpose of seeking such a generalization was to partially answer certain criticisms that had been levied against the LP, criticisms which seemed to prevent many statisticians from seriously considering the LP and its implications. The research effort seemed worthwhile because of the simplicity, central importance, and far reaching implications of the LP.

Background reading for the research revealed a wider than expected range of published criticisms of the LP. In an attempt to be complete and address all such criticisms, the research paper expanded considerably. Eventually it seemed sensible to enlarge the paper to a monograph. This also allowed for discussion of conditioning ideas in general and for a review of the implications of the LP. It was decided, however, to stop short of a general review of conditional *methods* in statistics. In particular, the monograph does not discuss the many likelihood based statistical methodologies that have been developed, although references to these methodologies will be given. This limitation was, in part, because such an endeavor would be far too ambitious, and, in part, because we feel (and indeed argue in Chapter 5) that Bayesian implementation of the LP is the correct conditional methodology.

The mathematical level of the monograph is, for the most part, kept at a nontechnical level. The main exception is the generalization of the LP in Section 3.4, which is (necessarily) presented at a measure-theoretic level, but can be skipped with no loss in continuity. Also, the monograph presupposes no familiarity with conditioning concepts. Indeed Chapter 2 provides an elementary review of conditioning, with many examples.

This second edition was produced under the rather severe constraint that the original manuscript, used for photo-offset printing, was inadvertantly destroyed; only the photos were kept. Thus changes could only be made by retyping entire pages or inserting new pages. A list of corrections that were too minor to justify the retyping of an entire page is given at the end of the monograph. Inserted pages received decimal page numbers: e.g. 74.1, 74.2. A list of additional references was added, and new discussions were kindly contributed by M. J. Bayarri and M. H. DeGroot, Bruce Hill, and Lucien Le Cam.

Substantial changes or additions were made in Sections 3.1, 3.5, 4.2.1, 4.4, and 4.5. The changes in Section 4.4 correct a glaring oversight in the first edition: the failure to emphasize the misleading conclusions that can result from violation of the Likelihood Principle in significance testing of a precise hypothesis. Another very weak part of the first edition was Section 3.5, which discussed prediction, design, and nuisance parameters. The new material incorporates recent substantive insights from the literature.

Numerous other minor changes and literature updatings were made throughout the monograph. We did not attempt complete coverage of recent literature, however.

We are grateful to a number of people for valuable discussions on this subject and/or for comments and suggestions on original drafts or the first edition of the monograph. In particular, we would like to thank George Barnard, M. J. Bayarri, Mark Berliner, Lawrence Brown, George Casella, Morris DeGroot, J. L. Foulley, Leon Gleser, Prem Goel, Clyde Hardin, Bruce Hill, Jiunn Hwang, Rajeev Karandikar, Lucien Le Cam, Ker-Chau Li, Dennis Lindley, George McCabe, Georges Monette, John Pratt, Don Rubin, Herman Rubin, Myra Samuels, Steve Samuels, and Tom Sellke. We are especially grateful to M. J. Bayarri and M. H. DeGroot for an exceptionally complete and insightful set of corrections and comments on the first edition. We are also grateful to Shanti Gupta for the encouragement to turn the material into a monograph.

Thanks are also due to the Alfred P. Sloan Foundation, the National Science Foundation (Grants MCS-7801737, MCS-8101670A1, and DMS-8702620), and the Center for Stochatic Processes at the University of North Carolina for support of the research in the monograph. Finally, we are extremely grateful to Norma Lucas, Teena Chase, and Betty Gick for excellent typing of the manuscript.

March, 1988

JAMES O. BERGER
Purdue University, West Lafayette

ROBERT WOLPERT
Duke University, Durham

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

Among all prescriptions for statistical behavior, the Likelihood Principle (LP) stands out as the simplest and yet most farreaching. It essentially states that all evidence, which is obtained from an experiment, about an unknown quantity θ, is contained in the likelihood function of θ for the given data. The implications of this are profound, since most non-Bayesian approaches to statistics and indeed most standard statistical measures of evidence (such as coverage probability, error probabilities, significance level, frequentist risk, etc.) are then contraindicated.

The LP was always implicit in the Bayesian approach to statistics, but its development as a separate statistical principle was due in large part to ideas of R. A. Fisher and G. Barnard (see Section 3.2 for references). It received major notice when Birnbaum (1962a) showed it to be a consequence of the more commonly trusted Sufficiency Principle (that a sufficient statistic summarizes the evidence from an experiment) and Conditionality Principle (that experiments not actually performed should be irrelevant to conclusions). Since then the LP has been extensively debated by statisticians interested in foundations, but has been ignored by most statisticians. There are perhaps several reasons for this. First, the consequences of the LP seem so absurd to many classical statisticians that they feel it a waste of time to even study the issue. Second, a cursory investigation of the LP reveals certain oft-stated objections, foremost of which is the apparent dependence of the principle on assuming exact knowledge of the (parametric) model for the experiment (so that an exact likelihood function exists). Since the model is rarely true, (hasty) rejection of the LP may result. Third, the LP does not say how one is to perform a statistical analysis; it merely gives a principle to which any method of analysis should adhere. Indeed Bayesian analysis is often presented as the way to implement the LP (with which we essentially agree), a very unappealing prospect to many classical statisticians.

The major purpose of this (mostly review) monograph is to address these concerns. A serious effort will be made, through examples and appeals to common sense, to argue that the LP is intuitively sensible, more so than the classical measures which it impunes. Also, a generalized version of the LP will be introduced, a version which removes the restriction of an exactly known likelihood function, and yet has essentially the same implications. (Other criticisms of the LP will also be discussed.) Finally, the question of implementation of the LP will be considered, and it will be argued that Bayesian analysis (more precisely robust Bayesian analysis) is the most sensible and realistic method of implementation. A thorough discussion of this issue is, however, outside the scope of the monograph, so the main thesis will simply be that the LP is believable and that behavior in violation of it should be avoided to the extent possible.

Acceptance of such a thesis radically alters the way one views statistics. Indeed, to many Bayesians, belief in the LP is the big difference between Bayesians and frequentists, not the desire to involve prior information. Thus Savage said (in the Discussion of Birnbaum (1962a))

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

Many Bayesians became Bayesians only because the LP left them little choice.

Sufficient time has passed since the axiomatic development of Birnbaum to hope that any valid objections to the LP would by now have been found. Indeed, there are numerous articles in the literature presenting examples, counterexamples, arguments, and counterarguments for the LP. We will attempt to discuss all major issues raised, and thus will necessarily cover much of the same ground as these other articles. The collection of relevant arguments in one place will hopefully make study of this crucial issue much easier.

Clearly, we cannot claim impartiality in this monograph; indeed the monograph is essentially aimed at promoting the LP. This can best be done, however, by purposely raising and answering all objections to it (of which we are aware), so a substantial accounting of the "other side" will be given. Also, although our criticism of classical modes of thought may seem rather severe at times, it would be wrong to conclude that we are completely rejecting classical statistics, as it is practiced. Most classical procedures work very well much of the time. Indeed, many classical procedures are exactly what an "objective conditionalist" would use, although for different reasons and with different interpretations. There are exceptions (e.g. significance testing and much of sequential analysis - see Chapter 4), where it can be argued that classical analyses often yield very misleading inferences because of their violation of the LP.

Of course, classical statisticians do (in practice) condition all the time; whenever an experimental protocol is altered or a look at the data reveals the necessity to alter the hypothesized model, conditioning has taken place. (Conditioning followed by the use of unconditional frequentist evaluations is, however, highly suspect, and is the source of much of the hostility towards the LP.) Conditioning seems unavoidable in practice, and so it is a wonderful practical implication of the LP that such conditioning is not only legitimate, but is proper, *providing* a suitable conditional analysis is then performed. Clinical trials is just one area where very desirable simplicity in experimentation and analysis results from adoption of the conditional viewpoint. Discussion of such practical implications is given in Chapter 4.

The mathematics and theoretical statistics used in the monograph will, for the most part, be kept at an easy-to-read level. (The exception is Section 3.4, where the general LP is developed.) Also, examples will frequently be given in simple artificial settings, rather than realistically complicated statistical situations, again for ease of reading and because complicated situations are often too involved to clearly reveal key issues. Advancement of a subject usually proceeds by applying to complicated situations truths discovered in simple settings.

Throughout the monograph, X will denote the random quantity to be observed, $\mathcal{X}$ the sample space, x (the observed data) a realization of X, and $P_\theta(\cdot)$ the probability distribution of X on $\mathcal{X}$, where $\theta \in \Theta$ is unknown. Although θ will be called the *parameter* and Θ the *parameter space*, the family $\{P_\theta(\cdot), \theta \in \Theta\}$ need not be a typical parametric family; θ could just denote some (possibly nonparametric) index. Also, θ will be understood to consist of *all* unknown features of the probability distribution. Often, therefore, only part of θ will be of interest, the remainder being a nuisance "parameter." In discussing sequential and prediction problems it will sometimes be convenient to consider unobserved random variables Z, as well as the unknown θ; z will then denote a possible value of Z. To simplify the exposition in the monograph, however, we will usually only consider the simpler case in which Z is absent. Note that for some statistical problems it is impossible to separate Z and $\{P_\theta(\cdot)\}$. See Section 3.5 for discussion of such problems.

When necessary, $\mathfrak{F}$ will denote the σ-field of measurable events in $\mathcal{X}$. If a density for X exists it will be denoted $f_\theta(x)$, and we will presume the existence of a single dominating σ-finite measure $\nu(\cdot)$ for $\{P_\theta(\cdot), \theta \in \Theta\}$ such that $P_\theta(B) = \int_B f_\theta(x)\, \nu(dx)$ for each $B \in \mathfrak{F}$. In all the examples ν will be taken to be counting measure in the discrete case and Lebesgue measure in the continuous case, when $\mathcal{X}$ is a subset of Euclidean space. Usually we will write the reference measure simply as "dx" (implicitly taking Lebesgue measure for ν); the formulas will require minor changes for cases (including those involving discrete distributions) in which other reference measures are more convenient.

Chapter 2. CONDITIONING

The most commonly used measures of accuracy of evidence in statistics are *pre-experimental*. A particular procedure is decided upon for use, and the accuracy of the evidence from an experiment is identified with the long run behavior of the procedure, were the experiment repeatedly performed. This long run behavior is evaluated by averaging the performance of the procedure over the sample space $\mathcal{X}$. In contrast, the LP states that *post-experimental* reasoning should be used, wherein only the actual observation x (and not the other observations in $\mathcal{X}$ that could have occured) is relevant. There are a variety of intermediate positions which call for partial conditioning on x and partial long run frequency interpretations. Partly for historical purposes, and partly to indicate that the case for at least some sort of conditioning is compelling, we discuss in this chapter various conditioning viewpoints.

2.1 SIMPLE EXAMPLES

The following simple examples reveal the necessity of at least sometimes thinking conditionally, and will be important later.

EXAMPLE 1. Suppose X_1 and X_2 are independent and

$$P_\theta(X_i = \theta-1) = P_\theta(X_i = \theta+1) = \frac{1}{2},\ i = 1,2.$$

Here $-\infty < \theta < \infty$ is an unknown parameter to be estimated from X_1 and X_2. It is easy to see that a 75% confidence set of smallest size for θ is

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

Thus, if repeatedly used in this problem, $C(X_1,X_2)$ would contain θ with probability .75.

Notice, however, that when $x_1 \neq x_2$ it is *absolutely certain* that $\theta = \frac{1}{2}(x_1+x_2)$, while when $x_1 = x_2$ it is equally uncertain whether $\theta = x_1-1$ or $\theta = x_1+1$ (assuming no prior knowledge about θ). Thus, from a post-experimental viewpoint, one would say that $C(x_1,x_2)$ contains θ with "confidence" 100% when $x_1 \neq x_2$, but only with "confidence" 50% when $x_1 = x_2$. Common sense certainly supports the post-experimental view here. It is technically correct to call $C(X_1,X_2)$ a 75% confidence set, but, if after seeing the data we know whether it is really a 100% or 50% set, reporting 75% seems rather silly.

The above example focuses the issue somewhat: does it make sense to report a pre-experimental measure when it is known to be misleading after seeing the data? The next example also seems intuitively clear, yet is the key to all that follows.

EXAMPLE 2. Suppose a substance to be analyzed can be sent either to a laboratory in New York or a laboratory in California. The two labs seem equally good, so a fair coin is flipped to choose between them, with "heads" denoting that the lab in New York will be chosen. The coin is flipped and comes up tails, so the California lab is used. After awhile, the experimental results come back and a conclusion must be reached. Should this conclusion take into account the fact that the coin could have been heads, and hence that the experiment in New York might have been performed instead?

This, of course, is a variant of the famous Cox example (Cox (1958)- see also Cornfield (1969)), which concerns being given (at random) either an accurate or an inaccurate measuring instrument (and knowing which was given). Should the conclusion reached by experimentation depend only on the instrument actually used, or should it take into account that the other instrument might have been obtained?

In symbolic form, we can phrase this example as a "mixed experiment" in which with probabilities $\frac{1}{2}$ (independent of θ) either experiment E_1 or experiment E_2 (both pertaining to θ) will be performed. Should the analysis depend only on the experiment actually performed, or should the possibility of having done the other experiment be taken into account?

The obvious intuitive answer to the questions in the above example is that only the experiment actually performed should matter. But this is counter to pre-experimental frequentist reasoning, which says that one should average over all possible outcomes (here, including the coin flip). One could argue that it is correct to condition on the coin flip, and then use the frequentist measures for the experiment actually performed, but the LP disallows this and is (surprisingly) derivable simply from conditioning on the coin flip plus sufficiency (see Chapter 3).

EXAMPLE 3. For a testing example, suppose it is desired to test H_0: $\theta = -1$ versus H_a: $\theta = 1$, based on $X \sim \mathcal{N}(\theta,.25)$. The rejection region $X \geq 0$ gives a test with error probabilities (type I and type II) of .0228. If $x = 0$ is observed, it is then permissible to state that H_0 is rejected, and that the error probability is $\alpha = .0228$. Common sense, however, indicates that the data $x = 0$ fails to discriminate at all between H_0 and H_a. Any sensible person would be equally uncertain as to the truth of H_0 or H_a (based just on the data $x = 0$). Suppose on the other hand, that $x = 1$ is observed. Then (pre-experimentally) one can still only reject at $\alpha = .0228$, but $x = 1$ is four standard deviations from $\theta = -1$, so the evidence against H_0 seems overwhelming.

Clearly, the actual intuitive evidence conveyed by x can be quite different from the pre-experimental evidence. This has led many frequentists to prefer the use of P-values to fixed error probabilities. The P-value (against H_0) would here be $P_{\theta=-1}(X \geq x)$, a measure of evidence against H_0 with much more dependence on the actual observation, x, than mere rejection at $\alpha = .0228$. (Even P-values can be criticized from a conditional viewpoint, however - see Section 4.4.)

Note that there is nothing logically wrong with reporting error probabilities in Example 3; it just seems to be an inadequate reflection of the evidence conveyed by the data to report α = .0228 for *both* x = 0 and x = 1. Pratt (1977) (perhaps somewhat tongue-in-cheek) thus coins

THE PRINCIPLE OF ADEQUACY. A concept of statistical evidence is (very) inadequate if it does not distinguish evidence of (very) different strengths.

EXAMPLE 4a. Suppose X is 1, 2, or 3 and θ is 1 or 2, with $P_\theta(x)$ given in the following table:

	X 1	2	3
P_0	.009	.001	.99
P_1	.001	.989	.01

.

The test, which accepts P_0 when x = 3 and accepts P_1 otherwise, is a most powerful test with *both* error probabilities equal to .01. Hence, it would be valid to make the frequentist statement, upon observing x = 1, "My test has rejected P_0 and the error probability is .01." This seems very misleading, since the likelihood ratio is actually 9 to 1 in *favor* of P_0, which is being *rejected.*

EXAMPLE 4b. One could object in Example 4a, that the .01 level test is inappropriate, and that one should use the .001 level test, which rejects only when x = 2. Consider, however, the following slightly changed version:

	X 1	2	3
P_0	.005	.005	.99
P_1	.0051	.9849	.01

.

Again the test which rejects P_0 when x = 1 or 2 and accepts otherwise has error probabilities equal to .01, and now it indeed seems sensible to take the indicated actions (if we suppose an action *must* be taken). It still seems unreasonable, however, to report an error probability of .01 upon rejecting P_0 when x = 1, since the data provides very little evidence in favor of P_1.

EXAMPLE 5. For a decision theoretic example, consider the interesting Stein phenomenon, concerned with estimation of a p-variate normal mean ($p \geq 3$) based on $X \sim \mathcal{N}_p(\theta, I)$ and under sum of squares error loss. The usual pre-experimental measure of the performance of an estimator δ is the risk function (or expected loss)

$$R(\theta,\delta) = E_\theta \sum_{i=1}^{p} (\theta_i - \delta_i(X))^2.$$

The classical estimator here is $\delta^0(x) = x$, but James and Stein (1960) showed that

$$\delta^{J-S}(x) = (1 - \frac{p-2}{\Sigma x_i^2})x$$

has $R(\theta,\delta^{J-S}) < R(\theta,\delta^0) = p$ for all θ. One can thus report δ^{J-S} as always being better than δ^0 from a pre-experimental viewpoint. However, if p = 3 and x = (0,.01,.01) is observed, then

$$\delta^{J-S}(x) = (0,-49.99,-49.99),$$

which is an absurd estimate of θ. Hence δ^{J-S} can be terrible for certain x. Of course the positive part version of δ^{J-S},

$$\delta^{J-S+}(x) = (1 - \frac{p-2}{\Sigma x_i^2})^+ x,$$

corrects this glaring problem, but the point is that a procedure which looks great pre-experimentally could be terrible for particular x, and it may not always be so obvious when this is the case.

Confidence sets for θ can also be developed (see Casella and Hwang (1982)) which have larger probabilities of coverage than the classical confidence ellipsoids, are never larger in size, and for small $|x|$ consist of the single point {0}. Indeed, these sets are of the simple form

$$C(x) = \begin{cases} \{\theta: \ |\theta-\delta^{J-S+}(x)|^2 \le \chi_p^2(1-\alpha)\} & \text{if } |x| > \varepsilon \\ \\ \{0\} & \text{if } |x| < \varepsilon, \end{cases}$$

where $\chi_p^2(1-\alpha)$ is the 1-α*th* percentile of the chi-square distribution with p degrees of freedom, and ε is suitably small. Although this confidence procedure looks great pre-experimentally, one would look rather foolish to conclude when p = 3 and x = (0,.01,.01) that θ is the point {0} with confidence 95%.

The above examples, though simple, indicate most of the intuitive reasons for conditioning. There are a wide variety of other such examples. The Uniform $(\theta-\alpha,\theta+\beta)$ distribution (α,β known) provides a host of examples where conditional reasoning differs considerably from pre-experimental reasoning (c.f. Welch (1939) and Pratt (1961)). The Stein 2-stage procedure for obtaining a confidence interval of fixed width for the mean of a $\mathcal{N}(\theta,\sigma^2)$ distribution is another example. A preliminary sample allows estimation of σ^2, from which it is possible to determine the sample size needed for a second sample in order to guarantee an overall probability of coverage for a fixed width interval. But what if the second sample indicates that the preliminary estimate of σ^2 was woefully low? Then one would really have much less *real* confidence in the proposed interval (c.f. Lindley (1958) and Savage et. al. (1962)). Another example is regression on random covariates. It is common practice to perform the analysis conditionally on the observed values of the covariates, rather than giving confidence statements, etc., valid in an average sense over all covariates that could have been observed. Robinson (1975) also gives extremely compelling (though artificial) examples of the need to condition. Piccinato (1981) gives some interesting decision-theoretic examples.

A final important example is that of robust estimation. A convincing case can be made that inference statements should be made conditionally on the residuals; if the data looks completely like normal data, use normal theory. Barnard (1981) says

> "We should recognise that 'robustness' of inference is a conditional property - some inferences from some samples are robust. But other inferences, or the same inferences from other samples, may depend strongly on distributional assumptions."

Dempster (1975) contains very convincing discussion and a host of interesting examples concerning this issue. Related to conditional robustness is large sample inference, which should often be done conditionally on shape features of the likelihood function. Thus, in using asymptotic normal theory for the maximum likelihood estimator, $\hat{\theta}$, one should generally use $I(\hat{\theta})^{-1}$, the inverse of *observed* Fisher information, as the covariance matrix, rather than $I(\theta)^{-1}$, the inverse of *expected* Fisher information. For extensive discussion of these and related issues see Jeffreys (1961), Pratt (1965), Andersen (1970), Efron and Hinkley (1978), Barndorff-Nielsen (1980), Cox (1980), and Hinkley (1980a,1982).

2.2 RELEVANT SUBSETS

Fisher (c.f. Fisher(1956a)) long advocated conditioning on what he called *relevant subsets* of $\mathcal{X}$ (also called "recognizable subsets", "reference sets", or "conditional experimental frames of reference"). There is a considerable literature on the subject, which tends to be more formal than the intuitive type of reasoning presented in the examples of Section 2.1. The basic idea is to find subsets of $\mathcal{X}$ (often determined by statistics) which, when conditioned upon, change the pre-experimental measure. In Example 1, for instance,

$$\mathcal{X} = \{x: x_1 = x_2\} \cup \{x: x_1 \neq x_2\},$$

and the coverage probabilities of $C(X_1,X_2)$ conditioned on observing X in the "relevant" subsets $\{x: x_1 = x_2\}$ or $\{x: x_1 \neq x_2\}$ are 1 and .5, respectively. In Example 2, the two outcomes of the coin flip determine two relevant subsets. In Examples 3, 4, and 5 it is not clear what subsets should be considered relevant, but many reasonable choices give conditional results quite different from the pre-experimental results.

Formal theories of relevant subsets (c.f. Buehler (1959)) proceed in a fashion analogous to the following. Suppose C(x) is a confidence procedure with confidence coefficient 1-α for all θ, i.e.,

(2.2.1) $$P_\theta(C(X) \text{ contains } \theta) = 1-\alpha \quad \text{for all } \theta.$$

Then B is called a relevant subset of $\mathcal{X}$ if, for some $\varepsilon > 0$, either

(2.2.2) $$P_\theta(C(X) \text{ contains } \theta | X \in B) \leq (1-\alpha) - \varepsilon \quad \text{for all } \theta$$

or

(2.2.3) $$P_\theta(C(X) \text{ contains } \theta | X \in B) \geq (1-\alpha) + \varepsilon \quad \text{for all } \theta.$$

When (2.2.2) or (2.2.3) holds and $x \in B$ is observed, it is questionable whether (2.2.1) should be the measure of evidence reported. This formed the basis of Fisher's objection (Fisher (1956b)) to the Aspin-Welch (1949) solution to the Behrens-Fisher problem (see also Yates (1964) and Cornfield (1969)). Another example follows. (For more examples, see Cornfield (1969), Olshen (1977), and Fraser (1977).)

EXAMPLE 6. (Brown (1967), with earlier related examples by Stein (1961) and Buehler and Fedderson (1963)). If $X_1,\ldots,X_n$ is a sample from a $\mathcal{N}(\theta,\sigma^2)$ distribution, both θ and σ^2 unknown, the usual 100(1-α)% confidence interval for θ is

$$C(\bar{x},s) = (\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}),$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively, and $t_{\alpha/2}$ is the appropriate critical value for the t-distribution with n-1 degrees of freedom. For n = 2 and α = .5 we thus have

$$P_{\theta,\sigma^2}(C(\bar{X},S) \text{ contains } \theta) = .5 \text{ for all } \theta,\sigma^2,$$

but Brown (1967) showed that

$$P_{\theta,\sigma^2}(C(\bar{X},S) \text{ contains } \theta \,\Big|\, |\bar{X}|/S \leq 1 + \sqrt{2}) \geq \frac{2}{3} \quad \text{for all } \theta,\sigma^2,$$

and hence the set

$$B = \{(x_1,\ldots,x_n): \ |\bar{x}|/s \leq 1 + \sqrt{2}\}$$

forms a relevant subset.

There is a considerable literature concerning the establishment of conditions under which relevant subsets do or do not exist (c.f. Buehler (1959), Wallace (1959), Stein (1961), Pierce (1973), Bondar (1977), Robinson (1976, 1979a, 1979b), and Pedersen (1978)). Though interesting, a study of these issues would take us too far afield. (See Section 3.7.3 for some related material, however.) Also, much of the theory is still based on frequentist (though partly conditional) measures, and hence violates the LP. Of course, many researchers in the field study the issue solely to point out inadequacies in the frequentist viewpoint, and not to recommend specific conditional frequentist measures. Indeed, it is fairly clear that the existence of relevant subsets, such as in Example 6, is not necessarily a problem, since when viewed completely conditionally (say from a Bayesian viewpoint conditioned on the data $(\bar{x},s)$), the interval $C(\bar{x},s)$ is very reasonable. Thus the existence of relevant subsets mainly points to a need to think carefully about conditioning.

2.3 ANCILLARITY

The most common type of partial conditioning advocated in statistics is conditioning on an ancillary statistic. An *ancillary statistic*, as introduced by Fisher (see Fisher (1956a) for discussion and earlier references), is a statistic whose distribution is independent of θ. (For a definition when nuisance parameters are present, see Section 3.5.5.) Thus, in Example 1, $S = |X_1-X_2|$ is an ancillary statistic which, when conditioned upon, gives "conditional confidence" for $C(X)$ of 100% or 50% as s is 1 or 0, respectively. And, in Example 2, the outcome of the coin flip is an ancillary statistic. The following is a more interesting example.

EXAMPLE 7. Suppose $X_1,\ldots,X_n$ are i.i.d. Uniform $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Then $T = (U,V) = (\min X_i, \max X_i)$ is a sufficient statistic, and $S = V-U$ is an ancillary statistic (having a distribution clearly independent of θ). The conditional distribution of T given S = s is uniform on the set

$$\mathcal{X}_s = \{(u,v): \ v-u = s \text{ and } \theta - \frac{1}{2} < u < \theta + \frac{1}{2} - s\}.$$

Inference with respect to this conditional distribution is straightforward. For instance, a 100(1-α)% (conditional) confidence interval for θ is

$$C(U,V) = \frac{1}{2}(U+V) \pm \frac{1}{2}(1-\alpha)(1-s),$$

one of the solutions proposed by Welch (1939). This conditional interval is considerably more appealing than various "optimal" nonconditional intervals, as discussed in Pratt (1961).

There are a number of difficulties in the definition and use of ancillary statistics (c.f. Basu (1964) and Cox (1971)). Nevertheless, conditioning on ancillaries goes a long way towards providing better conditional procedures. A few references, from which others can be obtained, are Fisher (1956a), Anderson (1973), Barnard (1974), Cox and Hinkley (1974), Cox (1975), Dawid (1975, 1981), Efron and Hinkley (1978), Barndorff-Nielsen (1978, 1980), Hinkley (1978, 1980), Seidenfeld (1979), Grambsch (1980), Amari (1982), Barnett (1982), and Buehler (1982).

2.4 CONDITIONAL FREQUENTIST PROCEDURES

An ambitious attempt to formalize conditioning within a frequentist framework was undertaken by Kiefer (1977). (See also Kiefer (1975, 1976), Brown (1977), Brownie and Kiefer (1977), and Berger (1984c, 1984d).) The formalization was in two distinct directions, which Kiefer called conditional confidence and estimated confidence.

2.4.1 Conditional Confidence

The basic idea of conditional confidence is to define a partition $\{\mathcal{X}_s: \ s \in \mathcal{S}\}$ of $\mathcal{X}$ (the sets in the partition are the relevant subsets of $\mathcal{X}$), and then associate with each set in the partition the appropriate conditional frequency measure for the procedure considered. In Example 1, the partition would be into the sets $\mathcal{X}_1 = \{x: \ x_1 = x_2\}$ and $\mathcal{X}_2 = \{x: \ x_1 \neq x_2\}$. In

Example 2, the partition would be into the sets where heads and tails are observed, respectively.

When dealing with a confidence procedure $\{C(X)\}$, the conditional frequency measure that would be reported, if $x \in \mathcal{X}_s$ were observed, is

$$\Gamma_s(\theta) = P_\theta(C(X) \text{ contains } \theta | X \in \mathcal{X}_s).$$

EXAMPLE 7 (continued). Let the partition be $\{\mathcal{X}_s: \ 0 \leq s \leq 1\}$ (see Example 7). Then, for the procedure $\{C(U,V)\}$,

$$\Gamma_s(\theta) = P_\theta(C(U,V) \text{ contains } \theta | (U,V) \in \mathcal{X}_s) \equiv 1-\alpha.$$

In Examples 1, 2, and 7 it is relatively clear what to condition on. In Examples 3, 4, 5, and 6, however, there is no clear choice of a partition. In a situation such as Example 3, the following choice is attractive.

EXAMPLE 3 (continued). Let $\mathcal{X}_s = \{-s,s\}$ (i.e., the two points s and -s) for $s > 0$. (We will ignore $x = 0$, since it has zero probability of occurring.) The "natural" measure of conditional confidence in a testing situation is the conditional error probability function, determined here by

(2.4.1)
$$\begin{aligned}\Gamma_s(1) = \Gamma_s(-1) &= P_{-1}(\text{Rejecting}|\mathcal{X}_s)\\ &= \frac{P_{-1}(X=s)}{P_{-1}(X=s)+P_{-1}(X=-s)}\\ &= 1/(1+e^{4s}).\end{aligned}$$

One would thus report the test outcome along with the conditional error probability $(1+e^{4|x|})^{-1}$. This conditional error probability has the appealing property of being close to 1/2 if $|x|$ is near zero, while being very small if $|x|$ is large. Thus it satisfies Pratt's "Principle of Adequacy."

The reason (from a frequency viewpoint) for formally introducing a partition is to prevent such "abuses" as conditioning on "favorable" relevant subsets, but ignoring unfavorable ones and presenting the unconditional measure when x is in an unfavorable relevant subset.

2.4.2 Estimated Confidence

An alternative approach to conditioning, which can be justified from a frequentist perspective (c.f. Kiefer (1977) or Berger (1984c)), is to present a data dependent confidence function. If a confidence set procedure C(x) is to be used, for instance, one could report $1-\alpha(x)$ as the "confidence" in C(x) when x is observed. Providing

$$(2.4.2) \qquad E_\theta(1-\alpha(X)) \leq P_\theta(C(X) \text{ contains } \theta) \quad \text{for all } \theta,$$

this "report" has the usual frequentist validity that, in repeated use, C(X) will contain θ with at least the average of the reported confidences. Thus, in Example 2, one could report $1-\alpha(x) = 1$ or $\frac{1}{2}$ as $x_1 \neq x_2$ or $x_1 = x_2$, respectively. Estimated confidence (or, more generally, estimated risk) can be very useful in a number of situations where conditional confidence fails (see Kiefer (1977) or Berger (1984c)).

2.5 CRITICISMS OF PARTIAL CONDITIONING

The need to at least sometimes condition seems to be well recognized, as the brief review in this chapter has indicated. The approaches discussed in Sections 2.2, 2.3, and 2.4.1 consider only partial conditioning, however; one still does a frequency analysis, but with the conditional distribution of X on a subset. There are several major criticisms of such partial conditioning. (The estimated confidence approach in Section 2.4.2 has a quite different basis; criticism of it will be given at the end of this section.)

First, the choice of a relevant subset or an ancillary statistic or a partition $\{\mathcal{X}_s: s \in \mathcal{S}\}$ can be very uncertain. Indeed, it seems fairly clear that it is hard to argue philosophically that one should condition on a certain set or partition, but not on a subset or subpartition. (After all, it seems somewhat strange to observe x, note that it is in, say, $\mathcal{X}_s$, and then forget about x and pretend only that $\mathcal{X}_s$ is known to have obtained.) Researchers working with ancillarity attempt to define "good" ancillary statistics to condition upon, but, as mentioned earlier, there appear to be no completely satisfactory definitions. Also, ancillary statistics do not exist in many situations where it seems important to condition, as the following simple example shows.

EXAMPLE 8. Suppose $\Theta = [0, \frac{1}{2})$, and

$$X = \begin{cases} \theta & \text{with probability } 1-\theta \\ 0 & \text{with probability } \theta. \end{cases}$$

(An instrument measures θ exactly, but will erroneously give a zero reading with probability equal to θ.) Consider the confidence procedure $C(x) = \{x\}$ (the point x). Here $P_\theta(C(X) \text{ contains } \theta) = 1-\theta$. It is clear, however, that one wants to condition on $\{x: \; x > 0\}$, since $C(x) = \{\theta\}$ for sure if $x > 0$. But there is no ancillary statistic which provides such a conditioning.

In situations such as Examples 3, 4, 5, and 6, the selection of a partition for a conditional confidence analysis seems quite arbitrary. Kiefer (1977) simply says that the choice of a partition must ultimately be left to the user, although he does give certain guidelines. The development of intuition or theory for the choice of a partition seems very hard, however (see also Kiefer (1976), Brown (1977), and Berger (1984c)).

Even more disturbing are examples, such as Example 4(b), where it seems impossible to perform the indicated sensible test and report conditional error probabilities reflecting the true uncertainty when x = 1 is observed. (A three point $\mathcal{X}$ cannot be partitioned into two nondegenerate sets, and on a degenerate set the conditional error probability must be zero or one.) Any theory which cannot handle such a simple example is certainly suspect.

The situation for estimated confidence theory is more ambiguous, because it has not been very extensively studied. In particular, the choice of a particular estimated confidence or risk is very difficult, in all but the simplest situations. And, in situations such as Examples 3 and 4(b), estimated confidence functions will have certain undesirable properties. In Example 3, for instance, any estimated error probability, $\alpha(x)$, which is decreasing in $|x|$ and satisfies the frequentist validity criterion (similar to (2.4.2))

$$E_\theta\ \alpha(X) \geq P_\theta(\text{Test is in error}) \qquad \text{for all } \theta,$$

must have $\alpha(0) > \frac{1}{2}$ (since $P_{\frac{1}{2}}$ (Test is in error) $= \frac{1}{2}$). It seems strange, however, to report an error larger than $\frac{1}{2}$ (which could, intuitively, be obtained by simple guessing). For more extensive discussion of estimated confidence, see Berger (1984c).

The final argument against partial conditioning is the already alluded to fact that the most clearcut and "obvious" form of conditioning (Example 2) implies (together with sufficiency) the LP, which states that complete conditioning (down to x itself) is necessary. Since this would eliminate the possible application of frequency measures, new measures of evidence would clearly be called for.

It should be mentioned that certain other forms of statistical inference are very conditional in nature, such as fiducial inference developed by Fisher (see Hacking (1965), Plackett (1966), Wilkinson (1977), Pedersen (1978), and Dawid and Stone (1982) for theory and criticisms), structural inference developed by Fraser (c.f. Fraser (1968, 1972, 1979)), and pivotal inference developed by Barnard (c.f. Barnard (1980, 1981) and Barnard and Sprott (1983)). (Barnett (1982) gives a good introduction to all of these approaches.) The similarities among these methods (and also "objective Bayesian" analysis and frequentist "invariance" analysis) are considerable, but the motivations can be quite different. These methods rarely result in unreasonable conclusions from a conditional viewpoint, and hence do have many useful implications for conditional analysis. Space precludes extensive discussion of these approaches. (Some discussion of structural and pivotal analysis will be given in Sections 3.6 and 3.7, in the course of answering a specific criticism of the LP.) Suffice it to say that they are based on "intuitive" principles which can be at odds with the LP (and Bayesian analysis), and hence leave us doubting their ultimate truth.

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

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

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

$$P_\theta(X^* = (1,x_1^*)|T = (1,x_1^*)) = 1-P_\theta(X^* = (2,x_2^*)|T = (1,x_1^*))$$

$$= \frac{\frac{1}{2}\cdot f_\theta^1(x_1^*)}{\frac{1}{2}\cdot f_\theta^1(x_1^*)+ \frac{1}{2}f_\theta^2(x_2^*)}$$

$$= c/(1+c),$$

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

(3.3.2) $$Ev(E^*,(1,x_1^*)) = Ev(E^*,(2,x_2^*)).$$

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

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

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

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 $(\Omega, \mathfrak{F}, \{\mu_\theta\})$ 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 $\Omega_x = \{$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*

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

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

$$\mathrm{Ev}(E_1,x_1) = \mathrm{Ev}(E_2,x_2)$$

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

Then $Ev(E_1,x_1) = Ev(E_2, \varphi(x_1))$ *for* $\{P^1_\theta\}$ *- 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* $E_1 = (X_1, \theta, \{P^1_\theta\})$ *and* $E_2 = (X_2, \theta, \{P^2_\theta\})$ *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 $f^i_\theta(x_i) = P^i_\theta(\{x_i\})$. First, assume the validity of the LP, and let

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

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 $f^2_\theta(x_2) = [1/c(x_1)]f^1_\theta(x_1)$ 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 $f^1_\theta(x_1) = cf^2_\theta(x_2)$ 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

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

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 $\{x_i \in \mathcal{X}_i: (i,x_i) \in A\}$. 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 $X_i = (U_i, V_i)$ $(i = 1,2,...)$ 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 $\mathcal{N}(\xi,1)$ 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

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

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

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

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* $\ell_x(\theta)$, $\mathrm{Ev}_\xi(E,x)$ *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,* $\mathrm{Ev}_{y,\xi}(E,x)$ *depends on* E *and* x *only through* $\ell_x(\theta)$ *as defined in (3.5.1).*

With this amendment, the LP says that $\mathrm{Ev}_\xi(E,x)$ (or more generally $\mathrm{Ev}_{y,\xi}(E,x)$) 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

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

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*

(3.5.3) $$\ell^B_x(y,\xi) = \int \ell^*_x(y,\xi,\eta^1)\ \pi(\eta^1|\xi)\ d\eta^1.$$

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 $\ell^1_x(\xi)$ is not always justified. If knowledge of η would communicate information about ξ, then $\ell^2_x(\eta)$ 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 $\ell^2_x(\eta)$ contains no information about ξ. This is clear, since then (3.5.3) becomes (ignoring y)

$$\ell^B_x(\xi) = \int \ell^1_x(\xi)\ \ell^2_x(\eta)\ \pi_2(\eta)\ d\eta \propto \ell^1_x(\xi).$$

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* *noninformative nuisance parameter* *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

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

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 $\mathfrak{Q}$. For simplicity, we assume Q is known.) This is actually more or less the "structural" formulation of the problem. The formulation in Pivotal Inference is based on "pivotals" $\omega = g(X,\theta)$ having known distributions. Typically g will be an appropriate inverse function of h, so the two approaches are very related. We will, for the most part, consider the structural formulation, although comments about differences for the pivotal model will be made. The structural model is sometimes called a functional model (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 $(n, \ddagger)$ 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 $\mathcal{G}$.) 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)

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

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 $P^1_\theta(A) > 0$ 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* __$\mathcal{C}$*-better*__ *than* δ^1.

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

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

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

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

$$(3.7.10) \qquad \delta^1((1,x_1')) \neq \delta^1((2,x_2')).$$

Consider, however, the procedure

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

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

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

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

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

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 $\underset{\sim}{\theta} \in C_C$. (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 $\underset{\sim}{\theta} \in C_{\mathcal{C}}$. 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 $\underset{\sim}{\theta} \in C_{\mathcal{C}}$, 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*,

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

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 $\underset{\sim}{\Sigma}$ (which is equivalent to θ) under the strictly convex loss

$$L(\delta,\underset{\sim}{\Sigma}) = \text{tr}(\delta\underset{\sim}{\Sigma}^{-1})-\log\det(\delta\underset{\sim}{\Sigma}^{-1})-2. \tag{3.7.17}$$

(The loss $L(\delta,\underset{\sim}{\Sigma}) = \text{tr}(\delta\underset{\sim}{\Sigma}^{-1}-I)^2$ 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

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

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

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

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

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

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

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 $\underset{\sim}{\ell}$.

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 contraindicated.

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 $\varepsilon > 0$,

$$P_{\underset{\sim}{\theta}}(\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i > \varepsilon) = 1 \tag{3.7.19}$$

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 $X \sim \mathcal{N}(\theta,1)$, $\Theta = \mathbb{R}^1$, $C(x) = (-\infty, x)$, 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

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

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.

CHAPTER 4. CONSEQUENCES AND CRITICISMS OF THE LIKELIHOOD PRINCIPLE AND RELATIVE LIKELIHOOD PRINCIPLE

Most people who reject the LP do so because it has consequences they do not like. Of course any theory deserves to be rejected if its consequences are erroneous, but great care must be taken in making sure that the consequences really are wrong and not just in opposition to the intuition currently dominant in the field. In this section we discuss some of the more surprising consequences of the LP and RLP, and investigate the conflicts with prevalent statistical intuition. It will come as no surprise that we feel that the conflicts are always resolved in favor of the LP and RLP.

4.1 INCOMPATIBILITY WITH FREQUENTIST CONCEPTS

4.1.1 Introduction

The philosophical incompatibility of the LP and the frequentist viewpoint is clear, since the LP deals only with the observed x, while frequentist analyses involve averages over possible observations. It cannot be said, however, that any particular frequentist procedure conflicts with the LP, since the procedure could happen to correspond to a sensible conditional procedure. Such a correspondence does, in fact, occur in many statistical situations. For instance, much of frequentist normal distribution theory inference provides the same numerical measures of "confidence" as does noninformative prior conditional Bayesian theory (because of the symmetries or group structure of the problem), although the interpretations of these measures are different. (A cynic might argue that frequentist statistics has survived precisely because of such lucky correspondences.) Nevertheless, enough direct conflicts have been (and will be) seen to justify viewing the LP as revolutionary from a frequentist perspective.

We have already alluded to the fact that a frequentist can logically dismiss the LP, essentially by rejecting the WCP and concluding that the concept of learning or drawing conclusions about θ, for a particular experiment, is meaningless. Thus Neyman (c.f. Neyman (1957, 1967, 1977)) espouses the viewpoint that only the performance of a procedure in repeated use is relevant, and that it is a mistake to think in terms of learning about particular θ. Though logically viable, this viewpoint is scientifically unappealing. Experiments are done precisely to obtain "evidence" about unknown θ, and investigators will not take kindly to being told that this is meaningless. Thus Birnbaum (1977) argues that Neyman-Pearson conclusions are virtually always used in an "evidentiary" fashion, rather than as measures of procedure performance in repeated use. Savage put this very succinctly when talking about confidence sets in Savage et. al. (1962):

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

Supposing then that we are going to use a frequency measure as a measure of evidence about θ, what classical justifications for such behavior can be advanced? There are at least the following four:

(i) Frequency measures are "objective", having a well defined physical interpretation, and science demands objective statistical measures.

(ii) The use of frequency measures (and procedures based on them) is reasonably sound and safe for nonspecialists.

(iii) One needs "repeatable" experiments in science, i.e., any evidence gathered about θ should also be likely to be found if the experiment is repeated; this will supposedly be true if frequency measures of evidence are used.

(iv) The following principle should be followed:

CONFIDENCE PRINCIPLE. Any statistician who uses a methodology in which he makes *statements or draws conclusions with specified accuracy, should be guaranteed that in the long run his actual accuracy will be at least that promised.*

We will briefly examine these four justifications.

4.1.2 Objectivity

It should be observed, first of all, that the LP is entirely objective, stating only that the evidence about θ is contained in the likelihood function. Also, the likelihood function has as much physical reality as any frequency measure calculated for a presumed model. It would thus be logically sound to pass on to the next issue. We dally, however, because of the problem of using the likelihood function. Indeed, since in Chapter 5 we will argue for Bayesian use of the likelihood function, issues of objectivity will become relevant.

The Bayesian answers to criticisms of objectivity are either (i) objectivity is a myth, or (ii) only through "noninformative" prior Bayesian analysis can objectivity be really attained. As an example of the first argument, Box (1980) states:

> "In the past, the need for probabilities expressing prior belief has often been thought of, not as a necessity for all scientific inference, but rather as a feature peculiar to Bayesian inference. This seems to come from the curious idea that an outright assumption does not count as a prior belief... I believe that it is impossible logically to distinguish between model assumptions and the prior distribution of the parameters."

A general review of this objectivity issue is given in Berger and Berry (1987). (See also Berger (1985).) The only portion of frequentist theory formally exempt from the argument is (completely) nonparametric analysis, and, even then, the choice of a particular procedure to use can be argued to be a highly subjective input.

If the model can be claimed to have some objective status, there is still argument (ii) (above) to contend with. The idea behind this argument is that one can lay claim to objectivity only by purposely striving for it, through use of what is deemed to be an "objective prior." Substantial support for this position can be found in Jeffreys (1961), Box and Tiao (1973), Zellner (1971), Rosenkranz (1977), Bernardo (1979), Berger (1980,1984e), and Jaynes (1981, 1982). Regardless of the validity of argument (ii), it is a fact that use of noninformative priors is objective, purposely not involving subjective prior opinions, and is consistent with the LP. The measures of evidence used are, of course, probabilistic statements about the unknown θ itself (through the formal posterior distribution of θ) and hence may be deemed less "real", but a very strong case can be made that "evidence" about uncertain quantities should only be quantified probabilistically (c.f. deFinetti (1972, 1974)). There are also other likelihood based methods which can be classified as objective, as will be seen in Chapter 5. Hence, even if deemed obtainable and desirable, objectivity is not a reason to reject the LP in favor of frequency measures.

4.1.3 Procedures for Nonspecialists

We accept the argument that it is important to develop reasonably simple statistical procedures which can be safely used by nonspecialists. However, it is not at all clear that this need be done from a frequency viewpoint. First, frequency methods often attain formal simplicity by obscuring difficult issues, such as the choice of error probabilities in a test or the choice of a partition in a conditional confidence procedure (see Section 2.5). Second, relatively simple procedures and methods of evaluation consistent with the LP can be developed (w/o the introduction of subjective priors) as the books of Jeffreys (1961), Box and Tiao (1973), and Zellner (1971) indicate. We are continually surprised at the ease with which the use of noninformative priors, as in these books, gives excellent (conditional) procedures. Indeed, as mentioned earlier, many reasonable frequentist procedures are, at least approximately, noninformative prior Bayes procedures, and "frequency confidence" then often coincides with "posterior confidence." When this correspondence does not occur, such as in unconditional frequentist approaches to the examples in Section 2.1, the frequentist approach is definitely suspect. Further discussion and references can be found in Berger (1980). Note that we are not maintaining that the use of noninformative priors solves all problems and is foolproof, but only that, if procedures which are simple to use and interpret are deemed necessary, then there are good conditional alternatives to frequentist development of procedures. We have also slighted the subjective Bayes solution to the problem, which will, however, be discussed in Chapter 5.

In this situation, where a procedure is developed for use by nonspecialists, the performance of the procedure in repeated use is certainly relevant (see Section 3.5.4), though not necessarily of primary importance. Good frequency performance can even be of interest to the strict conditionalist, as the following example indicates.

EXAMPLE 16. Suppose a confidence procedure $C(x)$ is to be used (i.e., when $X = x$ is observed it will be stated that $\theta \in C(x)$), having frequentist coverage probability

$$\Gamma(\theta) \equiv P_\theta(C(X) \text{ contains } \theta) \geq 1-\alpha.$$

A conditional Bayesian (for simplicity) would, for a prior distribution π on Θ, be interested in having good posterior probability that θ is in $C(x)$, i.e., would want

$$\lambda(x) \equiv P^{\pi(\theta|x)}(\theta \in C(x))$$

to be large, where $\pi(\theta|x)$ is the posterior probability distribution of θ given x. But, letting m denote the marginal distribution of X (i.e., $m(\cdot) = E^\pi P_\theta(\cdot)$) and $I_B(y)$ denote the usual indicator function on a set B, it is clear that

$$\begin{aligned} E^m\lambda(X) &= E^m P^{\pi(\theta|X)}(\theta \in C(X)) \\ &= E^{\text{joint distbn. }(\theta,X)}[I_{C(X)}(\theta)] \\ &= E^{\pi}P_\theta(C(X) \text{ contains } \theta) \\ &\geq 1-\alpha. \end{aligned}$$

Since this relationship holds regardless of π, a conditionalist could feel that $\lambda(x)$ is "likely" to be large if $C(x)$ is used and α is small, and hence be willing to use $C(x)$ when unable to carry out a trustworthy Bayesian analysis. See Pratt (1965) and Berger (1984b) for more general development and specific examples.

It is important to emphasize that the primary goal in situations such as Example 16 should still be good conditional performance, and that the frequentist measure does not guarantee this. Conceivably, $\lambda(x)$ could be very small for some x (and all m), which is certainly relevant since such x could be observed. Thus our view is that procedures should usually be developed from a conditional viewpoint, and their frequency properties perhaps investigated to ensure robustness. Of course the already existing classical procedures which have good conditional properties are fine. Other discussions of this point can be found in Godambe and Thompson (1977), Godambe (1982a,b), and Berger(1984e).

4.1.4 Repeatability

There is certainly truth to the observation that, if a scientific experiment claims to have obtained strong evidence about θ, then many scientists expect future similar experiments to also provide strong evidence. The frequency measures, based on imagining repetitions of the experiment, seem ideally suited to achieve this. There is a serious concern here, however, as the following example indicates.

EXAMPLE 17. Suppose X has the two point distribution given by $P_\theta(X = 0) = .99$ and $P_\theta(X = \theta) = .01$. (Either θ will be measured exactly, or no observation will be recorded.) If now $x = 5$ is observed, it should certainly be concluded that $\theta = 5$ exactly (very strong "evidence"), but repetitions of the experiment are very unlikely to reproduce the result.

It could perhaps be argued that science should not believe "lucky" observations, as in the previous example, and hence should not think conditionally on the data. This seems too severe a straightjacket, however. One can always be skeptical of lucky observations and seek possible alternative reasons for them, but their conditional evidential interpretation should be allowed. Such conditional interpretations can, of course, also be verified or disproved by future investigations.

4.1.5 The Confidence Principle

The Confidence Principle was implicit in much of Neyman's early development of the frequentist viewpoint (c.f. Neyman (1967) and also Neyman (1957, 1977) and Berger (1984c)), and was stated explicitly by Birnbaum (c.f. Giere (1977) and Birnbaum (1968, 1970, 1977)), who ultimately came to reject the LP because of its conflict with the Confidence Principle. Other discussions of this or related principles can be found in Cox and Hinkley (1974) (which distinguishes between strong and weak versions, the weak version allowing conditioning on relevant subsets), Kiefer (1977b), Le Cam (1977), and Barnard and Godambe (1982). Critical discussion can be found in Jeffreys (1961), Hacking (1965), Edwards (1972), deFinetti (1972, 1974), Pratt (1977), and Jaynes (1981, 1982). The following mathematical formulation of the Confidence Principle will be useful in the discussion, and is related to the Evaluation Game in Section 3.7.2.

THE FORMAL CONFIDENCE PRINCIPLE. A procedure δ *is to be used for a sequence of problems consisting of observing* $X_i \sim P_{\theta_i}$. *A criterion,* $L(\theta_i,\delta(x_i))$, *measures the performance of* δ *in each problem (small* L *being good). One should report, as the "confidence" in use of* δ*,*

$$\bar{R}(\delta) = \sup_{\underset{\sim}{\theta}} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} L(\theta_i,\delta(x_i)), \tag{4.1.1}$$

assuming the limit exists with probability one. (It can usually be shown that $\bar{R}(\delta) = \sup_\theta R(\theta,\delta)$, *where* $R(\theta,\delta) = E_\theta L(\theta,\delta(X))$.)

EXAMPLE 18. Suppose δ is a confidence procedure, so that $\delta(x_i) \subset \Theta$ will be the confidence set when x_i is observed. The natural measure of the performance of $\delta(x_i)$ is

$$L(\theta_i,\delta(x_i)) = 1-I_{\delta(x_i)}(\theta_i),$$

since this measures whether or not $\delta(x_i)$ *does* contain θ_i. The risk of δ is

$$R(\theta,\delta) = E_\theta L(\theta,\delta(X)) = 1-P_\theta(\delta(X) \text{ contains } \theta),$$

and it is easy to show, for this problem, that

$$\bar{R}(\delta) = \sup_\theta R(\theta,\delta) = 1-\inf_\theta P_\theta(\delta(X) \text{ contains } \theta).$$

Hence the "report," according to the Confidence Principle, should be one minus the minimum coverage probability of δ.

Although the Confidence Principle is formulated above only in terms of repetitive use of δ for problems of the same form (but possibly differing θ_i), it can easily be generalized to include use of δ for different types of problems. Such a generalization adds little conceptually, however. The appeal of the Confidence Principle is undeniable. By following it, the *actual* average performance of δ in repeated use will be at least as good as the *reported* performance $\bar{R}(\delta)$. There are several problems in following the Confidence Principle, however.

The first difficulty is that, in virtually all statistical investigations, extensive assumptions concerning the model, etc., are made. Thus a person claiming to err no more than 5% of the time because he follows the Confidence Principle, is really saying he errs no more than 5% of the time if all the model assumptions he makes are correct. This removes some of the lustre from the principle.

A second serious issue is the need to have a valid bound, $\bar{R}(\delta)$, on the performance of δ. This is an often unappreciated aspect of the frequentist position. Indeed, the frequentist position is often viewed as requiring only the reporting of the function $R(\theta,\delta)$. Without the bound, $\bar{R}(\delta)$, however, no guarantee of long run performance, in actual use of δ on different problems, can be given.

EXAMPLE 19. Consider simple versus simple hypothesis testing, and suppose one always uses the most powerful test of level $\alpha = .01$. One can make the frequentist statement that only 1% of true null hypotheses will be rejected (i.e., $R(\theta,\delta) = .01$ for θ equal to the null), but this says nothing about how often one errs when rejecting. For instance, if the test has power of .01 (admittedly terrible power, but useful for making the point) and the null and alternative hypotheses occur equally often in repetitive use of the test, then *half* of all rejections will be in error. Thus one can not make meaningful statements about actual error incurred in repetitive use, without an appropriate bound on $R(\theta,\delta)$ for all θ.

The problem with needing $\bar{R}(\delta)$ is, of course, that it could be a useless bound (or could even be infinite). Indeed, whenever $R(\theta,\delta)$ is highly variable as a function of θ, the reporting of $\bar{R}(\delta)$ is likely to be excessively conservative. The conditional frequentist approaches discussed in Section 2.4 have considerable promise in overcoming this difficulty, however, and can be given interpretations compatible with the Confidence Principle.

Ultimately, the only clear objection to the Confidence Principle is that it conflicts with the LP. This was indicated in the examples and discussion in Chapter 2, and will be seen in later examples also. Most conditionalists view the Confidence Principle, while attractive, as an unattainable goal. (Note, however, that a Bayesian conditionalist follows the Confidence Principle to the extent that his statements of accuracy will be correct, in the long run average sense, if his prior assumptions are correct; one could, indeed, argue that it is the Bayesian who is honestly trying to follow the Confidence Principle by clearly stating the beliefs and assumptions his assessments are based on.) In choosing between the LP and the Confidence Principle, it is important to recall the simple axiomatic basis of the LP, and to realize that no such basis has been found for the Confidence Principle. Indeed, the long run performance view is deemed rather peculiar by most uninitiated people (c.f., the discussions in the early papers of Neyman in Neyman (1967)).

4.2 THE IRRELEVANCE OF STOPPING RULES

4.2.1 Introduction

One of the most important applications of the LP and RLP is the Stopping Rule Principle (SRP). Stated informally, the SRP is simply that the reason for stopping experimentation (the *stopping rule*) should be irrelevant to evidentiary conclusions about θ. The theoretical and practical implications of the SRP to such fields as sequential analysis and clinical trials are enormous, and will be partially discussed in Sections 4.2.3 and 4.2.4. The SRP itself will be discussed at two levels: in Section 4.2.2 it will be presented in a relatively simple sequential setting, in which it will be shown to follow solely from the LP, while in Section 4.2.6 a very general version will be developed from the RLP. Section 4.2.7 discusses situations in which the SRP is *not* applicable, and Section 4.2.5 points out an interesting conflict between frequentist admissibility and the frequentist belief in the importance of considering stopping rules.

The Stopping Rule Principle was first espoused by Barnard (1947a, 1949), whose motivation at the time was essentially a reluctance to allow an experimenter's *intentions* to affect conclusions drawn from data. (More will be said of this shortly.) The principle was shown to be a consequence of the LP in Birnbaum (1962a), and Barnard, Jenkins and Winsten (1962), and argued to hold in essentially complete generality by Pratt (1965). Other good discussions of the principle can be found in Anscombe (1963), Cornfield (1966), Bartholomew (1967), Basu (1975), Berger (1980), and in many Bayesian works such as Edwards, Lindman, and Savage (1963).

Before formally introducing stopping rules and the stopping rule principle, it is useful to illustrate certain of the ideas through a simple example. The following example, from Berger and Berry (1987), demonstrates the possible extreme dependence of frequentist measures upon the intentions of the experimenter concerning stopping the experiment. The example clearly questions the sensibility of such extreme dependence. (Berger and Berry, 1987, also contains other simple examples, on both sides of the issue.)

EXAMPLE 19.1. A scientist enters the statistician's office with 100 observations, assumed to be independent and from a $N(\theta,1)$ distribution. The scientist wants to test H_0: $\theta = 0$ versus H_1: $\theta \neq 0$. The current average is $\bar{x}_n = 0.2$, so the standardized test statistic is $z = \sqrt{n}|\bar{x}_n - 0| = 2$. A careless classical statistician might simply conclude that there is significant evidence against H_0 at the 0.05 level. But a more careful one will ask the scientist, "Why did you cease experimentation after 100 observations?" If the scientist replies, "I just decided to take a batch of 100 observations," there would seem to be no problem, and very few classical statisticians would pursue the issue. But there is another important question that should be asked (from the classical perspective), namely: "What would you have done had the first 100 observations not yielded significance?"

To see the reasons for this question, suppose the scientist replies: "I would then have taken another batch of 100 observations." This reply does not completely specify a stopping rule, but the scientist might agree that he was implicitly considering a procedure of the form:

(a) take 100 observations;

(b) if $\sqrt{100}|\bar{x}_{100}| \geqq k$ then stop and reject H_0;

(c) if $\sqrt{100}|\bar{x}_{100}| < k$ then take another 100 observations and reject if $\sqrt{200}|\bar{x}_{200}| \geqq k$.

For this procedure to have level $\alpha = 0.05$,k must be chosen to be 2.18 (Pocock, 1977). Since the actual data had $\sqrt{100}|\bar{x}_{100}| = 2 < 2.18$, the scientist could not actually conclude significance, and hence would have to take the next 100 observations.

This strikes many people as peculiar. The interpretation of the results of an experiment depends not only on the data obtained and the way it was obtained, but also upon *thoughts* of the experimenter concerning plans for the future.

Of course, this can be carried further. Suppose the puzzled scientist leaves and gets the next 100 observations, and brings them back. Consider two cases. If $\sqrt{200}|\bar{x}_{200}| = 2.1 < 2.18$ then the results are not significant. But they would have been significant had the scientist not paused halfway through the study to calculate z! (It would certainly be tempting not to disclose this interim calculation, and essentially impossible to determine whether or not the scientist had made an interim calculation!) On the other hand, suppose $\sqrt{200}|\bar{x}_{200}| = 2.2 > 2.18$, so now significance has been obtained. But wait! Again the statistician asks what the scientist would have done had the results not been significant. Suppose the scientist says, "If my grant renewal were to be approved, I would then take another 100 observations; if the grant renewal were rejected, I would have no more funds and would have to stop the experiment in any case." The advice of the classical statistician must then be: "We cannot make a conclusion until we find out the outcome of your grant renewal; if it is not renewed, you can claim significant evidence against H_0, while if it is renewed you cannot claim significance and must take another 100 observations." The up-to-now honest scientist has had enough, and he sends in a request to have the grant renewal denied, vowing never again to tell the statistician what he would have done under alternative scenarios.

Note that we are not faulting the classical statistician here for ascertaining and incorporating the stopping rule in the analysis. If one insists on utilization of frequentist measures, such involvement of the stopping rule (even if it exists only in the imagination of the experimenter) is mandatory. The need here for involvement of the stopping rule clearly calls the basic frequentist premise into question, however.

4.2.2 The (Discrete) Stopping Rule Principle

So as not to obscure the essential nature of the SRP, the discussion in this section will be restricted to the following fairly simple situation. Suppose E^{τ} is a sequential experiment consisting of (i) a sequence of independent observations X_1, $X_2, \ldots$, which will be observed one at a time and which have common density $f_{\theta}(x)$; and (ii) a non-randomized *stopping rule,* τ, which can be represented by a sequence of sets,

$$A_m \subset \mathcal{X}^m = \mathcal{X} \times \mathcal{X} \times \ldots \times \mathcal{X} \qquad \text{(the m-fold Cartesian product of } \mathcal{X}\text{)},$$

having the property that

$$(4.2.1) \qquad \text{if } \underset{\sim}{x}^m = (x_1, \ldots, x_m) \in A_m, \quad \text{sampling stops;}$$

$$\text{if } \underset{\sim}{x}^m \in A_m^c, \quad \text{sampling continues.}$$

Since the observations will be observed sequentially, it is clearly unnecessary to have A_m contain points whose first j coordinates were in A_j for any $j < m$; thus we henceforth assume that

$$A_m \cap A_j \times \mathcal{X}^{m-j} = \emptyset \qquad \text{for } j < m.$$

The *stopping time,* N, corresponding to τ is that (random) m for which $\underset{\sim}{x}^m \in A_m$; the realization of N will be denoted by n. As usual, only proper stopping rules will be considered, i.e., those which have N finite with probability one for all θ. The probability density of the random experimental outcome $\underset{\sim}{X}^N = (X_1, \ldots, X_N)$ is then

$$(4.2.2) \qquad f_{\theta}^{\tau}(\underset{\sim}{x}^n) = I_{A_n}(\underset{\sim}{x}^n) \prod_{i=1}^{n} f_{\theta}(x_i).$$

EXAMPLE 20. Suppose the X_i are $\mathcal{N}(\theta, 1)$.

Case 1. Consider the stopping rule, τ^1, defined by

$$A_m^1 = \begin{cases} \emptyset & \text{if } m \neq k \\ \mathcal{X}^k & \text{if } m = k. \end{cases}$$

The experiment E^{τ^1} is thus the fixed sample size experiment which always observes precisely k observations.

Case 2. Consider the stopping rule, τ^2, defined by

$$A_m^2 = \{\underset{\sim}{x}^m \in \mathcal{X}^m: \ |\bar{x}_m| > Km^{-\frac{1}{2}}\}, \tag{4.2.3}$$

where $\bar{x}_m$ is the mean of $(x_1,\ldots,x_m)$ and K is a fixed positive constant. (By using the Law of the Iterated Logarithm, τ^2 can be shown to be a proper stopping rule.) This stopping rule is rather peculiar, in that it says to stop sampling when the sample mean is K standard deviations from zero.

EXAMPLE 21. Suppose the X_i are Bernoulli (θ).

Case 1. Let E^{τ^1} be the fixed sample size experiment which takes k observations, where $k \leq 2$.

Case 2. Let τ^2 be defined by

$$A_1^2 = \{1\}, \ A_2^2 = \{(0,0),(0,1)\}, \ A_j^2 = \emptyset \quad \text{for } j > 2$$

(i.e., stop if $X_1 = 1$, and otherwise stop after observing X_2), and let E^{τ^2} be the corresponding sequential experiment.

STOPPING RULE PRINCIPLE (SRP): In a sequential experiment E^τ, *with observed final data* $\underset{\sim}{x}^n$, $Ev(E^\tau,\underset{\sim}{x}^n)$ *should not depend on the stopping rule* τ.

The SRP would imply, in Example 20, that if the observation in Case 2 happened to have n = k, then the evidentiary content of the data would be the same as if the data had arisen from the fixed sample size experiment in Case 1. A similar conclusion would hold in Example 21 if n = k occurred.

When $\mathcal{X}$ is discrete, the SRP is an immediate consequence of the LP. This is immediate from (4.2.2) in that $\ell_{\underset{\sim}{x}^n}(\theta)$ is proportional to $\prod_{i=1}^{n} f_\theta(x_i)$, which does not depend on the stopping rule. For derivation of the SRP in general (from the RLP) see Section 4.2.6.

4.2.3 Positive Implications

A recurring problem in classical statistics is that of optional stopping. Ideally (from a classical viewpoint) an experimenter chooses his stopping rule before experimentation, and then follows it exactly. Actual practice is, however, acknowledged to be quite different. Experiments may end because the data looks convincing enough, because money runs out, or because the experimenter has a dinner date. Indeed, little or no thought may have been given to the stopping rule prior to experimentation, in which case, upon stopping for whatever reason, the data is often treated as having arisen from a fixed sample size design. Optional stopping may often be harmless (such as when the experimenter quits to have dinner), but stopping "when the data looks good" can be a serious error when combined with frequentist measures of evidence. For instance, if one used the stopping rule in Case 2 of Example 20, but analyzed the data as if a *fixed* sample had been taken, one could *guarantee* arbitrarily strong frequentist "significance" against H_0: $\theta = 0$ by merely choosing large enough K.

Optional stopping poses a significant problem for classical statistics, even when the experimenters are extremely scrupulous. Honest frequentists face the problem of getting extremely convincing data too soon (i.e., before their stopping rule says to stop), and then facing the dilemma of honestly finishing the experiment, even though a waste of time or dangerous to subjects, or of stopping the experiment with the prematurely convincing evidence and then not being able to give frequency measures of evidence. One could argue that experiments should be designed to allow for early stopping in response to clear evidence (and, indeed, many such stopping rules have been created, as in the theory of "repeated significance testing"), but there will often be unforeseen eventualities that crop up in sequential experimentation, leaving a strict frequentist in an embarassing position.

Contrast this enormous dilemma with the startling simplicity resulting from use of the SRP. The SRP says that it just doesn't matter; stop for whatever reasons, which (conditional on the data) do not depend on θ (see

Section 4.2.7), and use an appropriate conditional analysis based on $\ell_{\underset{\sim}{x}^n}(\theta)$ (or, alternatively, $\prod_{i=1}^{n} f_\theta(x_i)$). The reason for stopping is simply not relevant. As Edwards, Lindman, and Savage (1963) say

> "The irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design... Many experimenters would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money, or patience."

Anscombe (1963) simply makes the blunt statement "Sequential analysis is a hoax." These comments should be qualified, of course, to the extent that design will depend on the stopping rule. In other words, choosing between two sequential designs obviously involves consideration of stopping rules. Indeed, the most difficult part of (theoretical) sequential (decision) analysis is that of deciding, at a given stage, whether to stop sampling or to take another observation (i.e., choosing the stopping rule). Much of the work done in classical sequential analysis has addressed this problem, and is hence of considerable relevance.

The other desirable implication of the SRP is that analysis of an experiment can be done objectively, in the sense that it is no longer necessary to know the experimenter's intentions towards stopping. It seems very strange that a frequentist could not analyze a given set of data, such as $(x_1,\ldots,x_n)$ in Example 20,if the stopping rule is not given. If the experimenter forgot to record the stopping rule and then died, it is unappealing to have to guess his stopping rule in order to conduct the analysis. As mentioned earlier, it was apparently this feeling, that data should be able to speak for itself, that led Barnard to first support the Stopping Rule Principle.

The above idea is actually a general consequence of the LP, and is useful to apply in areas other than optional stopping. Consider the following example.

EXAMPLE 22. An experiment was conducted with two treatment groups (T_1 and T_2) and a control group (C), the outcomes for each experimental unit being simply success (S) or failure (F). The data was

	C	T_1	T_2
S	8	12	2
F	12	8	8

.

In analyzing the results, the experimenter noted that, in comparing T_1 with C, a standard analysis under the null hypothesis of no treatment effect was not significant at level α = .1 (one-tailed), but that if the patients in T_2 and C were pooled, then T_1 was significantly better at the α = .02 level. The experimenter went on to say that T_1 was really the treatment of interest and that T_2 was thought to have no effect but was just included for thoroughness, and hence that pooling T_2 and C is acceptable.

To the critical appraiser, this creates doubts concerning hypothesis selection and confirmation from the same set of data. On the other hand, maybe the experimenter really was planning to pool T_2 and C all along (and was sure T_2 was no worse than C), an especially plausible possibility considering that only 10 patients were given T_2. In any case, it is disconcerting that to analyze the problem from a frequentist perspective we have to know what the experimenter's *intentions* were. Trying to analyze hard data by guessing what the experimenter was thinking before doing the experiment seems rather strange. (Of course, a Bayesian won't necessarily be able to avoid such considerations, since the experimenter's statements may well affect prior probability judgements. The uncertainty will be up front in the prior where it belongs, however, with the data speaking for itself through the likelihood function.)

4.2.4 Criticisms

The rosy statements in the previous section concerning the SRP can be viewed as hopelessly misguided by frequentists, since frequency measures are so dependent on stopping rules. Consider Examples 20 and 21, for instance.

EXAMPLE 21 (continued). In the fixed sample size experiment, $\bar{X}_k$ would be an unbiased estimator of θ for either $k = 1$ or 2. If one were to ignore the stopping rule, τ^2, in Case 2, however, and still use the sample mean as the estimator, a "problem" of bias arises. Indeed, the sample mean, $\bar{X}_N$, has

$$\begin{aligned} E_\theta \bar{X}_N &= P_\theta(X_1=1)E_\theta[\bar{X}_1|X_1=1] + P_\theta(X_1=0)E_\theta[\bar{X}_2|X_1=0] \\ &= \theta + \frac{1}{2}\theta(1-\theta), \end{aligned}$$

which is biased upwards. Thus if a conditionalist stated he would be using $\bar{X}_N$, regardless of the stopping rule, the experimenter could use τ^2 and "make θ appear larger than it really is" (if desired).

EXAMPLE 20 (continued). This example has been extensively discussed, in terms of its relationship to the SRP and the LP. Armitage (1961) published (to our knowledge) the first such discussion. Basu (1975) gives a particularly thorough examination of a version of the example. For definiteness in highlighting the "paradox," let us assume that a 95% "confidence interval" for θ is desired, and that an "objective" conditionalist states that, if a fixed sample of size n were taken, he would use the interval

$$C_n(\bar{x}_n) = (\bar{x}_n-(1.96)n^{-\frac{1}{2}}, \bar{x}_n + (1.96)n^{-\frac{1}{2}}). \tag{4.2.4}$$

Of course, he would not interpret confidence in the frequency sense, but instead would (probably) use a posterior Bayesian viewpoint with the noninformative prior density $\pi(\theta) = 1$, which leads to a $\mathcal{N}(\bar{x}_n, n^{-\frac{1}{2}})$ posterior distribution for θ (also, the usual fiducial distribution and the likelihood function for θ).

Suppose now that the experimenter has an interest in seeing that $\theta = 0$ is not in the confidence interval. He could then use the stopping rule in (4.2.3) for some $K > 1.96$. The conditionalist, being bound to ignore the stopping rule, will still use (4.2.4) as his confidence interval, but this can *never* contain zero. Hence the frequentist probability of coverage of (4.2.4), namely

$$r(\theta) = P_\theta^{\tau^2}(C_N(\bar{X}_N) \text{ contains } \theta),$$

is such that $r(0) = 0$ and (by continuity) $r(\theta)$ is near zero for small θ. The experimenter has thus succeeded in getting the conditionalist to perceive that $\theta \neq 0$, and has done so honestly.

Examples 20 and 21 are typical of how the SRP (or the LP) seems to allow the experimenter to mislead a conditionalist. The "misleading", however, is solely from a frequentist viewpoint, and will not be of concern to a conditionalist. Before discussing why, two comments about Example 20 should be gotten out of the way.

(i) Use of a stopping rule, such as that in (4.2.3), can be chancy for an experimenter if $\theta = 0$ is a real possibility, since then N is likely to be extremely large. (This has no real bearing on the arguments here, however.)

(ii) A Bayesian conditionalist might not completely ignore a stopping rule such as that in (4.2.3), if he suspects it is being used because the *experimenter* thinks θ might be zero. The Bayesian might then assign some positive prior probability, λ, to θ being equal to zero, in recognition of the experimenter's possible knowledge. (The stopping rule is affecting only the prior, however, not "what the data says.") A Bayesian analysis in this situation is strikingly different than that in the "noninformative" case. Indeed, as a particular example, if the $\theta \neq 0$ are given prior density $(1-\lambda)$ times a $\mathcal{N}(0,\rho^2)$ density, then the posterior probability that $\theta = 0$, given the observation $\bar{x}_n = Kn^{-\frac{1}{2}}$, is

$$\pi(0|\bar{x}_n = Kn^{-\frac{1}{2}}) = [1+(\lambda^{-1}-1)(1+n\rho^2)^{-\frac{1}{2}}e^{K^2n\rho^2/2(1+n\rho^2)}]^{-1}.$$

For some specific numbers, suppose that $\rho^2 = 10$, $K = 3$, and $n = 10{,}000$. Then,

$$\pi(0|\bar{x}_n = 3n^{-\frac{1}{2}}) = [1+(\lambda^{-1}-1)(.285)]^{-1}.$$

For moderate λ, this says that $\theta = 0$ is quite plausible when n is large, even though $\bar{x}_n$ is three standard deviations from 0. (This is essentially "Jeffrey's" or "Lindley's" Paradox.)

Finally, let us return to Examples 20 and 21 and see if the conditional perspective might not after all be more intuitively appealing. The use of a biased estimator in Example 21 is really not that troubling, since bias has long been a suspect criterion (especially when compared to, say, the plausibility of the Weak Conditionality Principle). We will concentrate on the more disturbing Example 20, therefore.

EXAMPLE 20 (continued). First of all, the likelihood function for θ (when we stop at time n) is proportional to a $\mathcal{N}(\bar{x}_n, n^{-\frac{1}{2}})$ density. This clearly indicates that any particular value of θ near $\bar{x}_n$ is more plausible than a value far from $\bar{x}_n$. The interval in (4.2.4) is a reasonable choice from this viewpoint, although other conditionalists might vary the constant 1.96 or shift somewhat towards a suspected prior mean.

Contrast this with the rather unreasonable way in which a frequentist must behave to obtain, say, coverage probability of at least .95 for all θ when K is large. It can be shown that a frequentist should stick to connected intervals (to minimize size for a given coverage probability) and that, when (say) $\bar{x}_n$ is slightly bigger than $Kn^{-\frac{1}{2}}$ and n is fairly large (which will typically be the case for large K and the stopping rule (4.2.3)), these intervals must usually include both zero and $\bar{x}_n$. Hence, in order to ensure the desired coverage probability at zero when K is large, a frequentist will modify (4.2.4) by replacing a small portion of this interval of "likely" θ, such as $(\bar{x}_n + (1.96-\varepsilon_n)n^{-\frac{1}{2}}, \bar{x}_n + (1.96)n^{-\frac{1}{2}})$, with a big interval, $[0, \bar{x}_n-(1.96)n^{-\frac{1}{2}})$, of unlikely θ. This seems unreasonable. The conditionalist knows that an $\bar{x}_n$ satisfying $\bar{x}_n > Kn^{-\frac{1}{2}}$ (with n very large) could have arisen from $\theta = 0$, but values near $\bar{x}_n$ are so much more likely to be the true θ that he "bets" on these. It should be reemphasized that the conditional analysis is predicated on $\theta = 0$ having no special plausibility; if it does, the Bayesian conclusions (see (ii) above) will be quite different.

The above attempts are probably unlikely to satisfy a frequentist's violated intuition, if the frequentist is not practiced in thinking conditionally. As Savage said in Savage et. al. (1962)

> "I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right."

Of some force may be the argument that, if one's intuition gives contradictory insights, it should be trusted in simple situations, such as Example 2, rather than in extremely complex situations such as Example 20. The next section also lends support to the case for ignoring the stopping rule.

4.2.5 Stopping Rules and Inadmissibility

In Section 3.7 it was argued that behavior in violation of the LP, but consistent with the WCP, tends to be decision-theoretically inadmissible. We rephrase the conclusion, in this section, to show that behavior dependent on the stopping rule will often be inadmissible.

Suppose we have possible observations $X_1, X_2, \ldots,$ as in Section 4.2.2, and are considering two possible stopping rules, τ^1 and τ^2, with respective stopping sets $\{A_m^1\}$ and $\{A_m^2\}$. The stopping rules, τ^1 and τ^2, are presumed to have the possibility of yielding common data, $\underset{\sim}{x}^n$; i.e., there is presumed to be some n^* and $A \subset A_{n^*}^1 \cap A_{n^*}^2$ such that A has positive probability in both E^{τ^1} and E^{τ^2} for all θ. Examples 20 and 21 are of this type, since the sets A_k^2 have positive probability for all θ (under both E^{τ^1} and E^{τ^2}), so that $A = A_k^2$ works.

Suppose that we face a decision problem concerning θ, consisting of choice of an action $a \in \mathcal{A}$ under a loss function $L(a,\theta)$ which is strictly convex in "a" for each θ. (More general loss functions can often be handled also.) Proposed for use in E^{τ^1} and E^{τ^2}, respectively, are decision rules $\delta_1(\underset{\sim}{x}^n)$ and $\delta_2(\underset{\sim}{x}^n)$. If, now, the stopping rule is felt to make a difference, δ_1 and δ_2 should differ for at least some of the possible common observations. Thus we suppose that there is some $A^* \subset A$ for which

(4.2.5) $$\delta_1(\underset{\sim}{x}^{n^*}) \neq \delta_2(\underset{\sim}{x}^{n^*}) \qquad \text{for } \underset{\sim}{x}^{n^*} \in A^*.$$

Consider, next, the mixed experiment, E^*, consisting of observing $J = 1$ or 2 with probability $\frac{1}{2}$ each and then performing E^{τ^J}. This is a well defined sequential experiment with random observation $(J, \underset{\sim}{X}^{N_J})$, N_J being the stopping time for E^{τ^J}. If the WCP is followedfor E^* and (4.2.5) holds, then the decision rule, δ, used for E^* should satisfy

$$\delta((1,\underset{\sim}{x}^{n^*})) \neq \delta((2,\underset{\sim}{x}^{n^*})) \qquad \text{for } \underset{\sim}{x}^{n^*} \in A^*.$$

(Alternatively, this inequality should hold on some A^* if it is felt that the stopping rule actually used - i.e., the value of j - really is relevant to the decision.) But, the estimator

$$\delta^*((j,\underset{\sim}{x}^n)) = \begin{cases} \frac{1}{2}\,\delta((1,\underset{\sim}{x}^{n^*})) + \frac{1}{2}\,\delta((2,\underset{\sim}{x}^{n^*})) & \text{if } n = n^* \text{ and } \underset{\sim}{x}^n \in A^* \\ \delta((j,\underset{\sim}{x}^n)) & \text{otherwise} \end{cases}$$

satisfies (because of the strict convexity of L)

(4.2.6) $$L(\delta^*((j,\underset{\sim}{x}^{n^*})),\theta) < \frac{1}{2}\,L(\delta((1,\underset{\sim}{x}^{n^*})),\theta) + \frac{1}{2}\,L(\delta((2,\underset{\sim}{x}^{n^*})),\theta).$$

Hence, letting E_θ^*, E_θ^1, and E_θ^2 stand for expectation in experiments E^*, E^{τ^1}, and E^{τ^2}, respectively, the frequentist risk (in E^*) of δ^* satisfies

$$\begin{aligned} R(\theta,\delta^*) &= E_\theta^* L(\delta^*((J,\underset{\sim}{X}^{N_J})),\theta) \\ &= \frac{1}{2} E_\theta^1 L(\delta^*((1,\underset{\sim}{X}^{N_1})),\theta) + \frac{1}{2} E_\theta^2 L(\delta^*((2,\underset{\sim}{X}^{N_2})),\theta) \\ &< \frac{1}{2} E_\theta^1 L(\delta((1,\underset{\sim}{X}^{N_1})),\theta) + \frac{1}{2} E_\theta^2 L(\delta((2,\underset{\sim}{X}^{N_2})),\theta) \\ &= E_\theta^* L(\delta((J,\underset{\sim}{X}^{N_J})),\theta) \\ &= R(\theta,\delta). \end{aligned}$$

(The inequality above is strict because of (4.2.6), the fact that A* has positive probability for all θ in E^{τ^1} and E^{τ^2}, and providing $R(\theta,\delta)$ is finite.) This establishes the inadmissibility of allowing the stopping rule to affect the decision making.

EXAMPLE 21 (continued). Suppose that the goal is to estimate θ under squared error loss, and that, because of the bias in use of $\bar{X}_N$ for the stopping rule τ^2, an estimator $\delta_2(\underset{\sim}{x}^n)$ would be used (in E^{τ^2}) such that $\delta_2(\underset{\sim}{x}^n)$ is *not* equal to $\bar{x}_n$ for at least one possible observation, say, $n = 1$, $x_1 = 1$. Let E^{τ^1} be the fixed sample size experiment of size $k = 1$, and suppose that $\delta_1(x_1) = x_1$ would be used for this experiment. However, the experimenter chooses between performing E^{τ^1} and E^{τ^2} on the basis of a fair coin flip ($J = 1$ or 2). This is exactly the situation discussed above, and if the experimenter follows his "instincts" and uses different estimates (depending on J or the actual stopping rule employed)when $x_1 = 1$ is observed, he will be behaving in an inadmissible fashion.

The development above is just a special case of that in Section 3.7, which in turn is basically just a version of the Rao-Blackwell theorem. (Here, J is *not* part of the sufficient statistic for θ in E* when $\underset{\sim}{x}^{n^*} \in A^*$, and decision rules should be based only on the sufficient statistic.) The reason for explicitly giving the development in the sequential framework is to clearly exhibit the conflict between the frequentist desire for admissibility and the intuitive notion that the stopping rule used should matter.

4.2.6 The General Stopping Rule Principle

The SRP can be generalized to an essentially arbitrary sequence of experiments, and shown (in this generality) to follow from the RLP. Thus suppose we have available a sequence $E_1, E_2, \ldots$ of experiments (replacing the i.i.d. observations, $X_1, X_2, \ldots$, of Section 4.2.2) consisting of observing X_j on $\mathcal{X}_j$. We can consider, for each m, the composite experiment $E^m = (\underset{\sim}{X}^m, \theta, \{P_\theta^m\})$ consisting of observing $\underset{\sim}{X}^m = (X_1, \ldots, X_m)$ on $\mathcal{X}^m = \prod_{j=1}^{m} \mathcal{X}_j$ with probability distribution P_θ^m. (If the experiments are independent, P_θ^m will simply by the product measure of the individual distributions on $\mathcal{X}_j$.)

We consider sequential procedures in which we decide, after performing experiments $E_1, \ldots, E_m$, whether or not to perform E_{m+1}. As usual, we can allow this decision to depend upon the outcome of an auxilliary chance mechanism, leading to the following general notion of a stopping rule.

DEFINITION. *A* *__stopping rule__* *is a sequence* $\underset{\sim}{\tau} = (\tau_0, \tau_1, \ldots)$ *in which* $\tau_0 \in [0,1]$ *is a constant and* $\tau_m\colon\ \mathcal{X}^m \to [0,1]$ *a measurable function on* $\mathcal{X}^m$ *for* $m \geq 1$.

The intention is that $\tau_m(\underset{\sim}{x}^m)$ represent the conditional probability of stopping after only m observations, given that we have taken m observations and have observed $\underset{\sim}{x}^m = (x_1, \ldots, x_m)$. The nonrandomized stopping rules discussed in Section 4.2.2 are the special case where the τ_m can only assume the values 0 and 1. When convenient, we shall regard τ_0 as a function on the one-point set $\mathcal{X}^0 = \{\emptyset\}$, the "sample space" for the null experiment $E^0 = (\mathcal{X}^0, \theta, \{P_\theta^0\})$, with P_θ^0 the point mass at $\mathcal{X}^0$'s only point for all θ.

Now define $\mathcal{X}^* = \{(m, \underset{\sim}{x}^m)\colon\ m \in \mathbb{N}, \underset{\sim}{x}^m \in \mathcal{X}^m\}$. For $\underset{\sim}{x}^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$ and $0 \leq j \leq m$, let $\underset{\sim}{x}^{m,j} = (x_1, \ldots, x_j) \in \mathcal{X}^j$ denote the initial segment; of course $\underset{\sim}{x}^0 = \emptyset \in \mathcal{X}^0$ no matter what $\underset{\sim}{x}^m \in \mathcal{X}^m$ might be. For each stopping rule, $\underset{\sim}{\tau}$, determine a family $\{P_\theta^{\underset{\sim}{\tau}}\}$ of measures on $\mathcal{X}^*$ by setting

$$P_\theta^{\underset{\sim}{\tau}}(m,A) = \int_A \prod_{j=0}^{m-1} (1-\tau_j(\underset{\sim}{x}^{m,j}))\tau_m(\underset{\sim}{x}^m)P_\theta^m(d\underset{\sim}{x}^m)$$

for each m and Borel set $A \subset \mathcal{X}^m$. With this definition, τ_0 is the probability of performing E^0, i.e. of taking no data at all. After observing $\underset{\sim}{x}^m$, $\tau_m(\underset{\sim}{x}^m)$ is the conditional probability of taking no more observations.

If $P_\theta^{\underset{\sim}{\tau}}(\mathcal{X}^*) = 1$ for all θ, then the procedure is certain to stop eventually and $\underset{\sim}{\tau}$ is called *proper*; otherwise $\underset{\sim}{\tau}$ is improper and, for at least one θ, there is a positive probability $(1-P_\theta^{\underset{\sim}{\tau}}(\mathcal{X}^*))$ that the sequential procedure would require sampling an infinite number of times. For a proper stopping rule, τ, we can consider the sequential experiment

$$E^{\underset{\sim}{\tau}} = ((N,\underset{\sim}{x}^N),\ \theta,\ \{P_\theta^{\underset{\sim}{\tau}}\}),$$

where N denotes the (random) stopping time. (It is notationally convenient to include N as part of the observation although it could, of course, be recovered from $\underset{\sim}{x}^N$.)

The Stopping Rule Principle for this general setting is formalized in the following theorem, and is shown to follow from the RLP.

THEOREM 5 (The Stopping Rule Principle). *From the* RLP, *it follows that, for any (proper) stopping rule* $\underset{\sim}{\tau}$,

$$Ev(E^{\underset{\sim}{\tau}},(n,\underset{\sim}{x}^n)) = Ev(\underset{\sim}{E}^n,\underset{\sim}{x}^n)$$

for $\{P_\theta^{\underset{\sim}{\tau}}\}$-a.e. $(n,\underset{\sim}{x}^n)$, *i.e. the evidence concerning* θ *in* $E^{\underset{\sim}{\tau}}$ *is identical with that for the fixed sample size experiment* $\underset{\sim}{E}^n$ *(with the observed* n*), so that* $\underset{\sim}{\tau}$ *is irrelevant.*

Proof. Pick $n \in \mathbb{N}$ and let $U_1 \subset \mathcal{X}^*$ be the set of points $(n,\underset{\sim}{x}^n)$ with $\underset{\sim}{x}^n \in \mathcal{X}^n$ satisfying $0 < \tau_n(\underset{\sim}{x}^n)\prod_{j=0}^{n-1}(1-\tau_j(\underset{\sim}{x}^{n,j}))$, and let $c\colon\ U_1 \to (0,\infty)$ be the indicated product. Map U_1 onto $U_2 = \{\underset{\sim}{x}^n \in \mathcal{X}^n\colon\ (n,\underset{\sim}{x}^n) \in U_1\}$ by setting $\varphi(n,\underset{\sim}{x}^n) = \underset{\sim}{x}^n$. Then φ is one-to-one and bimeasurable, and

$$P_\theta^n(A) = \int_{\varphi^{-1}(A)} [1/c(\underset{\sim}{y})]P_\theta^{\underset{\sim}{\tau}}(d\underset{\sim}{y}).$$

The assertion of the theorem now follows from the RLP. ||

Notice that Θ was not required to be a subset of some Euclidean space, nor was $\{P_\theta^m\}$ required to be a dominated family; thus even in situations where no version of the usual LP can apply, the SRP is valid (provided, of course, that the WCP and SP, and hence the RLP, are accepted). This was observed in Pratt (1965).

4.2.7 Informative Stopping Rules

Even the definition of a stopping rule given in the last section may seem somewhat narrow when compared with the vast possibilities for informal stopping discussed in Section 4.2.3. Stopping rules which appear to be more general can be created by introducing an auxilliary variable Y (possibly random), and allowing τ_m, the conditional probability of stopping at stage m, to depend on the value of Y, as well as on $\underset{\sim}{x}^m$. This actually adds very little generality, however, since the values of Y at each stage could simply be incorporated into the data X_i. The following example illustrates the importance of sometimes doing this.

EXAMPLE 23. Suppose $X_1, X_2, \ldots$ are independent Bernoulli (θ) random variables, with $\theta = .49$ or $\theta = .51$. The observations, however, arrive randomly. Indeed, if $\theta = .49$, the observations arrive as a Poisson process with mean rate of 1 per second, while if $\theta = .51$, the observations will arrive as a Poisson process with mean rate of 1 per hour. The "stopping rule" that will be used is to stop the experiment at the first observation that arrives *after* 1 minute has elapsed. One can here introduce Y = time, and write down the stopping rule in terms of Y and the X_i.

It is intuitively clear that this stopping rule cannot be ignored since, if one ends up with 60 observations, knowing whether the experiment ran for 1 minute or $2\frac{1}{2}$ days is crucial knowledge. Incorporating Y into the data resolves all ambiguities, however. Thus, simply define Y_i as the (random) time at which the $i^{\underline{th}}$ observation arrives, and consider the experiment to consist of observing (X_1,Y_1), $(X_2,Y_2), \ldots$. The stopping rule will be given by

$$\tau_m(((x_1,y_1),\ldots,(x_m,y_m))) = \begin{cases} 0 & \text{if } y_m < 1 \\ 1 & \text{if } y_m \geq 1, \end{cases}$$

and is of the form discussed in Section 4.2.6 (or even Section 4.2.2). The importance of the number of observations arriving during the time span of the experiment will be reflected in the portion of the likelihood function due to the y_i.

Slightly more generality might be needed than afforded by simply observing the auxilliary variables at the observation times (as in Example 23) and including them as part of the observations, but the idea is clear: consider *all* available observational information as part of the data X_i. (Of course, some auxilliary information may be considered too informal to include as part of the data, and yet may have some effect on stopping, but such information should only be ignored if it seems relatively unimportant, in which case its effect on stopping can probably also be ignored.)

Even within the above more general perspective on stopping rules, a difficulty might still arise. This difficulty is that the stopping rule might be unknown or partially unknown, in that cessation of the sequential experiment could depend on unobservable random quantities whose probability distributions are not completely known. Following the convention of Section 3.5 and letting θ denote *all* unknown quantities, we could thus write a general stopping rule in terms of $\tau_m(\underset{\sim}{x}^m,\theta)$. (Actually, by including a uniform random variable in θ, it would be possible to have the τ_m assume only the values zero or one.) The general density on $\mathcal{X}^*$ (densities, and discreteness if necessary, being assumed to retain compatibility with Section 3.5) would then be

$$f_\theta^*((n,\underset{\sim}{x}^n)) = [\prod_{j=0}^{n-1}(1-\tau_j(\underset{\sim}{x}^{n,j},\theta))]\tau_n(\underset{\sim}{x}^n,\theta)f_\theta^n(\underset{\sim}{x}^n),$$

where f_θ^n is the density corresponding to P_θ^n. Again following Section 3.5, one could write $\theta = (\xi,\eta)$, where ξ is of interest and η is a nuisance variable. If, for the observed $(n,\underset{\sim}{x}^n)$, $\tau_j(\underset{\sim}{x}^{n,j},\theta)$ does not depend on ξ for $j \leq n$, and if η is a noninformative nuisance parameter (see Section 3.5) for the fixed sample size experiments involving observation of $\underset{\sim}{X}^n$, then the LP and NNPP (see Section 3.5) imply that $\underset{\sim}{\tau}$ is irrelevant. Such a $\underset{\sim}{\tau}$ is called *noninformative*; otherwise $\underset{\sim}{\tau}$ is said to be *informative* and the SRP will not apply. (Raiffa and Schlaifer (1961) introduced these terms.)

We do not pursue the matter further because informative stopping rules occur only rarely in practice (providing all observational information is incorporated into the X_i, as in Example 23). There exists a certain amount of disagreement concerning this point, but the disagreement seems to be primarily due to the misconception that an informative stopping rule is one for which N carries information about θ. This is *not* the definition of an informative stopping rule. Very often N *will* carry information about θ, but to be informative a stopping rule must carry information about θ additional to that available in $\underset{\sim}{X}^N$, and this last will be rare in practice.

4.3 THE IRRELEVANCE OF CENSORING MECHANISMS

4.3.1 Introduction

Another great simplification that application of the LP (or RLP) makes possible is in the handling of censoring. Data is often observed in censored form, and the mechanisms causing the censoring can be quite involved. In most such cases, the LP (or RLP) will imply that only the result of the censoring, and not the censoring mechanism or distribution, is relevant to conclusions about θ.

Section 4.3.2 considers the situation of fixed (nonrandom) censoring, and establishes a version of the irrelevance of censoring mechanisms called Censoring Principle 1. One of the implications of Censoring Principle 1 is that the evidential import of an uncensored observation, from an experiment in which censoring was possible, is the same as the identical observation from an uncensored version of the experiment.

Section 4.3.3 considers random censoring, and establishes conditions under which the distribution of the censoring random variable is irrelevant. The main condition is on the censoring mechanism itself, and leads to the concept of a *noninformative censoring mechanism.* This concept is surprisingly simple and powerful. It is not the case, however, that all sensible censoring mechanisms are noninformative, although many common ones are. This issue is discussed in Section 4.3.4.

The Censoring Principle, as it applies to uncensored observations in nonrandom censoring, seems to be due to John Pratt (see Pratt (1961, 1965), his discussion in Birnbaum (1962a), and the discussion in Savage et. al. (1962)). The general Censoring Principles developed here and the concept of a noninformative censoring mechanism appear to be new, however. Before proceeding with these general developments, it is worthwhile to present an illuminating (and entertaining) example from Pratt's discussion of Birnbaum (1962a). The example makes a simple version of the Censoring Principle seem intuitively obvious.

EXAMPLE 24 (Pratt). A sample of 25 observations was taken from a $\mathcal{N}(\theta,\sigma^2)$ population, and inference about the population mean was desired. All observations were found to lie between 72 and 99, and a standard normal analysis was performed by a frequentist statistician. The statistician reported the analysis to the experimenter, but, curious about the observed 99, asked the experimenter how high his measuring instrument (assumed to be perfectly accurate) read. The experimenter said that the instrument only read to 100, but that, if he had observed a reading of 100, he would have switched to another instrument which had a range of 100 to 1000. The statistician was happy with this response, and satisfied with a job well done.

The next day the experimenter called about something else, and mentioned that he had just checked the high range instrument and found that it was broken. The statistician asked if the experimenter would have had the instrument repaired before completing the previous experiment, to which the experimenter said no. The statistician then said that what were really being observed were observations, X_i, from the truncated distribution with the usual normal density for $x_i < 100$ but the point mass

$$P_{\theta,\sigma^2}(100) = \int_{100}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\theta)^2\}dx$$

at $x_i = 100$. This, said the statistician, calls for a different analysis; for instance, the usual $100(1-\alpha)\%$ confidence interval for θ in the normal situation would no longer have probability of coverage of at least $1-\alpha$ in the truncated situation. The experimenter reacted to this with outrage, saying that he observed precisely what he would have observed had the high range instrument been working (all observations *were* less than 100), and that the condition of an instrument never used in the experiment hardly seemed relevant to the information about θ obtained from the experiment. The frequentist statistician merely shook his head at the naivete of experimenters.

4.3.2 Fixed Censoring and Equivalent Censoring Mechanisms

Consider an experiment $E = (X, \theta, \{P_\theta\})$. Fixed censoring occurs when, instead of X, one observes $Y = g(X)$, where g is a known function from $\mathcal{X}$ into $\mathcal{Y}$. Thus the experiment really performed is $E^g = (Y, \theta, \{P_\theta \circ g^{-1}\})$. (As usual, if $A \subset \mathcal{Y}$, $g^{-1}(A) = \{x \in \mathcal{X}: g(x) \in A\}$.)

EXAMPLE 25. Suppose $X = (X_1,\ldots,X_n)$, where the X_i represent the times of death due to cancer of patients in a cancer survival experiment. Suppose, however, that the experiment will last only ten years, so that the real data will, for the $i^{\underline{th}}$ patient, be

$$Y_i = (Y_i^1, Y_i^2) \equiv (\min\{X_i, 10\}, I_{[0,10]}(X_i)) \tag{4.3.1}$$

(i.e., the truncated survival time and an indicator as to whether the observation is or is not truncated). Thus

$$Y = g(X) \equiv (Y_1,\ldots,Y_n). \tag{4.3.2}$$

This is an example of what is commonly called type I censoring. Example 24 is also of this type.

EXAMPLE 26. Suppose that $X = (X_1,\ldots,X_n)$, but that the n-r largest of the X_i will be truncated at the $r^{\underline{th}}$ largest. Thus let

$$Y_i = (Y_i^1, Y_i^2) \equiv (\min\{X_i, X_{(r)}\},\ I_{[-\infty, X_{(r)}]}(X_i)), \tag{4.3.3}$$

where $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ are the order statistics for X. Again

$$Y = g(X) \equiv (Y_1,\ldots,Y_n). \tag{4.3.4}$$

This is an example of what is commonly called type II censoring.

EXAMPLE 27. Suppose $\mathcal{X} = R^n$, $\mathcal{Y} = \mathcal{X} \times \{0,1\}$, and for some fixed $\rho > 0$,

$$g(X) = \begin{cases} (X,0) & \text{if } |X| \leq \rho \\ (\rho X/|X|,1) & \text{if } |X| > \rho. \end{cases} \tag{4.3.5}$$

Then E^g represents the experiment in which the radius of X is truncated at ρ, but the direction, $X/|X|$, of X is faithfully reported. This is not a standard "type" of censoring, but fits easily within our framework.

Our goal in this section is to indicate that the only effect a censoring mechanism should have on a conclusion is to convey knowledge concerning the actual location of x in $\mathcal{X}$. This may seem intuitively obvious, but Example 24 is a prime illustration of how this is not the case classically. We formalize this notion in the following definition.

DEFINITION. *Let* $E = (X, \Theta, \{P_\theta\})$ *be a given uncensored experiment, and consider two fixed censoring mechanisms* g_1 *and* g_2. *These mechanisms will be said to be* _equivalent on_ $A \subset \mathcal{X}$ *if, for all* $x \in A$,

$$g_1^{-1}(g_1(x)) = g_2^{-1}(g_2(x)). \tag{4.3.6}$$

As a special case, a single fixed censoring mechanism, g, *will be said to be* _equivalent to no censoring on_ $A \subset \mathcal{X}$ *if* $g^{-1}(g(x)) = x$ *for all* $x \in A$.

The idea in the above definition is that, for censoring mechanism g_i, one observes $Y_i = g_i(X)$ and that the only information communicated by the censored data, y_i, is that x was in $g_i^{-1}(y_i)$. If (4.3.6) is satisfied, then g_1 and g_2 will always convey the same information (for $x \in A$). And a g which is equivalent to no censoring (for $x \in A$) conveys exactly the same information that x does. In Example 24, it is clear that the censoring mechanism is equivalent to no censoring on $A = \{x: \ x_i < 100, i = 1,\dots,25\}$; in Example 25, g is equivalent to no censoring on $A = \{x: \ x_i < 10, i = 1,\dots,n\}$; and, in Example 27, g is equivalent to no censoring on $A = \{x: \ |x| < \rho\}$. As an example of possible equivalence of two different censoring mechanisms, consider the following combination of Examples 25 and 26.

EXAMPLE 28. Suppose $X = (X_1,\dots,X_n)$, where the X_i can assume only positive integer values. Let g_1 be as in (4.3.1) and (4.3.2), g_2 be as in (4.3.3) and (4.3.4), and $A = \{x: \ x_{(r)} = 10\}$. It is easy to check that, for $x \in A$,

$$g_1^{-1}(g_1(x)) = g_2^{-1}(g_2(x)) = \{z \in A: \ z_i = x_i \ \text{ if } x_i \leq 10\}.$$

Hence the type I and type II censoring would, in this case, be equivalent on A. (Note that classical analysis tends to treat the two types of censoring differently.)

We now formally state, and justify, the principle that equivalent censoring mechanisms convey the same information about θ, for $x \in A$.

CENSORING PRINCIPLE 1. *If* E^{g_1} *and* E^{g_2} *are two experiments arising from censoring mechanisms equivalent on* A *for an experiment* E, *then*

$$\text{(4.3.7)} \qquad Ev(E^{g_1}, g_1(x)) = Ev(E^{g_2}, g_2(x))$$

for all $x \in A$ *if* $\mathcal{X}$ *is discrete, and for* $\{P_\theta\}$ - a.e. $x \in A$ *in general. As a special case, if* $g^{-1}(g(x)) = x$ *for all* $x \in A$, *then* (4.3.7) *can be replaced by*

$$\text{(4.3.8)} \qquad Ev(E^g, g(x)) = Ev(E,x).$$

Censoring Principle 1 follows from the LP in the discrete case since, by definition, the probabilities of $g_1(x)$ and $g_2(x)$ are equal (to $P_\theta(g_i^{-1}(g_i(x))))$ for all θ. In the general case it follows from the RLP by setting $U_1 = \{g_1(x): \ x \in A\}$, $U_2 = \{g_2(x): \ x \in A\}$, $\varphi(g_1(x)) = g_2(x)$ and $c(g_1(x)) = 1$ for $x \in A$.

The greatest practical use of Censoring Principle 1 is in the case where a censoring mechanism, g, is equivalent to no censoring on A, as was the case in Examples 24, 25, and 27 when no censoring happened to occur. The censoring mechanism can then be completely ignored.

4.3.3 Random Censoring

To generalize the notion of censoring to include random censoring, let $\lambda \in \Lambda$ be a censoring variable with probability density ν on Λ. (To avoid technicalities, discreteness of Λ and $\mathcal{X}$ will be assumed until the end of the section.) Suppose that X and λ are independent (without which very little progress can be made), and that

$$Y = g(X,\lambda) \in \mathcal{Y}$$

is observed. The actual experiment performed can thus be written

$$E^{g,\nu} = (Y,\ \theta,\ \{f_\theta^{g,\nu}\}),$$

where the density of Y is

$$(4.3.9) \qquad f_\theta^{g,\nu}(y) = \sum_{\{(x,\lambda):\ g(x,\lambda) = y\}} f_\theta(x)\nu(\lambda) \quad .$$

EXAMPLE 29. Suppose X represents the time at which a patient in a cancer survival study would suffer death due to cancer, and let λ represent the death time due to competing risks. (We will sidestep the issue of whether or not X and λ can be well-defined.) The actual observation for the patient will be

$$(4.3.10) \qquad Y = (Y^1,Y^2) = g(X,\lambda) \equiv (\min\{X,\lambda\},\ I_{[0,\lambda]}(X)),$$

i.e., the actual time of death, Y^1, and an indicator, Y^2, as to the cause of death. Generalization to involve data from n patients and a variety of competing risks is straightforward, and all the subsequent theory will apply equally well to such a generalization.

The LP, of course, implies that the likelihood function, determined from (4.3.9) for the observed y, contains all the information about θ available from the experiment. The difficulty in utilizing this likelihood function lies in the presence of ν in the expression: typically, ν will be unknown (and complicated). If, however, ν were judged to convey no information about θ (see Section 3.5 and Section 4.3.4) *and* $f_\theta^{g,\nu}(y)$ could be shown to factor into separate terms involving θ and ν, then the difficulty would disappear. This would result in an enormous simplification of the analysis, and is another of the great practical gains that can be realized through adoption of the LP. The following definition gives the key characterization of censoring mechanisms for which this program is possible.

DEFINITION. *A censoring mechanism* $g\colon \mathcal{X} \times \Lambda \to \mathcal{Y}$ *is* <u>*noninformative*</u> *at* $y \in \mathcal{Y}$ *if* $g^{-1}(y)$ *is a product set, i.e., if*

$$g^{-1}(y) = A_y \times B_y, \textit{ where } A_y \subset \mathcal{X} \textit{ and } B_y \subset \Lambda.$$

EXAMPLE 29 (continued). Here

$$g^{-1}((y^1,y^2)) = \begin{cases} (y^1,\infty) \times \{y^1\} & \text{if } y^2 = 0 \\ \\ \{y^1\} \times [y^1,\infty) & \text{if } y^2 = 1, \end{cases}$$

so that g is a noninformative censoring mechanism at *all* $y \in \mathcal{Y}$.

EXAMPLE 27 (continued). Consider the situation in Example 27, but assume that ρ is now a random variable (and, hence, replace $g(X)$ by $g(X,\rho)$). Since

$$g^{-1}((y^1,y^2)) = \begin{cases} \{y^1\} \times [|y^1|,\infty) & \text{if } y^2 = 0 \\ \\ \{cy^1\colon\ c > 1\} \times \{|y^1|\} & \text{if } y^2 = 1, \end{cases}$$

g is a noninformative censoring mechanism at *all* $y \in \mathcal{Y}$.

If g is noninformative at y, then (4.3.9) becomes (employing also the independence of X and λ)

$$(4.3.11) \qquad f_\theta^{g,\nu}(y) = [\sum_{x \in A_y} f_\theta(x)][\sum_{\lambda \in B_y} \nu(\lambda)],$$

so that (for known ν), the LP implies that all information concerning θ from the experiment is contained in

$$(4.3.12) \qquad \ell_y^*(\theta) = \sum_{x \in A_y} f_\theta(x).$$

If ν is unknown but "noninformative" for θ (see Sections 3.5 and 4.3.4), the same conclusion follows from the NNPP in Section 3.5.5. These conclusions can be summarized as follows.

CENSORING PRINCIPLE 2. *If $\mathcal{X}$ and Λ are discrete,* X *and* λ *are independent,* g *is noninformative at the observed* y, *and either* ν *is known or it is unknown but noninformative, then* $\mathrm{Ev}(E^{g,\nu},y)$ *depends only on* $\ell_y^*(\theta)$ (*from* (4.3.12)).

Note that this principle does *not* say that censoring has no effect on the analysis. Indeed, $\ell_y^*(\theta)$ will often fail to be proportional to $\ell_x(\theta) = f_\theta(x)$, which would be used if no censoring occurred. Another point is that the only censoring mechanisms which can *guarantee* that $\mathrm{Ev}(E^{g,\nu},y)$ does not depend on ν (for ν as in the principle) are noninformative censoring mechanisms. This is established in the following theorem.

THEOREM 6. *If* g: $\mathcal{X} \times \Lambda \to \mathcal{Y}$ *is* <u>*not*</u> *a noninformative censoring mechanism at* y, *then there exists* $\{f_\theta\}$ *on* $\mathcal{X}$ *such that* $\mathrm{Ev}(E^{g,\nu},y)$ *depends on* ν.

Proof. If $g^{-1}(y)$ is not a product set, it follows that there exist two points λ_1, $\lambda_2 \in \Lambda$ such that either

$$\Omega_1 = \{x: \ g(x,\lambda_1) = y \text{ and } g(x,\lambda_2) \neq y\},$$

or

$$\Omega_2 = \{x: \ g(x,\lambda_1) \neq y \text{ and } g(x,\lambda_2) = y\},$$

or both are nonempty. Consider ν that are concentrated on $\{\lambda_1,\lambda_2\}$, and define

$$\Omega_3 = \{x: \ g(x,\lambda_1) = g(x,\lambda_2) = y\}.$$

Equation (4.3.9) can then be written

$$f_\theta^{g,\nu}(y) = \nu(\lambda_1)P_\theta(\Omega_1) + \nu(\lambda_2)P_\theta(\Omega_2) + P_\theta(\Omega_3).$$
$$= \nu(\lambda_1)[P_\theta(\Omega_1) - P_\theta(\Omega_2)] + P_\theta(\Omega_2 \cup \Omega_3).$$

Thus, as long as $\{f_\theta\}$ is chosen so that $[P_\theta(\Omega_1)-P_\theta(\Omega_2)]$ and $P_\theta(\Omega_2 \cup \Omega_3)$ are not proportional as functions of θ, the likelihood function will depend on $\nu(\lambda_1)$. ||

Finally, we leave the discrete setting and develop a very general version of Censoring Principle 2, based on the RLP. We will assume that Λ and $\mathcal{Y}$ are LCCB spaces, that ν is a Borel probability measure, and that $g: \ \mathcal{X} \times \Lambda \to \mathcal{Y}$ is a Borel function. The actual experiment of observing $Y = g(X,\lambda)$ is $E^{g,\nu} = (Y, \theta, \{P_\theta^{g,\nu}\})$, where

$$(4.3.13) \qquad P_\theta^{g,\nu}(C) = (P_\theta\times\nu)(\{(x,\lambda): \ g(x,\lambda) \in C\}).$$

The definition of a noninformative censoring mechanism at y remains unchanged, and leads to the following principle.

CENSORING PRINCIPLE 2'. *Let* $C \subset \mathcal{Y}$ *be a Borel set such that* g *is a noninformative censoring mechanism at all* $y \in C$. *Suppose* ν_1 *and* ν_2 *are Borel probability measures (for* λ*) which are mutually absolutely continuous on* $C^* = \bigcup_{y\in C} B_y$ *(where* $g^{-1}(y) = A_y \times B_y$*). Then, if either* (i) ν_1 *and* ν_2 *are known, or* (ii) *they are unknown but noninformative (see Sections* 3.5 *and* 4.3.4*), it should be the case that*

$$(4.3.14) \qquad Ev(E^{g,\nu_1},y) = Ev(E^{g,\nu_2},y) \quad \textit{for } \{P_\theta^{g,\nu_1}\}\text{-a.e. } y \in C.$$

The conclusion in Censoring Principle 2' is not quite as strong as that in the original Censoring Principle 2, in that evidentiary equivalence is only stated to hold among equivalence classes of ν (on C). Of course, if the possible ν under consideration are known to be absolutely continuous with respect to some measure μ, then it can be stated that ν is irrelevant (if it is noninformative). For instance, in Example 29 it may be reasonable to assume that ν is absolutely continuous with respect to Lebesgue measure, and is thus ignorable (if noninformative).

It seems likely that Censoring Principle 2' is a general consequence of the RLP. This is because one can define (see the RLP) $U_1 = U_2 = C$, φ to be the identity map, and

$$c(y) = c((x,\lambda)) = \nu_2(d\lambda)/\nu_1(d\lambda), \tag{4.3.15}$$

and seek to show that (for any Borel subset, D, of C)

$$P_\theta^{g,\nu_2}(D) = \int_D c(y)P_\theta^{g,\nu_1}(dy). \tag{4.3.16}$$

Since (4.3.16) is essentially (3.4.1) of the RLP (where 1/c has been replaced by c for convenience in what follows), Censoring Principle 2' would be an immediate consequence of the RLP (and the NNPP of Section 3.5, if the ν_i are unknown but noninformative). And (4.3.16) seems to be a correct equation: it can trivially be verified to hold in the discrete setting, for instance. Unfortunately, severe measurability difficulties (due to the possible nasty nature of g) prevented us from verifying (4.3.16), in general. Under additional conditions, however, we were able to show that (4.3.16) does hold for some positive c, which suffices, by the above argument, to establish Censoring Principle 2' as a consequence of the RLP. Furthermore, though somewhat technical, these additional conditions involve only the censoring mechanism, g, and not the P_θ or ν. This makes general verification of the irrelevance of any specific censoring mechanism possible.

THEOREM 7. *Let* g *be a noninformative censoring mechanism at all* $y \in C$, *and suppose that there exist sequences* $\{\varphi_n\}$ *and* $\{\psi_n\}$ *of measurable mappings* $\varphi_n\colon \mathcal{X} \to \mathcal{X}$ *and* $\psi_n\colon \Lambda \to \Lambda$, *such that the functions* $g_n(x,\lambda) \equiv g(\varphi_n(x),\psi_n(\lambda))$ *are countably valued and the* σ*-algebras,* $\mathcal{B}$, $\mathfrak{F}_n$, *and* $\mathcal{B}_n$, *generated on* $\mathcal{X} \times \Lambda$ *by* $g(x,\lambda)$, $\varphi_n(x)$, *and* $g_n(x,\lambda)$, *respectively, satisfy the conditions*

i) $$\mathfrak{F}_n \vee \mathcal{B}_n \subset \mathfrak{F}_{n+1} \vee \mathcal{B}_{n+1}$$

ii) $$\bigcap_{m=1}^{\infty} \bigvee_{n=m}^{\infty} \mathscr{B}_n = \mathscr{B}.$$

Then for any two probability measures ν_1 *and* ν_2 *on* Λ*, which are mutually absolutely continuous on* C,

$$(4.3.17) \qquad \int_C h(y)P^{g,\nu_2}(dy) = \int_C h(y)c(y)P^{g,\nu_1}(dy)$$

for every bounded measurable function h *on* $\mathcal{Y}$ *and every probability measure* P *on* $\mathcal{X}$. *(Note that* (4.3.16) *follows trivially from* (4.3.17). *Hence, under the above conditions, Censoring Principle* 2' *is a consequence of the* RLP.*)*

Proof. We will prove the theorem for $C = \mathcal{Y}$. The modifications needed for arbitrary C are obvious. For $n \geq 1$ let $\{y_j^n\}_{j\geq 1}$ be the countably many values of g_n; the σ-algebra $\mathscr{B}_n$ is generated by the countable partition $\wp^n = \{A_j^n \times B_j^n\}$ of $\mathcal{X} \times \Lambda$ into the measurable rectangles (or product sets) $A_j^n \times B_j^n = g_n^{-1}(y_j^n)$, where $A_j^n = \varphi_n^{-1}(A_{y_j^n})$ and $B_j^n = \psi^{-1}(B_{y_j^n})$; here (as before) A_y and B_y are determined by the relation $g^{-1}(y) = A_y \times B_y$. For $(x,\lambda) \in \mathcal{X} \times \Lambda$, define

$$(4.3.18) \qquad \bar{c}_n(x,\lambda) = \begin{cases} \nu_2(B_j^n)/\nu_1(B_j^n) & \text{if } \nu_1(B_j^n) > 0, \\ 1 & \text{if } \nu_1(B_j^n) = \nu_2(B_j^n) = 0, \end{cases}$$

$$\bar{c}(x,\lambda) = \limsup_{n\to\infty} \bar{c}_n(x,\lambda),$$

where j is determined by the relation $g^n(x,\lambda) = y_j^n$.

A direct computation verifies that, for any probability measure P on $\mathcal{X}$,

$$(4.3.19) \qquad \bar{c}_n = E^{P\times\nu_1}\left[\frac{\nu_2(d\lambda)}{\nu_1(d\lambda)}\middle|\mathcal{F}_n \vee \mathscr{B}_n\right].$$

Indeed, to show this it is sufficient to take any bounded measurable function, h, on $\mathcal{X} \times \mathcal{Y}$ and note that

$$\int_{\mathcal{X}} \int_{\Lambda} h(\varphi_n(x), g_n(x,\lambda))\bar{c}_n(x,\lambda)P(dx)\nu_1(d\lambda)$$

$$= \sum_{j=1}^{\infty} \int_{A_j^n} h(\varphi_n(x), y_j^n) \frac{\nu_2(B_j^n)}{\nu_1(B_j^n)} P(dx)\nu_1(B_j^n)$$

$$= \int_{\mathcal{X}} \int_{\Lambda} h(\varphi_n(x), g_n(x,\lambda))P(dx)\nu_2(d\lambda).$$

By (4.3.19) and Condition (i), $\bar{c}_n$ is a uniformly integrable martingale on $(\mathcal{X} \times \Lambda, (\mathfrak{F}_n \vee \mathscr{G}_n)_{n\geq 1}, P\times\nu_1)$, for every P. Hence $\bar{c}_n$ converges to $\bar{c}$ with $P\times\nu_1$-measure 1 for every P, and satisfies

(4.3.20) $$\bar{c}_n = E^{P\times\nu_1}[\bar{c}|\mathfrak{F}_n \vee \mathscr{G}_n] \qquad \text{for every } n \geq 1.$$

Since we may take P to be concentrated on any single point $x \in \mathcal{X}$, we have actually shown that $\bar{c}_n(x,\lambda)$ converges to $\bar{c}(x,\lambda)$ for every $x \in \mathcal{X}$ and ν_1-almost every λ in Λ (where the exceptional set of ν_1-measure zero may depend on x).

It is obvious from the definition of $\bar{c}_n$ that $\bar{c}_n(x,\lambda)$ depends on x and λ only through $y_j^n = g_n(x,\lambda)$, and therefore that $\bar{c}_n$ is $\mathscr{G}_n$-measurable. It follows that $\bar{c}$ is measurable over $\bigvee_{n=m}^{\infty} \mathscr{G}_n$ for each m, and so (by Condition (ii)) $\bar{c}$ is measurable over $\mathscr{G}$. Since any $\mathscr{G}$-measurable function may be written as a Borel-measurable function of g, there exists some positive function, c, on $\mathcal{Y}$ with

(4.3.21) $$\bar{c}(x,\lambda) = c\circ g(x,\lambda).$$

Now let h be bounded and measurable on $\mathcal{Y}$, let P be the probability measure on $\mathcal{X}$, and set

(4.3.22) $$\bar{h}_n = E^{P\times\nu_2}[h\circ g|\mathfrak{F}_n \vee \mathscr{G}_n].$$

Again the martingale convergence theorem implies that $\bar{h}_n(x,\lambda)$ converges to $h\circ g(x,\lambda)$ for $P\times\nu_1$-almost every (x,λ), since $h\circ g$ is $\mathscr{G}$-measurable and Conditions (i) and (ii) imply that $\mathscr{G} \subset \bigvee_{n=1}^{\infty} \mathscr{G}_n \subset \bigvee_{n=1}^{\infty} (\mathfrak{F}_n \vee \mathscr{G}_n)$. By Lebesgue's dominated convergence theorem, (4.3.20), and (4.3.19),

$$\begin{aligned}
\int_{\mathcal{Y}} hc \; dP^{g,\nu_1} &= \int_{\mathcal{X}} \int_{\Lambda} h\circ g \; c\circ g \; P(dx)\nu_1(d\lambda) \\
&= \lim_{n\to\infty} \int_{\mathcal{X}} \int_{\Lambda} \bar{h}_n \bar{c} \; P(dx)\nu_1(d\lambda) \qquad \text{(by DCT)} \\
&= \lim_{n\to\infty} \int_{\mathcal{X}} \int_{\Lambda} \bar{h}_n \bar{c}_n \; P(dx)\nu_1(d\lambda) \qquad \text{(by (4.3.20))} \\
&= \lim_{n\to\infty} \int_{\mathcal{X}} \int_{\Lambda} \bar{h}_n \; P(dx)\nu_2(d\lambda) \qquad \text{(by (4.3.19))} \\
&= \int_{\mathcal{X}} \int_{\Lambda} h\circ g \; P(dx)\nu_2(d\lambda) \qquad \text{(by DCT)} \\
&= \int_{\mathcal{Y}} h \; dP^{g,\nu_2}.
\end{aligned}$$

This verifies (4.3.17) and completes the proof. ||

Remark 1. In case it is possible to find $\{\varphi_n\}$ and $\{\psi_n\}$ so that $\mathscr{B}_n \subset \mathscr{B}_{n+1}$, Condition (i) in the theorem may be eliminated and Condition (ii) can be simplified to $\bigvee_{n=1}^{\infty} \mathscr{B}_n = \mathscr{B}$.

Remark 2. If φ_n and ψ_n are themselves countably-valued, then obviously g_n is also, so the theorem applies if Conditions (i) and (ii) are satisfied.

EXAMPLE 29 (continued). Letting $\langle a \rangle$ denote the closest integer to a (the larger integer in case of a tie), define

$$\varphi_n(x) = 2^{-n} \langle 2^n x \rangle \text{ and } \psi_n(\lambda) = 2^{-n} \langle 2^n \lambda \rangle.$$

It is straightforward to verify that Conditions (i) and (ii) in Theorem 7 are satisfied, and hence that Censoring Principle 2' follows in complete generality from the RLP (for this situation).

EXAMPLE 27 (continued). Let $\rho_n = \{A_j^n\}_{j \le J_n}$ be a sequence of partitions of the unit sphere (in R^n) into finitely many Borel sets such that ρ_{n+1} refines ρ_n and

$$\lim_{n\to\infty} \max_{j \le J_n} \text{diam}(A_j^n) = 0.$$

Let $\{\xi_j^n\}$ be a collection of points such that $\xi_j^n \in A_j^n$, and define

$$\varphi_n(x) = i2^{-n}\xi_j^n \qquad \text{if} \quad i \leq 2^n|x| < i+1 \text{ and } x/|x| \in A_j^n,$$

$$\psi_n(\rho) = k2^{-n} \qquad \text{if} \quad k \leq 2^n\rho < k+1.$$

Again the Conditions (i) and (ii) of Theorem 7 are easily verified, so that this censoring mechanism is also generally irrelevant.

4.3.4 Informative Censoring

It is, of course, not always the case that the censoring mechanism or distribution can be ignored. There are very few instances of fixed censoring wherein the mechanisms can be labeled informative, so we will concentrate in this section on random censoring.

The most common reason for being unable to ignore the censoring distribution, ν, in random censoring is dependence of the random variable X and the random censoring variable λ. In Example 29, for instance, one may have a non-cancer death which occurred because cancer substantially lowered overall health. Indeed in competing risk theory, in general, dependence between X and the censoring variables may be the rule rather than the exception. Such dependence makes Censoring Principle 2 inapplicable, and indeed $\ell_y(\theta)$ will typically depend upon ν in such situations. (The LP is still valid, of course.)

A second possible reason that the censoring distribution might be informative is that the censoring mechanism, g, might fail to be noninformative. As a very simple example, suppose the actual observation is

$$Y = g(X,\lambda) = X+\lambda,$$

where $X \in \mathcal{X} = (0,\infty)$ and $\lambda \in \Lambda = (0,\infty)$. It is easy to check that $g^{-1}(y)$ is *not* a product set in $\mathcal{X} \times \Lambda$ for any y, so that g clearly fails to be noninformative. For such g, $\ell_y(\theta)$ will typically depend on ν.

A third reason that ν might not be ignorable is that ν will often be unknown, and there could be some "prior" relationship between ν and θ. Again, the notation of Section 3.5 is convenient here. Thus let θ stand for *all* unknown aspects of the situation and write $\theta = (\xi,\eta)$, where ξ is of interest and η is a nuisance variable (presumably containing unknown aspects of the distribution, ν, of λ). For instance, if $X \sim P_\xi$ and $\lambda \sim \nu_\eta$ are similar competing risks, there might well be suspected relationships between ξ and η which prevent ν_η from being ignored (even if X and λ are independent and g is noninformative). We will not repeat the discussion of Section 3.5 concerning when and why η (and hence ν_η) can be ignored in such situations.

A final kind of informative censoring should be mentioned, even though it is not censoring in the formal sense we have defined. This is censoring in which censored data is simply not observed or recorded. Thus, for the censoring mechanism described in (4.3.1) and (4.3.2), it could be the case that an $X_i > 10$ is not observed or even known to have existed. Such a situation is easily dealt with by recognizing that the relevant probability distribution of the *observed* X_i is the conditional distribution, given that $X_i \leq 10$. The censoring mechanism will usually enter into this conditional distribution in a nonignorable fashion, however.

Interestingly enough, this omission of data due to censoring can arise from the methods of *reporting* data (c.f. Dawid and Dickey (1977)). An obvious example is that of a trade journal which only publishes results of experiments which provide "significant" evidence according to some criteria. The data of interest, for a given issue, would be all data from experiments on that issue, but only that data leading to "significance" will become available; the rest will be censored. This is a very complicated problem, and it is not at all clear how to analyze the situation. The censoring of the journal clearly can not be ignored, however.

4.4 SIGNIFICANCE TESTING

4.4.1 Conflict with the LP

Significance testing of a hypothesis (used here in the sense of P-values, rather than α-level testing) is viewed by many as a crucial element of statistics, yet it provides a startling and practically serious example of conflict with the LP. A significance test of the hypothesis H_0, that X has distribution P^0, proceeds by defining some statistic T(X), where large values of T supposedly cast doubt on H_0, and then calculating, for the given observation x, the significance level (or P-value) of x,

$$p = P^0(T(X) \geq T(x)) = \int_{\{y: T(y) \geq T(x)\}} P^0(dy) \qquad (4.4.1)$$

(i.e., the probability under P^0 of observing x or something more "extreme"). If this is small, then one supposedly doubts that H_0 could be true. General discussions of significance testing (including discussions of important practical issues such as "real" versus "statistical" significance) can be found in Edwards, Lindman, and Savage (1963), Hacking (1965), Morrison and Henkel (1970), Edwards (1972), Cox and Hinkley (1974), Dempster (1974a,b), Pratt (1976,1977), Cox (1977), Barnard (1980), Good (1981), Barnett (1982), Berger (1985), Hall and Selinger (1986), and Berger and Delampady (1987).

A very common setting for significance testing is the parametric framework of testing H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$. Then the null distribution, P^0, is simply P_{θ_0} in our usual notation (or $f_{\theta_0}(\cdot)$ if densities exist). In this parametric setting it is clear that reporting significance levels violates the LP, since significance levels involve averaging over sample points other than just the observed x (see (4.4.1)). The extremely serious practical problems that can result are discussed in Section 4.4.2.

Significance testing is also frequently used when only a single model P^0 is being contemplated. Testing of fit to a specified model is a common example. Since only one probability distribution is then involved, there is no likelihood function; it is hence often argued that the LP cannot apply to such a situation. Arguments to the contrary will be given in Section 4.4.3.

4.4.2 Averaging Over "More Extreme" Observations

The logic behind including all data "more extreme" than the given x, when calculating p, is not particularly convincing. Consider the following artificial example, related to an example in Cox (1958).

EXAMPLE 30. Suppose, under P_0 and P_1, respectively, that X has the distributions given in the following table.

x	0	1	2	3	4
$P_0(x)$	.75	.14	.04	.037	.033 .
$P_1(x)$	.70	.25	.04	.005	.005

If T(x) = x were used as the test statistic for a significance test of either P_0 or P_1 (i.e., if large x were considered "extreme"), and if x = 2 were observed, then the significance level against P_0 alone would be

$$p_0 = P_0(X \geq 2) = .11,$$

while the significance level against P_1 alone would be

$$p_1 = P_1(X \geq 2) = .05.$$

(We are not thinking here of testing P_0 versus P_1; the focus is on comparing significance tests of each separate hypothesis.) Thus x = 2 would provide "significant evidence against P_1 at the 5% level," but would not even provide "significant evidence against P_0 at the 10% level."

The concern here, of course, is that were P_0 and P_1 being considered simultaneously as possible models, likelihood reasoning would argue that they are equally supported by x = 2; their likelihood ratio is then equal to one. When considered in isolation therefore, it is definitely strange that x = 2 provides such different significance levels for P_0 and P_1.

Jeffreys (1961) clearly exposed the questionable logic behind significance levels, stating

> "...a hypothesis which may be true may be
> rejected because it has not predicted
> observable results which have not occurred."

In the example here, neither P_0 nor P_1 "predicts" that x = 3 or x = 4 will occur, and indeed they do not occur, but P_1 would be rejected at the 5% level, while P_0 would not, because P_1 "predicts" these *unobserved* results even less than P_0.

Questionable logic could perhaps be overlooked if it made little difference in practice, but here the averaging over other observations will virtually *always* have a profound effect. Consider the following example from Edwards, Lindman, and Savage (1963).

EXAMPLE 30.1. Suppose $X = (X_1,\ldots,X_n)$, where the X_i are i.i.d. $\mathcal{N}(\theta,\sigma^2)$, σ^2 known. The usual test statistic for testing H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$ is

$$T(X) = \sqrt{n}|\bar{X} - \theta_0|/\sigma,$$

where $\bar{X}$ is the sample mean. If $t = T(x)$ is the observed test statistic, the significance level is then

$$p = 2(1 - \Phi(t)),$$

where Φ is the standard normal c.d.f..

Consider, now, this testing scenario from a likelihood perspective. Were H_1 given by H_1: $\theta = \theta_1$, it would have been natural to use, as the comparative evidence for the two hypotheses, the observed likelihood ratio

$$L_{\theta_1} = f_{\theta_0}(x)/f_{\theta_1}(x).$$

Unfortunately, the actual H_1 consists of all $\theta \neq \theta_0$, making it difficult to define a true likelihood ratio, L, of H_0 to H_1. It seems clear, however, that a lower bound on L is

$$\underline{L} = f_{\theta_0}(x)/\sup_{\theta\neq\theta_0} f_\theta(x).$$

The evidence against H_0 is certainly no stronger than $\underline{L}$.

An easy calculation shows that, in this example,

$$\underline{L} = \exp\{-\tfrac{1}{2}t^2\}.$$

The following table gives values of $\underline{L}$ for various t, and also gives the significance levels associated with these t. (The $\underline{L}g$ row is discussed later.)

Table 1. Likelihood Ratio Bounds and Significance Levels

t	1.645	1.960	2.576	3.291
p	.10	.05	.01	.001
$\underline{L}$	.258	.146	.036	.0044
$\underline{L}g$	.644	.409	.123	.018

The surprise here is that $\underline{L}$ is much larger than p. When p is .05 for instance, $\underline{L}$ is .146, indicating that the data provides *no more* than 1 to 7 evidence against H_0.

$\underline{L}$ itself can be argued to be misleadingly small because it is based on maximizing the "likelihood of H_1." More reasonable is to use, as the "likelihood of H_1", an average of $f_\theta(x)$ over all $\theta \neq \theta_0$. This leads to a *weighted likelihood ratio*

$$Lg = f_{\theta_0}(x)/\int_{\{\theta\neq\theta_0\}} f_\theta(x)\ g(\theta)d\theta,$$

where g is some density (or "weight function"). A Bayesian would choose g to be the conditional prior density on H_1, in which case Lg would be the *Bayes factor*.

Regardless of interpretation, one can gain insight into the impact of such evidence measures by calculating lower bounds on Lg over reasonable classes of g. For instance, in Berger and Sellke (1987) it is shown that for *any* density g which is a nonincreasing function of $|\theta-\theta_0|$, Lg is at least as large as $\underline{L}g$, given in the last row of Table 1. The indication is thus that, when p = .05 say, the evidence against H_0 is actually no stronger than 1 to $2\frac{1}{2}$. (And if one tried "natural" functions g, one would find that Lg is typically 1 or more when p = .05; see, e.g., Jeffreys (1961).)

The above example is quite disturbing. It indicates that the classical statistician and the conditionalist will often reach very different conclusions with the same data, precisely because one averages over all "extreme" sample points while the other uses only the observed data. (Berger and Sellke (1987) specifically show that this averaging is the source of the discrepancy.) Furthermore, the discrepancy between significance levels and conditional measures of evidence (e.g., $\underline{L}$, Lg or $\underline{L}$g, the posterior probability of H_0, and even conditional frequentist measures -cf. Berger and Sellke (1987)) has been shown to hold in a huge variety of significance testing problems involving a "precise" hypothesis. (H_0 need not be a point null for the discrepancy to arise - see Berger and Sellke, 1987, and Berger and Delampady, 1987 - but if H_0 is, say, a one-sided hypothesis, then the discrepancy may not arise - see Casella and Berger, 1987.) Note also that this discrepancy is very related (but not identical) to "Jeffreys's Paradox" or "Lindley's Paradox". These issues are explored, in depth, in Edwards, Lindman, and Savage (1963), Berger and Sellke (1987) and Berger and Delampady (1987). Other relevant works include Lindley (1957, 1977), Jeffreys (1961), DeGroot (1973), Dempster (1974b), Dickey (1977), Smith and Spiegelhalter (1980), Good (1981, 1984), Shafer (1982), Zellner (1984), Berger (1985), Delampady and Berger (1987), and Delampady (1986a,b).

One defense of averaging over other observations (and at the same time an attack on the LP) that is sometimes advanced is the claim that it is necessary to consider what observations *might have* occurred. It is, however, a misconception to believe that the LP fails to do this. Indeed, in determining the likelihood function (or family of distributions for X), it is crucial to consider and compare the possible x that might be observed. Once this has been done, however, and the data obtained, the LP states that only the observed $\ell_x(\theta)$ is needed.

4.4.3 Testing A Single Null Model

When only P^0 has been formulated, it has been argued that significance testing does not violate the LP because nothing resembling a likelihood function exists. Although correct in a certain formal sense, there are several weaknesses to the argument.

Perhaps the most serious weakness follows from the observations in the previous section: if averaging over "extreme" sample points is virtually *always* bad in testing a "precise" null when alternatives are given, it seems incredibly optimistic to believe that such averaging will be reasonable when alternatives are not given. The argument that "significance testing is the only available statistical procedure" is hardly persuasive when it is known that this available statistical procedure is bad for testing precise hypotheses.

A second weakness of the argument that only P^0 exists is that implicit alternatives to P^0 often are present. Indeed, alternatives must enter, at least informally, into the choice of the test statistic $T(x)$. For instance, in Example 30 it seems justifiable to use $T(x) = x$ to measure "extreme" only if the alternatives that one has in mind are, say, alternatives which are stochastically larger than P_0 (so that a large x tends to support the alternatives more than it tends to support P_0.) As another example of the implicit presence of alternatives, consider chi-square testing of fit.

EXAMPLE 30.2. Consider a statistical experiment in which n independent and identically distributed random quantities $X_1, X_2, \ldots, X_n$ are observed from a distribution F. It is desired to conduct a significance test of the hypothesis H_0: $F = F_0$, where F_0 is a specified distribution. A common test procedure, when no alternatives are specified, is the chi-square test of fit. *Chi-Square Test Procedure*: First, a partition $\{a_i\}_{i=0}^{m}$ of the real line is selected. Then the sample frequencies of the n observations in the cells of the partition are calculated. Let $\underset{\sim}{z} = (z_1, \ldots, z_m)^t$ denote these frequencies; thus z_i = number of X_i's in $(a_{i-1}, a_i]$. Define

$$\theta_i = F(a_i) - F(a_{i-1}) = P^F(a_{i-1} < X \leq a_i),$$

$$\theta_i^0 = F_0(a_i) - F_0(a_{i-1}) = P^{F_0}(a_{i-1} < X \leq a_i),$$

and

$$\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_m)^t, \ \underset{\sim}{\theta}^0 = (\theta_1^0, \ldots, \theta_m^0)^t.$$

Then the chi-square test procedure is to calculate the test statistic

$$t = \sum_{i=1}^{m} \frac{(z_i - n\theta_i^0)^2}{n\theta_i^0} ,$$

and approximate the significance level by

$$p = P(\chi^2_{m-1} \geq t),$$

where χ^2_{m-1} is a chi-square random variable with m-1 degrees of freedom.

The implied alternatives here arise from the fact that $\underset{\sim}{Z}$ has a *Multinomial* $(n,\underset{\sim}{\theta})$ distribution, so that basing the test on $\underset{\sim}{z}$ is equivalent to acknowledging the test to be that of H_0: $\underset{\sim}{\theta} = \underset{\sim}{\theta}^0$ versus H_1: $\underset{\sim}{\theta} \neq \underset{\sim}{\theta}^0$. (Use of t can be argued to further imply that the alternatives, $\underset{\sim}{\theta} \neq \underset{\sim}{\theta}^0$, are roughly ordered in plausibility according to $\eta = \sum_{i=1}^{m} (\theta_i - \theta_i^0)^2/\theta_i^0$, so that one is really testing H_0: $\eta = 0$ versus H_1: $\eta > 0$.) But this is a parametric problem with specified alternatives (and hence a likelihood function) so that LP-compatible testing methods can apply. Indeed, in Delampady and Berger (1987) it is shown that the same type of difficulty for significance testing, that was discussed in Section 4.4.2, exists here: the significance level is typically much smaller than sensible conditional measures of the evidence for H_0.

The above argument, that there are implicit alternatives in significance testing, can actually be given a quite general formal foundation. It has previously been mentioned that the actual sample space $\mathcal{X}$ will be discrete in practice. But then, as discussed in Section 3.6.1, even the set $\{P_\theta\}$ of *all* distributions on $\mathcal{X}$ actually results in a definable likelihood function. Furthermore, a significance test of P^0 can be identified with a test of H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$, where $P_{\theta_0} = P^0$. Thus the LP can apply, and argues against the use of significance levels.

Although formally correct, we do not ascribe much practical importance to this last argument, because the class of *all* alternatives to P^0 is typically much too big to suggest a sensible analysis. In practice, some consideration of the type of alternatives that are expected is necessary, even in classical significance testing. In choosing a test statistic T(x), for instance, we earlier observed that it is often necessary to consider alternatives when defining "extreme." It has been argued that it may be easier to guess a reasonable T, reflecting intuitive judgements as to which observations support H_0 and which support alternatives, than to attempt explicit consideration of alternatives and construction of T by, say, likelihood ratio comparisons of P^0 with the alternatives. The argument that one can do better by use of intuition, than by explicit consideration of important relevant features of a problem (here, the alternatives), is difficult to refute, but is an argument that we would feel very uncomfortable having as a basis for our approach to science and understanding. Even more troubling is the fact that significance testing allows one to "hide" this use of personal intuition. Thus, while Pratt (1965) admits that consideration of alternatives can be hard and a source of controversy in many situations dealt with by significance testing, he argues that

> "Computing a P-value runs the danger of hiding this real uncertainty and legitimate disagreement behind a screen of irrelevant precision."

As a final point, it has been extensively argued (cf. Hacking (1965)) that one can never really reject P^0 until one has something better, namely another model P^1 which is both "reasonable" and better supported by the data. In Example 30, for instance, the observation $x = 2$ is quite unlikely to occur under P_0, but it is equally unlikely to occur under P_1; thus if P_0 and P_1 are *known* to be the only possibilities, then $x = 2$ provides no evidence against P_0. Thus consideration of alternatives is imperative if one actually seeks to reject P^0.

4.4.4 Conclusions

What is to be concluded about significance testing? First of all, it should be admitted that, as the significance level (or P-value) decreases, the evidence against H_0 will be increasing (assuming that T has been chosen appropriately). Indeed, in a few special situations (primarily one-sided testing situations) the significance level can correspond to a reasonable conditional (Bayesian) measure of the validity of H_0 (cf. Jeffreys (1961), Pratt (1965), DeGroot (1973), Fraser and Mackay (1976), Dickey (1977), Zellner (1982), and Casella and Berger (1987)). In general though, the magnitude of a significance level need bear no relationship (from problem to problem) to the actual amount of evidence against H_0, and significance levels in testing precise hypotheses are typically so misleadingly small that their use for actually rejecting a hypothesis is strongly contraindicated.

Although a given significance level can mean vastly different things in different situations, it can be argued that, through frequent use in various situations, insight into its true strength of evidence against H_0 can be obtained. This is perhaps true: capable people can become very good at doing tasks with grossly inadequate tools. This is not to say, however, that better tools should be ignored or, more importantly, that inexperienced people will do well with the inadequate tools.

One possibly valid use of significance testing is to provide an alert that further investigation (in particular consideration of alternatives) is needed. As Barnard (1981) says

> "The question to be answered is whether the feature (T(x)) presented is so improbable on H_0 as to justify the effort involved in exercising our imagination to produce an hypothesis that could account for it."

There is no guarantee from a small significance level that P^0 is wrong (i.e., that an alternative hypothesis can be found which is substantially more supported by the data), but without a small significance level there may be no need to look past P^0. This use of significance testing can be argued to be important even to Bayesians, as extensively discussed in Box (1980): for a given model and prior, the marginal (or predictive) density of X can be used to conduct a significance test which could alert one to question the model or prior.

Of course, even this use of significance testing as an alert could be questioned, because of the matter of averaging over unobserved x. It is hard to see what else could be done with P^0 *alone*, however, and it is sometimes argued that time constraints preclude consideration of alternatives. This may occasionally be true, but is probably fairly rare. Even cursory consideration of alternatives and a few rough likelihood ratio calculations will tend to give substantially more insight than will a significance level, and will usually not be much more difficult than sensibly choosing T and calculating the significance level. (See also Dempster (1974b).) Admittedly, such an approach will be somewhat imprecise, but what is the advantage of "irrelevant precision"?

4.5 RANDOMIZATION ANALYSIS

4.5.1 Introduction

In classical finite population sampling (or survey sampling) and randomization testing, the randomization in the experimental design (used to select the sample or allocate treatments) is a dominant factor in the construction of measures of evidence about θ. These measures are pre-experimental in nature, and their use directly violates the LP and RLP. (The outcome of the randomization is usually known, and hence averaging over samples or treatment allocations that might have occurred is supposedly irrelevant.) Hence, belief in the LP would have a profound effect on one's view of these areas of statistics.

Perhaps not surprisingly, it is in these areas, so drastically affected by the LP, that some of the strongest intuitive arguments against the LP can be raised. The issues involved are very complex, so much so that all we can hope to do is skim the surface of the subject. Indeed, we will essentially restrict ourselves to a defence of the LP in a few simple examples, trying to establish, as plausible, the argument that anything sensible in randomization analysis is sensible precisely because it has a sensible interpretation from a likelihood viewpoint.

Although our main emphasis will not be on criticizing randomization analysis, it is important to keep several issues in mind. First, randomization analysis can clearly be very silly conditionally, if followed blindly. Even proper randomization can result (by bad luck) in treatment groups unbalanced with respect to unanticipated (but observed to be important) covariates, or in a sample which is clearly unrepresentative of the population, and yet classical randomization analysis does not treat such situations any differently than situations where the outcome of the randomization is "good". Thus, if one randomly samples from the population of voters in a survey on preference in the next Presidential election and finds that, unfortunately, all members of the random sample happen to be Republican, it is permissible (classically) to ignore this fact and proceed with the usual analysis. A second problem with randomization analysis (or at least randomization testing) is that it is often implemented through significance testing, and the serious concerns of the previous section then apply. The third, and most important, problem is that randomization analysis *does violate* the LP. In murky situations, where intuition stumbles, it seems especially necessary to depend on foundations.

Because of the above (and various specific) criticisms of randomization analysis, such analysis is usually advanced, not as an always sound way of proceeding, but as the most useful practical method of obtaining a reasonable answer. We will try to argue that the case for this is weak, at best.

Of course, even though we argue that the basis of randomization analysis is fundamentally in error, many of the specific procedures used in survey sampling and randomization testing are perfectly satisfactory. (If so, however, it is probably because they have some sensible interpretation consistent with the LP.) Also, the value of randomization itself, in treatment allocation and the choice of a sample, is not being addressed here.

Such randomization is often argued to be valuable (even by many conditionalists) in helping to reduce systematic effects that perhaps might unwittingly be introduced by the experimental design or sampling plan. An experiment in which randomization is used properly will, most of the time, turn out to be reasonably balanced with respect to experimentally induced (and unanticipated) covariates. Randomization also helps greatly in convincing others, who do not have access to the experimental setup or data, that no systematic biases were present. Employing measures of evidence based on the randomization probabilities is an entirely different matter, however. Indeed, a conditionalist will not only ignore the randomization probabilities, since the outcome of the randomization is known, but will also check to see that balance with respect to important new covariates was indeed obtained.

4.5.2 Finite Population Sampling

A typical classical setup is that of having a population $\mathcal{Y} = \{y_1,\ldots,y_N\}$ of N units, where each unit y_i can be represented as a vector $y_i = (u_i,v_i)$, u_i representing a label (or other known information) about the unit and v_i representing something unknown (but observable). It is desired to infer something about $\theta = (v_1,\ldots,v_N) \in \Theta$, from a sample $\mathcal{Y}_s = (y_{i_1}, y_{i_2},\ldots,y_{i_m})$, which is a subset of $\mathcal{Y}$; here $s = \{i_1,\ldots,i_m\} \subset I = \{1,\ldots,N\}$ indicates which units from the population are selected to be part of the sample. Note that it is typically also possible to use the known labels $\underset{\sim}{u} = (u_1,\ldots,u_N)$ in making inferences about θ. Let $\mathcal{S}$ denote the collection of all subsets of I, and suppose P is a probability distribution on $\mathcal{S}$. A procedure $\delta(\mathcal{Y}_s,\underset{\sim}{u})$ is to be used, and some criterion function $L(\delta(\mathcal{Y}_s,\underset{\sim}{u}),\theta)$ employed. Finally, the overall statistical procedure (P,δ), by which it is meant that s will be chosen according to the probability distribution P on $\mathcal{S}$ and $\delta(\mathcal{Y}_s,\underset{\sim}{u})$ will be used, is evaluated classically by the frequentist measure of performance

$$R(P,\delta,\theta) = \sum_{s\in\mathcal{S}} L(\delta(\mathcal{Y}_s,\underset{\sim}{u}),\theta)P(s).$$

EXAMPLE 31. Suppose it is desired to estimate the population total $\lambda = \sum_{i=1}^{N} v_i$, using an estimator $\delta(\mathcal{Y}_s,\underset{\sim}{u})$ and squared error loss $L(\delta(\mathcal{Y}_s,\underset{\sim}{u}),\theta) = (\delta(\mathcal{Y}_s,\underset{\sim}{u})-\lambda)^2$. Suppose $P(s) = 1/\binom{N}{n}$ for all $s \in \mathcal{S}$ of size n, corresponding to selection of a simple random sample of size n. The estimator

$$(4.5.1) \qquad \delta(\mathcal{Y}_s,\underset{\sim}{u}) = \frac{N}{n}\sum_{j=1}^{n} v_{i_j}$$

(recall that $\mathcal{Y}_s = ((u_{i_1},v_{i_1}),\ldots,(u_{i_n},v_{i_n}))$) is unbiased in the sense that

$$\sum_{s \in \mathcal{S}} \delta(\mathcal{Y}_s,\underset{\sim}{u})P(s) = \lambda,$$

and hence $R(P,\delta,\theta)$ can be considered to be the variance of the procedure (P,δ), were it repeatedly used.

To investigate this situation from the viewpoint of likelihood, note that the only randomness here is in the generation of s, and hence that

$$(4.5.2) \qquad P_\theta(\mathcal{Y}_s) = \begin{cases} P(s) & \text{if } \mathcal{Y}_s \in \Omega_\theta \\ 0 & \text{otherwise,} \end{cases}$$

where Ω_θ is the set of all possible vectors, $\mathcal{Y}_s$, which could arise as samples for the given $\underset{\sim}{u}$ and θ. (Note that the implicit sample space is the union, over all θ, of such Ω_θ.) Thus the likelihood function for θ, when $\mathcal{Y}_s$ is observed, is simply

$$(4.5.3) \qquad \ell(\theta) = P(s)I_{\Lambda(\mathcal{Y}_s)}(\theta),$$

where (for $\mathcal{Y}_s = ((u_{i_1},v_{i_1}),\ldots,(u_{i_m},v_{i_m}))$)

$$\Lambda(\mathcal{Y}_s) = \{\theta \in \Theta\colon \text{ for } j = 1,\ldots,m, \text{ the } i_j \text{ component of } \theta \text{ equals } v_{i_j}\}.$$

Since $\ell(\theta)$ is constant for $\theta \in \Lambda(\mathcal{Y}_s)$, it conveys no information about θ, other than that the part of θ observed (in $\mathcal{Y}_s$) is known. This is deemed by some to be a failure of the LP, in that the statistical procedure is thought to provide considerable information about that part of θ not observed in $\mathcal{Y}_s$, call it θ^*.

The likelihood (or maybe Bayesian) view is indeed, that the data contains no inherent information about θ^*, and that the only way of infering anything about θ^* is to relate it somehow to the observed sample. Various relationships which might be deemed reasonable are:

(i) All v_i are thought to be similar, and the labels u_i contain no information. Of the many ways to model this, a simple (often too simple) possibility is to presume that the v_i are independent observations from a $\mathcal{N}(\mu,\tau^2)$ distribution. Then estimate μ and τ^2, using the sample $\mathcal{Y}_s$, and infer whatever is desired about θ. In the situation of Example 31, the answers would be essentially the same as the classical answers.

(ii) Suppose the v_i are thought to be linearly related to the u_i, say

$$v_i = \alpha + \beta u_i + \varepsilon_i,$$

where the ε_i are presumed to have some distribution. Clearly a quite different analysis would be appropriate.

(iii) Suppose two distinct similar groups within the population can be identified from $\mathcal{Y}_s$. Knowledge about each group can be obtained from $\mathcal{Y}_s$, as in (i), and the proportion of each group in the population estimated. (Of course, a stratified sample would probably have been desirable had the groups been identifiable solely from the labels.)

(iv) Suppose it is felt that the sample does not look typical of the remainder of the population. (An unlucky sample was drawn, or the sample revealed an unanticipated bias in the sampling plan.) It is not clear what to do, but it certainly cannot be right to proceed with a classical analysis, as if the sample was satisfactory.

In the situations above, classical sampling theorists would, of course, recommend different procedures for the various presumed models. The point of the discussion is to indicate that the data, $\mathcal{Y}_s$, really doesn't say anything about θ^*, unless there is some background information relating the data to the population. It might be argued that, even when nothing is known about the population, a simple random sample will probably produce a representative subset of the population, so that an estimator such as (4.5.1) is reasonable for the population total. We do not disagree, but judge that (4.5.1) is then reasonable precisely because the sample is thought to be representative, in which case (4.5.1) would be justifiable from a variety of Bayesian arguments. The randomization *may* help to convince one that the sample is representative, but, once convinced of that fact, there is no further need to consider the sample selection probabilities.

Modeling the population is often called the superpopulation approach to survey sampling. Although we have presented it as Bayesian in nature, the modeling of the population can also be argued to be as "objective" as any modeling usually done in statistics (cf. the discussion by Royall of Basu (1971)), in which case one can argue that a directly meaningful likelihood function for the superpopulation parameters will exist. To a Bayesian, the choice of a model is just part of the prior specification (and often the most important and uncertain part), so the distinction seems unnecessary.

This discussion has assumed that the selection probabilities, P(s), are known. If they are partially unknown and depend on θ or on an informative nuisance parameter (see Section 3.5) they could be relevant to conclusions about θ. Rubin (1984) addresses this issue, distinguishing between "ignorable" and "nonignorable" sample selection mechanisms, and raises the related point that the P(s) may be useful as crude covariates in certain situations of stratified sampling.

Another issue that has been raised is the possibility of involving the P(s) by purposely ignoring the randomization outcome. Indeed, Rao (1971) argues that one can obtain an "informative" likelihood function by ignoring the labels u_i in the sample $\mathcal{Y}_s$. The available data is then only $\underset{\sim}{v}$, an m-vector of the observed v_i, with no record of which elements of the population it is associated with. It is easy to calculate, using (4.5.2) and

(4.5.3), that the likelihood function corresponding to $\underset{\sim}{v}$ is

$$(4.5.4) \qquad \ell(\theta) = \sum_{\text{all } \mathcal{Y}_s \text{ of size } m} P(s) I_{\Lambda(\mathcal{Y}_s)}(\theta).$$

This likelihood function may seem to contain more information about θ. In Example 31, for instance, it is easy to see that, if N/m is an integer, the M.L.E. for θ is any vector containing N/m copies of $\underset{\sim}{v}$. The M.L.E. for λ would thus be (4.5.1).

In discussing the reasonableness of the above proposal, it is important to first note that ignoring data is often a sensible practical necessity, as the following example indicates.

EXAMPLE 32. Suppose we observe (X,Y) having a joint density $f(x|\theta)g(y|\theta)$ (i.e., X and Y are independent), but that f is known while g is completely unknown. If we have very little prior information about g, so little that y conveys no clear knowledge about θ, then basing the analysis on x alone seems reasonable. Of course, ignoring y can be viewed as a formal violation of the LP, since it essentially involves integrating y out of the joint density of X and Y. It is not a violation of the spirit of the LP, however, providing $\ell_X(\theta) = f(x|\theta)$ is felt to be reasonably close to what would have been obtained were y included (say, by putting a prior distribution on g and integrating out over this prior). Further discussion and references on this issue can be found in Pratt (1965), who calls X an "insufficient statistic," and in Berger (1983).

While ignoring data may often be a practical necessity, there is a crucial difference between doing so in Example 32 and doing so in the sample survey problem. In Example 32 an *unknown* element g was eliminated by ignoring data, while Rao (1971) suggests replacing the *known* likelihood function in (4.5.3) by the version in (4.5.4) that would result if the labels in s were ignored. No real simplification is involved in the latter situation; indeed (4.5.4) seems more complicated than (4.5.3). In some situations a non-Bayesian likelihood analysis of (4.5.4) may seem easier than a similar analysis of (4.5.3), but such is probably only the case in simple situations like that of Example 31 where P(s) is constant, (and then direct reasoning of a model construction or Bayesian nature with (4.5.3) is also easy). And it is easy to construct examples where the use of (4.5.4) with highly variable P(s) can give completely unreasonable answers for particular observed $\mathcal{Y}_s$.

We have barely touched the surface of survey sampling. Deeper discussions of these issues and other references can be found in Godambe (1966, 1982a, 1982b), Cornfield (1969), Basu (1969, 1971, 1978), Ericson (1969), Kalbfleisch and Sprott (1969), Rao (1971), Royall (1971, 1976), Godambe and Thompson (1976), Smith (1976), Cassel et. al. (1977), and Thompson (1980). A particularly convincing case for the Bayesian view can be found in Basu (1978).

4.5.3 Randomization Testing

Randomization testing was introduced by Fisher (cf. Fisher (1960)) and was further developed by Kempthorne and others. (See Kempthorne and Folkes (1971) and Basu (1980) for some of these developments and other references). The basis of randomization testing is using the randomization mechanism involved in treatment allocation to experimental units to form probability assessments of evidence. The following simple example exhibits the key features of the approach. See Basu (1980), and the discussants thereof, for a more general discussion.

EXAMPLE 33. In an experiment, n independent pairs of matched subjects $\{(S_1^0,S_1^1),\ldots,(S_n^0,S_n^1)\}$ are to be utilized to compare two treatments, T_0 (the "standard") and T_1 (the "new treatment"). Within each pair, the two treatments are randomly assigned: let r_i equal 0 or 1 as treatment T_0 or T_1, respectively, is assigned to S_i^0 (so that treatment $T_{(1-r_i)}$ is assigned to S_i^1), and define $\underset{\sim}{r} = (r_1,\ldots,r_n)$. Note that $P(r_i = 0) = P(r_i = 1) = \frac{1}{2}$. The result of the experiment will be a vector $\underset{\sim}{X} = (X_1,\ldots,X_n)$, where, for the ith pair,

$$X_i = \begin{cases} 0 & \text{if } T_0 \text{ is judged to have worked better} \\ 1 & \text{if } T_1 \text{ is judged to have worked better.} \end{cases}$$

(For simplicity of discussion, we assume that equality of treatments is not a possible observation, and that only the crude measures X_i are observable.)

Randomization testing, here, would involve consideration of the hypothesis (H_0) that the treatments have an identical effect, in the sense that a given subject in each pair, say subject $S_i^{\delta_i}$ ($\delta_i = 0$ or 1), would do best no matter which treatment it received. It is easy to check that H_0 can be written mathematically as

$$(4.5.5) \qquad H_0\colon\ X_i = (r_i+\delta_i)_{\mathrm{mod}\ 2} \quad \text{for } i = 1,\ldots,n.$$

Also, letting $\underset{\sim}{\delta} = (\delta_1,\ldots,\delta_n)$, it is clear that, *pre-experimentally*, $\underset{\sim}{X}$ has density (under H_0) $f_{\underset{\sim}{\delta}}(\underset{\sim}{x}) = 2^{-n}$ (since there is only one assignment $\underset{\sim}{r}$ which will match $\underset{\sim}{x}$ to $\underset{\sim}{\delta}$, and each $\underset{\sim}{r}$ has probability 2^{-n} of occurring).

Suppose that it is desired to perform a significance test of H_0 against the one-sided alternative that T_1 is a better treatment than T_0. The natural test statistic would be $X = \sum_{i=1}^{n} X_i$, with large values of X providing evidence against H_0. The significance level (or P-value) of an observation, $\underset{\sim}{x}$, would then be

$$\alpha = P_{H_0,\underset{\sim}{\delta}}(X \geq x = \sum_{i=1}^{n} x_i) = \sum_{j=x}^{n} \binom{n}{j} 2^{-n}.$$

If, for example, all $x_i = 1$, then $\alpha = 2^{-n}$ which, for large n, would seem to cast doubt on H_0.

The pre-experimental measure of evidence, α, in the above example is based on the randomization probabilities. Since the actual randomization outcome $\underset{\sim}{r}$ becomes known, however, conditional reasoning would argue that such probabilities are irrelevant. A conditional analysis of the problem might go as follows.

EXAMPLE 33 (continued). Because of the pairing (and the randomization) it might be deemed reasonable to pretend that the subjects within each pair are identical. If the pairs can be considered to be a random sample from the entire population of pairs, and θ denotes the (hypothetical) proportion of the population for which treatment T_1 would be better than T_0, then one could write the joint density of $\underset{\sim}{x}$ and $\underset{\sim}{r}$ as

$$f_\theta(\underset{\sim}{x},\underset{\sim}{r}) = 2^{-n}\theta^x(1-\theta)^{n-x}.$$

A likelihood analysis could then be performed, based on this (binomial) likelihood for θ. (Of course, a significance test of $\theta = \frac{1}{2}$ would give the same result as the randomization analysis, and we will argue that this is really why the randomization analysis is, at all, sensible.)

The randomization mechanism plays no direct role in the above likelihood argument. Indeed, the use of randomization is limited to making more believable the assumption that the paired subjects are equivalent: the randomization hopefully eliminates the possibility of experimenter induced bias that might be introduced by, say, giving treatment T_0 to the subjects (perhaps subconsciously) thought to be healthiest. It might be argued, by some, that the classical randomization analysis seems intuitively more sensible than the modeled likelihood analysis. The following illustration of biased randomization (as discussed in Basu (1980)) casts doubt on the validity of such an argument.

EXAMPLE 33 (continued). Suppose the treatments are assigned by a randomization mechanism having the property that the subjects S_i^0 (independently) receive treatment T_0 with probability $\frac{1}{4}$ and treatment T_1 with probability $\frac{3}{4}$. Suppose, further, that the randomization outcome happens to be that each S_i^0 receives treatment T_0, and the experimental outcome happens to be that each $x_i = 1$. If the null hypothesis is true, then it must be the case that $\delta_i = 1$ for all i (see (4.5.5)). But it follows that the significance level against H_0 is

$$\alpha = P_{H_0,\underset{\sim}{\delta} = (1,\ldots,1)}(X \geq x = n) = P(\text{all } r_i = 0) = 4^{-n}.$$

This significance level seems misleadingly low, due to the "unlikely" randomization outcome. The evidence against H_0 certainly seems no stronger than it would have been had an unbiased randomizer been used. The modeled likelihood analysis would, of course, be unaffected by the use of the biased randomizer. Thus it seems that the randomization analysis may be rather suspect, unless it corresponds to a sensible modeled likelihood analysis.

As with finite population sampling, the likelihood approach tends to involve further modeling of the situation under investigation. While to some extent unappealing (more assumptions must be introduced), there seems to be little choice. In Example 33, if one were not comfortable in treating the subjects within a pair as identical, or the pairs as representative of the population, then the randomization analysis would also be very suspect. (If it so happened that a certain subject in each pair could be identified as "healthier", a careful investigation of the matchups of treatments and subjects would be indicated.) Extensive discussions of these issues can be found (with other references) in Savage et. al. (1962), Hill (1970), Good (1976), Rubin (1978), Lindley and Novick (1981), and especially Basu (1980).

CHAPTER 5. IMPLEMENTATION OF THE LIKELIHOOD PRINCIPLE

5.1 INTRODUCTION

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

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

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

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

This is certainly overstating the case, somewhat, in that, at the very least, foundational arguments can be invaluable in giving direction to our efforts. Thus the "consistency" theory strongly suggests that truth lies in a Bayesian direction, and the LP strongly suggests that truth lies in the direction of methods based on determination and utilization of the likelihood function (for the observed x). Luckily (or inevitably) these two directions are compatible.

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

5.2 NON-BAYESIAN LIKELIHOOD METHODS

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

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

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

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

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

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

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

5.3 ARGUMENTS FOR BAYESIAN IMPLEMENTATION

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

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

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

5.3.1 General Considerations

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

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

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

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

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

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

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

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

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

In some situations, it is clearly imperative to determine and introduce π. One such situation is that of Section 4.5.2, in which the likelihood function itself conveys almost no information unless θ is severely restricted through π (i.e., a suitable model for the population is introduced). Indeed the nonparametric situation discussed in Section 3.6.1 is the general prototype for this situation, in that the likelihood function is very difficult to use unless θ is substantially restricted a priori, corresponding to proposing a model (or class of models) for the distribution of X. The "generalized inverse" problems discussed in Jaynes (1981) also have this same flavor.

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

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

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

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

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

The likelihood function is easily seen to be

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

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

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

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

Now

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

and

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

while, amazingly,

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

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

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

while

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

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

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

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

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

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

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

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

$$\theta = \text{N N E S W S W W}.$$

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

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

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

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

for each of the four possible θ.

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

To us, this example again serves to indicate that the likelihood function can really be utilized only through Bayesian analysis. For instance, forget for a moment the amusing structure of Stone's example, and just consider the statistical problem involving θ and X. Suppose θ was in actuality generated according to the (prior) distribution

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

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

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

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

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

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

(for the θ of length m+1) and

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

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

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

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

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

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

(The soldier trails an *elastic* string, and after burying the treasure at the end of $\theta \neq x_0$ he passes out and has a 75% chance of being snapped back to 0; the end of x_0, however, is very slippery, so if the soldier buries the treasure there and passes out he will be snapped back to 0 for sure. There also happens to be a good samaritan who walks the streets within one block of a shelter at x_0, and if the soldier passes out at x_0N, x_0S, x_0W, or x_0E and doesn't get snapped back to 0, the good samaritan will take him back to x_0.) Suppose now that the observation from this model just happens to be x_0, so that the likelihood function for θ is the same as that obtained from Stone's model for the observation x_0. The LP says that the conclusions in each case should be the same, and we concur. Since θ is still the path generated by the drunken soldier, the prior defined by $\pi(\theta) = p_n/N_n$ is still appropriate, and the resulting Bayesian analysis sensible. (Alternatively, if θ had been generated in such a way that the prior was felt to be locally uniform near x_0 - note that any proper prior could only be locally uniform near some of the possible observations - the Bayesian analysis with $\pi(\theta) = K$ would be appropriate.)

This Bayesian reasoning is in conflict with frequentist reasoning, which states in the situation of Stone that, conditional on θ, X is three times as likely to extend the path as to backtrack, while in the situation of Fraser there is no reason to think this. Such reasoning seems to be the basis of the claim by Fraser (in the discussion of Hill (1981)) that the situation provides a counterexample to the LP. To us it instead provides yet another counterexample to frequentist reasoning. If there is doubt as to this, imagine that θ really was generated according to one of the priors considered here (and a compelling case can be made that the drunken soldier actually does generate θ according to $\pi(\theta) = p_n/N_n$), in which case there seems little doubt that the two models give the same answer. (See also Berger (1984a).)

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

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

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

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

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

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

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

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

$$(x-(1.96)\sigma,\ x+(1.96)\sigma) \tag{5.3.4}$$

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

$$(y-(1.96)[y/d],\ y+(1.96)[y/d]) \tag{5.3.5}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.4 ROBUST BAYESIAN ANALYSIS

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

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

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

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

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

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

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

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

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

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

A second reason for possible violation of the LP, as discussed in Section 4.1.3, is that many users of statistics will be unable to perform careful robust Bayesian analyses. For these users we must provide simple Bayesian procedures with "built in" robustness. In part, this robustness should be measured in a frequency sense, since the procedures will be used repeatedly (i.e., for different X). Of course, good conditional performance of these procedures should still be of paramount concern. Note, in particular, that the noninformative prior or "objective" Bayesian procedures are usually very good procedures from this perspective of use by nonspecialists, and may formally violate the LP through dependence of the noninformative prior on E (and not just the observed $\ell_x(\theta)$). In a similar vein, Hajék (1967, 1971) argues that asymptotic theory (which can provide useful simple procedures for nonspecialists in complicated situations) can sometimes be more difficult if one is restricted to basing it only on the given likelihood function, and not on E as a whole.

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

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

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

5.5 CONCLUSIONS

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

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

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

References

AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis.* Cambridge University Press, Cambridge.

AKAIKI, H. (1982). On the fallacy of the likelihood principle. *Statistics and Probability Letters 1,* 75-78.

AMARI, S. (1982). Geometrical theory of asymptotic ancillarity and conditional inference. *Biometrika 69,* 1-18.

ANDERSEN, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. Roy. Statist. Soc. B 32,* 283-301.

ANDERSEN, E. B. (1971). A strictly conditional approach in estimation theory. *Skand. Aktuarietidskr. 54,* 39-49.

ANDERSEN, E. B. (1973). *Conditional Inference and Models for Measuring.* Mentalhygiejnisk Forlag, Copenhagen.

ANSCOMBE, F. J. (1963). Sequential medical trials. *J. Amer. Statist. Assoc. 58,* 365-383.

ARMITAGE, P. (1961). Contribution to the discussion of C.A.B. Smith 'Consistency in statistical inference and decision'. *J. Roy. Statist. Soc. B 23,* 1-37.

ASPIN, A. A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated (with an appendix by B. L. Welch). *Biometrika 36,* 290-296.

BAHADUR, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist. 25,* 423-462.

BARNARD, G. A. (1947a). A review of 'Sequential Analysis' by Abraham Wald. *J. Amer. Statist. Assoc. 42,* 658-669.

BARNARD, G. A. (1947b). The meaning of significance level. *Biometrika 34,* 179-182.

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

BARNARD, G. A. (1962). Comments on Stein's 'A remark on the likelihood principle'. *J. Roy. Statist. Soc. A 125,* 569-573.

BARNARD, G. A. (1967a). The use of the likelihood function in statistical inference. In *Proc. 5th Berkeley Symp. on Math. Statist. and Prob.,* University of California Press, Berkeley.

BARNARD, G. A. (1967b). The Bayesian controversy in statistical inference. *J. Inst. Actuaries 93,* 229-269.

BARNARD, G. A. (1971). Scientific inferences and day-to-day decisions. In *Foundations of Statistical Inference,* V. P. Godambe and D. A. Sprott (eds.). Holt, Rinehart and Winston, Toronto.

BARNARD, G. A. (1974). On likelihood. In the *Proceedings of the Conference on Foundational Questions in Statistical Inference,* O. Barndorff-Nielsen, P. Blaesild, and G. Schou (eds.). Department of Theoretical Statistics, University of Aarhus.

BARNARD, G. A. (1975). Conditional inference is not inefficient. *Scandinavian J. of Statist. 3,* 132-134.

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

BARNARD, G. A. (1981). A coherent view of statistical inference. Presented at the Symposium on Statistical Inference and Applications, University of Waterloo, August, 1981.

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

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

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

BARNDORFF-NIELSEN, O. (1971). *On Conditional Statistical Inference.* Matematisk Institute, Aarhus University.

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

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

BARNDORFF-NIELSEN, O. (1980). Conditionality resolutions. *Biometrika 67,* 293-310.

BARNETT, V. (1982). *Comparative Statistical Inference (2nd Edition).* John Wiley and Sons, New York.

BARTHOLOMEW, D. J. (1967). Hypothesis testing when the sample size is treated as a random variable. *J. Roy. Statist. Soc. B 29,* 53-82.

BARTLETT, M. S. (1936). Statistical information and properties of sufficiency. *Proc. Royal Soc. A 154,* 124.

BARTLETT, M. S. (1953). Approximate confidence intervals, I, *Biometrika 40,* 13-19. II, *Biometrika 40,* 306-317.

BASU, D. (1964). Recovery of ancillary information. In *Contributions to Statistics,* C. R. Rao (ed.), 7-20. Pergamon Press, Oxford.

BASU, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā A 31*, 441-454.

BASU, D. (1971). An essay on the logical foundations of survey sampling, part one. In *Foundations of Statistical Inference,* V. P. Godambe, and D. A. Sprott (eds.). Holt, Rinehart and Winston, Toronto.

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

BASU, D. (1978). On the relevance of randomization in data analysis (with discussion). In *Survey Sampling and Measurement*, N. K. Namboodiri (ed.). Academic Press, New York.

BASU, D. (1980). Randomization analysis of experimental data: the Fisher randomization test. *J. Amer. Statist. Assoc. 75*, 575-595.

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

BERGER, J. (1984a). In defense of the likelihood principle: axiomatics and coherency. In *Bayesian Statistics II*, J. M. Bernardo, M. H. DeGroot, D. Lindley, and A. Smith (eds.).

BERGER, J. (1984b). Bayesian salesmanship. In *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno deFinetti*, P. K. Goel and A. Zellner (eds.). North-Holland, Amsterdam.

BERGER, J. (1984c). The frequentist viewpoint and conditioning. To appear in the *Proceedings of the Berkeley Conference in Honor of J. Kiefer and J. Neyman*, L. LeCam and R. Olshen (eds.). Wadsworth, Belmont California.

BERGER, J. (1984d). A review of J. Kiefer's work on conditional frequentist statistics. To appear in *The Collected Works of Jack Kiefer* (L. Brown, I. Olkin, J. Sacks, H. Wynn, eds.).

BERGER, J. (1984e). The robust Bayesian viewpoint (with discussion). In *Robustness in Bayesian Statistics*, J. Kadane (ed.), 63-144. North-Holland, Amsterdam.

BERGER, J. and BERLINER, L. M. (1983). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. Technical Report #83-35, Statistics Dept., Purdue Univ., W. Lafayette.

BERNARDO, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. B 41*, 113-147.

BIRNBAUM, A. (1961a). On the foundations of statistical inference: binary experiments. *Ann. Math. Statist. 32*, 414-435.

BIRNBAUM, A. (1961b). Confidence curves: an omnibus technique for estimation and testing statistical hypotheses. *J. Amer. Statist. Assoc. 56,* 246-249.

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

BIRNBAUM, A. (1962b). Intrinsic confidence methods. *Bulletin of the Int. Statist. Inst. 39,* 375-383.

BIRNBAUM, A. (1968). Likelihood. In *International Encyclopedia of the Social Sciences, Vol. 9.*

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

BIRNBAUM, A. (1970a). Statistical methods in scientific inference. *Nature 225,* 1033.

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

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

BIRNBAUM, A. (1977). The Neyman-Pearson theory as decision theory and as inference theory: with a criticism of the Lindley-Savage argument for Bayesian theory. *Synthese 36,* 19-49.

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

BONDAR, J. V. and MILNES, P. (1981). Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups. *Zeitschr. Wahrsch. Verw. Geb. 57,* 103-128.

BOX, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. B 143,* 383-430.

BOX, G.E.P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading.

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

BROWN, L. D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist. 6,* 59-71.

BROWNIE, C. and KIEFER, J. (1977). The ideas of conditional confidence in the simplest setting. *Commun. Statist. A6(8),* 691-751.

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

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

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

BUEHLER, R. J. (1982). Some ancillary statistics and their properties (with discussion). *J. Amer. Statist. Assoc. 77,* 581-594.

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

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

CASELLA, G. and HWANG, J. T. (1982). Zero-radius confidence procedures. Technical Report, Cornell University.

CASSEL, C. M., SÄRNDAL, C. E., and WRETMAN, J. H. (1977). *Foundations of Inference in Survey Sampling.* Wiley, New York.

CORNFIELD, J. (1966). Sequential trials, sequential analysis, and the likelihood principle. *The American Statist. 20,* No. 2, 18-23.

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

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

COX, D. R. (1971). The choice between ancillary statistics. *J. Roy. Statist. Soc. B 33,* 251-255.

COX, D. R. (1975). Partial likelihood. *Biometrika 62,* 269-276.

COX, D. R. (1977). The role of significance tests. *Scand. J. Statist. 4,* 49-70.

COX, D. R. (1978). Foundations of statistical inference: the case for eclecticism. *Austral. J. Statist. 20,* 43-59.

COX, D. R. (1980). Local ancillarity. *Biometrika 67,* 279-286.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

DAWID, A. P. (1975). On the concepts of sufficiency and ancillarity in the prescence of nuisance parameters. *J. Roy. Statist. Soc. B 37,* 248-258.

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

DAWID, A. P. (1980). A Bayesian look at nuisance parameters. In *Bayesian Statistics,* J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A.F.M. Smith (eds.). University Press, Valencia.

DAWID, A. P. (1981). Statistical inference. In *Encyclopedia of Statistical Sciences,* S. Kotz and N. L. Johnson (eds.). Wiley, New York.

DAWID, A. P. and DICKEY, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Amer. Statist. Assoc. 72,* 845-850.

DAWID, A. P. and STONE, M. (1982). The functional-model basis of fiducial inference (with discussion). *Ann. Statist. 10,* 1054-1074.

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

DE FINETTI, B. (1972). *Probability, Induction, and Statistics.* Wiley, New York.

DE FINETTI, B. (1974). *Theory of Probability, Volumes 1 and 2.* Wiley, New York.

DE GROOT, M. H. (1973). Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. *J. Amer. Statist. Assoc. 68,* 966-969.

DEMPSTER, A. P. (1974a). Remarks on inference. In the *Proceedings of the Conference on Foundational Questions in Statistical Inference,* O. Barndorff-Nielsen, P. Blaesild, and G. Schou (eds.). Department of Theoretical Statistics, University of Aarhus.

DEMPSTER, A. P. (1974b). The direct use of likelihood for significance testing. In the *Proceedings of the Conference on Foundational Questions in Statistical Inference,* O. Barndorff-Nielsen, P. Blaesild, and G. Schou (eds.). Department of Theoretical Statistics, University of Aarhus.

DEMPSTER, A. P. (1975). A subjectivist look at robustness. *Bull. of the International Statist. Inst. 46,* 349-374.

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

DICKEY, J. M. (1977). Is the tail area useful as an approximate Bayes factor? *J. Amer. Statist. Assoc. 72,* 138-142.

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

EDWARDS, A.W.F. (1972). *Likelihood.* C.U.P., Cambridge.

EDWARDS, A.W.F. (1974). The history of likelihood. *Int. Statist. Rev. 42,* 9-15.

EDWARDS, W., LINDMAN, H., and SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review 70,* 193-242.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika 65,* 457-482.

ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations. *J. Roy. Statist. Soc. B 31,* 195-233.

FISHER, R. A. (1921). On the 'Probable Error' of a coefficient of correlation deduced from a small sample. *Metron. I,* part 4, 3-32.

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Phil. Soc. 22,* 700-725.

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Royal Soc. A 144,* 285-307.

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

FISHER, R. A. (1956b). On a test of significance in Pearson's Biometrika Tables (no. 11). *J. Roy. Statist. Soc. B 18,* 56-60.

FISHER, R. A. (1960). *The Design of Experiments (7th ed.).* Oliver and Boyd, Edinburgh.

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

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

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

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

FRASER, D.A.S. (1976). Necessary analysis and adaptive inference (with discussion). *J. Amer. Statist. Assoc. 71,* 99-113.

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

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

FRASER, D.A.S. and MACKAY, J. (1976). On the equivalence of standard inference procedures. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II,* W. L. Harper and C. A. Hooker (eds.). Reidel, Dordrecht.

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

GEISSER, S. (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott (eds.). Holt, Rinehart, and Winston, Toronto.

GIERE, R. N. (1977). Allan Birnbaum's conception of statistical evidence. *Synthese 36*, 5-13.

GODAMBE, V. P. (1966). A new approach to sampling from finite populations, I. *J. Roy. Statist. Soc. B 28*, 310-319.

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

GODAMBE, V. P. (1982a). Likelihood principle and randomization. In *Statistics and Probability: Essays in Honor of C. R. Rao*, G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh (eds.). North-Holland, Amsterdam.

GODAMBE, V. P. (1982b). Estimation in survey sampling: robustness and optimality. *J. Amer. Statist. Assoc. 77*, 393-406.

GODAMBE, V. P. and THOMPSON, M. E. (1976). Philosophy of survey-sampling practice. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II*, W. L. Harper and C. A. Hooker (eds.). Reidel, Dordrecht.

GODAMBE, V. P. and THOMPSON, M. E. (1977). Robust near optimal estimation in survey practice. *Bulletin of the Internat. Statist. Inst. 47*, 127-170.

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

GOOD, I. J. (1965). *The Estimation of Probabilities*. M. I. T. Press, Cambridge.

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

GOOD, I. J. (1981). Some logic and history of hypothesis testing. In *Philosophy in Economics*, J. C. Pitt (ed.). Reidel, Dordrecht.

GRAMBSCH, P. (1980). Likelihood inference. Ph.D. Dissertation, University of Minnesota.

HACKING, I. (1965). *Logic of Statistical Inference.* Cambridge University Press, Cambridge.

HAJEK, J. (1967). On basic concepts of statistics. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* LeCam and J. Neyman (eds.). University of California Press, Berkeley.

HAJEK, J. (1971). Limiting properties of likelihoods and inference. In *Foundations of Statistical Inference,* V. P. Godambe and D. A. Sprott (eds.). Holt, Rinehart, and Winston, Toronto.

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

HILL, B. (1970). Some contrasts between Bayesian and classical inference in the analysis of variance and in the testing of models. In *Bayesian Statistics,* D. L. Meyer and R. O. Collier, Jr. (eds.). Peacock Publishers, Itasca, Illinois.

HILL, B. (1973). Review of "Likelihood" by A.W.F. Edwards. *J. Amer. Statist. Assoc. 68,* 487-488.

HILL, B. (1974a). Review of"Bayesian Inference in Statistical Analysis" by G.E.P. Box and G. Tiao. *Technometrics 16,* 478-479.

HILL, B. (1974b). On coherence, inadmissibility, and inference about many parameters in the theory of least squares. In *Studies in Bayesian Econometrics and Statistics,* S. Fienberg and A. Zellner (eds.). North-Holland, Amsterdam.

HILL, B. (1975). Abberant behavior of the likelihood function in discrete cases. *J. Amer. Statist. Assoc. 70,* 717-719.

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

HINDE, J. and AITKIN, M. (1984). Nuisance parameters, canonical likelihoods, and direct likelihood inference. Technical Report, Centre for Applied Statistics, University of Lancaster.

HINKLEY, D. V. (1978). Likelihood inference about location and scale parameters. *Biometrika 65*, 253-262.

HINKLEY, D. V. (1979). Predictive likelihood. *Ann. Statist. 7*, 718-728.

HINKLEY, D. V. (1980a). Fisher's development of conditional inference. In *R. A. Fisher: An Appreciation*, S. E. Fienberg and D. V. Hinkley (eds.). Springer-Verlag, New York.

HINKLEY, D. V. (1980b). Likelihood as approximate pivotal distribution. *Biometrika 67*, 287-292.

HINKLEY, D. V. (1983). Can frequentist inferences be very wrong? A conditional 'yes'. In *Scientific Inference, Data Analysis, and Robustness*, G.E.P. Box, T. Leonard, and C. F. Wu (eds.). Academic Press, New York.

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

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

JAYNES, E. T. (1982). *Papers on Probability, Statistics, and Statistical Physics, a reprint collection.* D. Reidel, Dordrecht.

JEFFREYS, H. (1961). *Theory of Probability (3rd edn.).* Clarendon Press, Oxford.

JEFFREYS, H. (1973). *Scientific Inference (3rd edn.).* C.U.P., Cambridge.

JOSHI, V. M. (1976). A note on Birnbaum's theory of the likelihood principle. *J. Amer. Statist. Assoc. 71*, 345-346.

KALBFLEISCH, J. D. (1971). Likelihood methods of prediction. In *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott (eds.). Holt, Rinehart, and Winston, Toronto.

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

KALBFLEISCH, J. D. (1975). Sufficiency and conditionality. *Biometrika 62*, 251-268.

KALBFLEISCH, J. D. (1978). Likelihood methods and nonparametric tests. *J. Amer. Statist. Assoc. 73*, 167-170.

KALBFLEISCH, J. D. and SPROTT, D. A. (1969). Applications of likelihood and fiducial probability to sampling finite populations. In *New Developments in Survey-Sampling*. Wiley, New York.

KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. Roy. Statist. Soc. B 32*, 175-208.

KEMPTHORNE, O. and FOLKES, J. L. (1971). *Probability, Statistics, and Data Analysis*. Iowa State University Press, Ames.

KIEFER, J. (1975). Conditional confidence approach in multi-decision problems. In *Proceedings of the Fourth Dayton Multivariate Conference*, P. R. Krishnaiah (ed.). 143-158. North-Holland, Amsterdam.

KIEFER, J. (1976). Admissibility of conditional confidence procedures. *Ann. Math. Statist. 4*, 836-865.

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

KIEFER, J. (1977b). The foundations of statistics - are there any? *Synthese 36*, 161-176.

KIEFER, J. (1980). Conditional inference. In the *Encyclopedia of Statistics*, S. Kotz and N. Johnson (eds.). Wiley, New York.

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

LAURITZEN, S. L. (1974). Sufficiency, prediction, and extreme models. *Scand. J. Statist. 1*, 128-134.

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

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

LINDLEY, D. V. (1958). A survey of the foundations of statistics. *Appl. Statist. 7*, 186-198.

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

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

LINDLEY, D. V. and NOVICK, M. (1981). The role of exchangeability in inference. *Ann. Statist. 9*, 45-58.

LINDLEY, D. V. and PHILLIPS, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *American Statist. 30*, 112-119.

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

MORRISON, D. F. and HENKEL, R. E. (1970). *The Significance Test Controversy.* Aldine, Chicago.

NEYMAN, J. (1957). 'Inductive behavior' as a basic concept of philosophy of science. *Rev. Intl. Statist. Inst. 25*, 7-22.

NEYMAN, J. (1967). *A Selection of Early Statistical Papers of J. Neyman.* University of California Press, Berkeley.

NEYMAN, J. (1977). Frequentist probability and frequentist statistics. *Synthese 36*, 97-131.

OLSHEN, R. A. (1973). The conditional level of the F-test. *J. Amer. Statist. Assoc. 68*, 692-698.

PEDERSEN, J. G. (1978). Fiducial inference. *Int. Statist. Review 46*, 147-170.

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

PICCINATO, L. (1981). On the orderings of decision functions. *Symposia Mathematica XXV*. Academic Press, London.

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

PLACKETT, R. L. (1966). Current trends in statistical inference. *J. Roy. Statist. Soc. A 129*, 249-267.

PLANTE, A. (1971). Counter-examples and likelihood. In *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott (eds.). Holt, Rinehart, and Winston, Toronto.

PRATT, J. W. (1961). Review of Lehmann's Testing Statistical Hypotheses. *J. Amer. Statist. Assoc. 56*, 163-166.

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

PRATT, J. W. (1976). A discussion of the question : for what use are tests of hypotheses and tests of significance. *Commun. Statist.-Theor. Meth. A5*, 779-787.

PRATT, J. W. (1977). 'Decisions' as statistical evidence and Birnbaum's 'confidence concept'. *Synthese 36*, 59-69.

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. Graduate School of Business Administration, Harvard University.

RAO, C. R. (1971). Some aspects of statistical inference in problems of sampling from finite populations. In *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott (eds.). Holt, Rinehart, and Winston, Toronto.

ROBINSON, G. K. (1975). Some counterexamples to the theory of confidence intervals. *Biometrika 62*, 155-161.

ROBINSON, G. K. (1976). Properties of Student's t and of the Behrens-Fisher solution to the two means problem. *Ann. Statist. 5*, 963-971.

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

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

ROSENKRANTZ, R. D. (1977). *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science.* Reidel, Boston.

ROYALL, R. (1971). Linear regression models in finite population sampling theory. In *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott (eds.). Holt, Rinehart, and Winston, Toronto.

ROYALL, R. (1976). Likelihood functions in finite population sampling survey. *Biometrika 63*, 605-617.

RUBIN, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist. 6*, 34-58.

RUBIN, D. B. (1984). The use of probability scores in applied Bayesian inference. In *Bayesian Statistics II*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A.F.M. Smith (eds.).

SAVAGE, L. J. (1954). *The Foundations of Statistics.* John Wiley and Sons, New York.

SAVAGE, L. J. (1970). Comments on a weakened principle of conditionality. *J. Amer. Statist. Assoc. 65*, 399-401.

SAVAGE, L. J. (1976). On rereading R. A. Fisher (with discussion). *Ann. Statist. 4*, 441-500.

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

SEIDENFELD, T. (1979). *Philosophical Problems of Statistical Inference.* Reidel, Boston.

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

SMITH, T.M.F. (1976). The foundations of survey sampling: a review. *J. Roy. Statist. Soc. A 139*, 183-204.

SPROTT, D. A. (1973a). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika 60*, 457-465.

SPROTT, D. A. (1973b). Practical uses of the likelihood function. In *Inference and Decision*, University Press of Canada, Toronto.

SPROTT, D. A. (1975). Marginal and conditional sufficiency. *Biometrika 62*, 599-605.

SPROTT, D. A. and KALBFLEISCH, J. D. (1969). Examples of likelihoods and comparisons with point estimates and large sample approximations. *J. Amer. Statist. Assoc. 64*, 468-484.

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

STEIN, C. (1962). A remark on the likelihood principle. *J. Roy. Statist. Soc. A 125*, 565-568.

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

THOMPSON, M. E. (1980). Likelihood principle and randomization in survey sampling. Report 78-04, Dept. of Statistics, University of Waterloo.

TJUR, T. (1978). Statistical inference under the likelihood principle. Preprint 1, Institute of Mathematical Statistics, University of Copenhagen, Copenhagen.

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

WELCH, B. L. (1939). On confidence limits and sufficiency with particular reference to parameters of location. *Ann. Math. Statist. 10*, 58-69.

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

YATES, F. (1964). Fiducial probability, recognizable subsets, and Behrens' test. *Biometrics 20*, 343-360.

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

ZELLNER, A. (1982). Applications of Bayesian analysis in econometrics. Presented at the Institute of Statisticians International Conference on Practical Bayesian Statistics at St. John's College, Cambridge.

Additional or Updated References

BAYARRI, M. J., DEGROOT, M. H., and KADANE, J.B. (1987). What is the likelihood function? (With Discussion). In *Statistical Decision Theory and Related Topics IV*, Volume 1, S. S. Gupta and J. Berger (eds.). Springer-Verlag, New York.

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

BERGER, J. (1987). Robust Bayesian analysis: sensitivity to the prior. To appear in the *Proceedings of the Conference in Honor of I. J. Good*, K. Hinkelmann (ed.).

BERGER, J. and BERLINER, L. M. (1986; previously 1983). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *Ann. Statist. 14*, 461-486.

BERGER, J. and BERRY, D. (1987). The relevance of stopping rules in statistical inference. In *Statistical Decision Theory and Related Topics IV*, Volume 1, S. S. Gupta and J. Berger (eds.). Springer-Verlag, New York.

BERGER, J. and BERRY, D. (1988). Statistical analysis and the illusion of objectivity. To appear in *American Scientist.*

BERGER, J. and DELAMPADY, M. (1987). Testing precise hypotheses (with Discussion). *Statistical Science 2*, 317-352.

BERGER, J. and SELLKE, T. (1987). Testing a point null hypothesis: the irreconcilability of P-values and evidence (with Discussion). *J. Amer. Statist. Assoc. 82*, 112-139.

BERLINER, M. (1987). Discussion of Bayarri, DeGroot, and Kadane (1987). In *Statistical Decision Theory and Related Topics IV*, Volume 1, S. S. Gupta and J. Berger (eds.). Springer-Verlag, New York.

BHAVE, S. V. (1984). Two concepts of conditionality. *J. Statist. Plann. and Inf. 10*, 131-135.

BUTLER, R. W. (1986). Predictive likelihood inference with applications (with Discussion). *J. Roy. Statist. Soc. B 47*, 1-38.

BUTLER, R. W. (1987). A likely answer to 'what is the likelihood function?' Discussion of Bayarri et. al. in *Statistical Decision Theory and Related Topics IV*, Volume 1, S. S. Gupta and J. Berger (eds.). Springer-Verlag, New York.

CASELLA, G. and BERGER, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc. 82*, 106-111.

DAWID, A. P. (1986). Discussion of Evans, Fraser, and Monette (1986). *Canad. J. Statist. 14*, 196-197.

DELAMPADY, M. and BERGER, J. (1987). Lower bounds on posterior probabilities for multinomial and chi-squared tests. Technical Report #86-37, Department of Statistics, Purdue University, West Lafayette.

EVANS, M., FRASER, D. A. S., and MONETTE, G. (1985a). Mixtures, embeddings, and ancillarity. *Canad. J. Statist. 13*, 1-6.

EVANS, M., FRASER, D. A. S., and MONETTE, G. (1985b). Regularity conditions for statistical models. *Canad. J. Statist. 13*, 137-144.

EVANS, M., FRASER, D. A. S., and MONETTE, G. (1985c). On the role of principles in statistical inference. In *Statistical Theory and Data Analysis*, K. Matusita (ed.). North-Holland, Amsterdam.

EVANS, M., FRASER, D. A. S., and MONETTE, G. (1986). On principles and arguments to likelihood (with discussion). *Canad. J. Statist. 14*, 181-199.

FRASER, D. A. S. (1984). Structural models. In *Encyclopedia of Statistical Science*, Volume 6, N. L. Johnson and S. Kotz (eds.). Wiley, New York.

FRASER, D. A. S., MONETTE, G., and NG, K. W. (1984). Marginalization, likelihood, and structural models. In *Multivariate Analysis VI*, P. R. Krishnaiah (ed.). North-Holland, Amsterdam.

GOOD, I. J. (1984). Notes C140, C144, C199, C200, and C201. *J. Statist. Computation and Simulation 19*.

HALL, P. and SELINGER, B. (1986). Statistical significance: balancing evidence against doubt. *Australian J. of Statist. 28*, 354-370.

HARTIGAN, J. A. (1967). The likelihood and invariance principles. *Ann. Math. Statist. 3*, 533-539.

HILL, B. (1987a). On the validity of the likelihood principle. In *Statistical Decision Theory and Related Topics IV*, Volume 1, S. S. Gupta and J. Berger (eds.). Springer-Verlag, New York.

HILL, B. (1987b). The validity of the likelihood principle. *The American Statistician 41*, 95-100.

HINDE, J. and AITKIN, M. (1987, previously 1984). Canonical likelihood: a new likelihood treatment of nuisance parameters. *Biometrika 74*, 45-58.

JOSHI, V. M. (1983). Likelihood principle. In *Encycl. Statist. Sci. 4*, 644-647, S. Kotz and N. L. Johnson (eds.). Wiley, New York.

KAY, R. (1985). Partial likelihood. In *Encycl. Statist. Sci. 6*, 591-593. S. Kotz and N. L. Johnson (eds.). Wiley, New York.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika 44*, 187-192.

LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika 64*, 207-213.

POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika 64*, 191-199.

SHAFER, G. (1982). Lindley's paradox. *J. Amer. Statist. Assoc. 77*, 325-351.

SMITH, A. F. M. and SPIEGELHALTER, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. B 42*, 213-220.

ZELLNER, A. (1984). Posterior odds ratios for regression hypotheses: general considerations and some specific results. In *Basic Issues in Econometrics.* University of Chicago Press, Chicago.

Discussion: Auxiliary Parameters and Simple Likelihood Functions

By Professors M. J. Bayarri and M. H. DeGroot

(University of Valencia and Carnegie-Mellon University)

We are grateful to our friends Jim Berger and Robert Wolpert for giving us this opportunity to contribute to their valuable and comprehensive study of the likelihood principle. Our prior distribution was highly concentrated on their writing an excellent monograph, and the evidence provided by the data we now have confirms our prior opinion. Their treatment is careful and thoughtful (this means that we agree with them) and leaves little room for further discussion. Nevertheless, haremos todo lo posible; we will try.

Our comments will be restricted to the material in Section 3.5 pertaining to the construction of a likelihood function to be used in statistical problems involving nuisance variables, nuisance parameters, and future observations. We will use the notation x, y, w, ξ, and η to represent the same quantities as in Section 3.5.2. Here, x is the observation and all the other quantities are unobserved, y and w are regarded as variables, ξ and η are regarded as parameters, and y and ξ are of interest. It will be convenient for us to use the notation $f(x,y,w|\xi,\eta)$ rather than $f_{(\xi,\eta)}(x,y,w)$ to denote a conditional density.

The basic purpose of a likelihood function is to serve as a function that relates observed and unobserved quantities, and conveys all the relevant information provided by the observed data about the unobserved quantities. From the Bayesian point of view, which we shall adopt in this discussion, we are interested in finding $f(y,\xi|x)$. If the design of the experiment is not under consideration, we could simply wait until x is observed and then assess the density $f(y,\xi|x)$ directly. However, to guide our thinking and to help make our conclusions more convincing to others, we would typically introduce some structure into our learning process by writing $f(y,\xi|x)$ in the form

$$(1) \qquad f(y,\xi|x) \propto f(x|\xi)\ f(y|x,\xi)\ f(\xi).$$

If there is general agreement about the form of the densities $f(x|\xi)$ and $f(y|x,\xi)$, then these densities can be regarded as "given" and in the spirit of (3.5.1), a likelihood function could be defined as

$$(2) \qquad \ell_x(y,\xi) = f(x|\xi)\ f(y|x,\xi) = f(x,y|\xi).$$

In BDK we referred to this likelihood function as LF_{rv} because it is derived from the conditional density of the "variables" given the "parameters".

We regard the likelihood function (2) as unsuitable as a general definition because we do not believe there is a clear-cut distinction between unobserved variables and parameters. The form of (2) relies on the density $f(x,y|\xi)$ being given or agreed on. We can rewrite this density as

$$(3) \qquad f(x,y|\xi) = f(x|y,\xi)\ f(y|\xi),$$

and agreement about $f(x,y|\xi)$ is equivalent to agreement about both factors on the right-hand side of (3). In this case the likelihood function would be given by (2). However, it is possible that there is agreement about the form of $f(x|y,\xi)$ while the form of $f(y|\xi)$ is considered as highly subjective. In this case, a likelihood function for y and ξ based on the observation x would be simply

$$(4) \qquad \ell_x(y,\xi) = f(x|y,\xi).$$

Thus, when an experimenter reports to us some particular function of y and ξ which is his or her likelihood function based on the observed data, we would still need further information from the experimenter in order to be able to make inferences or calculate posterior distributions. We must know whether or not the density $f(y|\xi)$ has been included in the likelihood function. In other words, in order to be able to use this likelihood function, we must know not only the function itself but also which factors have been used to derive it. (We will argue later in this discussion that it is unnecessary ever to include the factor $f(y|\xi)$ in the likelihood function in order to convey the evidence provided by the data x, since this factor does not involve x.)

It should be noted that the factors on the right-hand side of (2) contain only the variables y and parameters ξ that are of interest. In many problems, the densities $f(x|\xi)$ and $f(y|x,\xi)$ can still be difficult to specify or can still be considered highly subjective by others. These difficulties are usually reduced by introducing further structure into the learning process by means of a more detailed specification of the "parameter space" of ξ and the "sample space" of y. These specifications are represented by a "nuisance parameter" η and a "nuisance variable" w. As a result, (1) now becomes

$$(5) \qquad f(y,\xi|x) \propto \int\int f(x|\xi,\eta)\ f(y,w|x,\xi,\eta)\ f(\xi,\eta)\ dw\ d\eta.$$

It should be emphasized that η and w are selected by us for our convenience. If we have been successful in our selection, then there will be general agreement among others on the form of $f(x|\xi,\eta)$ and $f(y,w|x,\xi,\eta)$. It is presumably because of such agreement that Berger and Wolpert regard these densities as being "given", and define a likelihood function (3.5.1) to be their product

$$(6) \qquad \begin{aligned} \ell_x(y,\omega,\xi,\eta) &= f(x|\xi,\eta)\ f(y,w|x,\xi,\eta) \\ &= f(x,y,w|\xi,\eta). \end{aligned}$$

In view of these comments, the traditional expression "nuisance parameter" for η seems inappropriate. Because it helps us to build models and to achieve agreement about those models, η might better be called an auxiliary parameter.

When x is a vector of observations, one typical way in which a convenient choice of the auxiliary parameter η can simplify the density $f(x|\xi,\eta)$ is making the components of x conditionally independent. More importantly, a convenient choice of η may make y and w conditionally independent of x given ξ and η. In this case, $f(y,w|x,\xi,\eta)$ reduces to $f(y,w|\xi,\eta)$ and the likelihood function (6) becomes

$$(7) \qquad \ell_x(y,w,\xi,\eta) = f(x|\xi,\eta)\, f(y,w|\xi,\eta).$$

Regardless of whether the density $f(y,w|\xi,\eta)$ is given or subjective, it does not involve the data x, so all the evidence in x about the unknowns is contained in the first factor $f(x|\xi,\eta)$ on the right-hand side of (7). Thus, we believe that it is the only factor that should be included in the likelihood function. The inclusion of other functions of the unknowns, such as $f(y,w|\xi,\eta)$ or the prior $f(\xi,\eta)$, which do not depend on the data, seems artificial.

It should be noted that the likelihood function that we are recommending in this situation, namely

$$(8) \qquad \ell_x(y,w,\xi,\eta) = f(x|\xi,\eta),$$

can also be expressed because of conditional independence as

$$(9) \qquad \ell_x(y,w,\xi,\eta) = f(x|y,w,\xi,\eta).$$

In other words, this likelihood function is simply the conditional density of the observations given the unobserved quantities. In BDK this likelihood function was called LF_{obs} and is in accord with the basic definition (3.1.1) given by Berger and Wolpert.

More generally, every Bayesian analysis proceeds from a specification of the joint density $f(x,y,w,\xi,\eta)$. If we let s denote the set $\{x,y,w,\xi,\eta\}$ of all the components of all the quantities considered in the problem, and let s_1 and s_2 denote non-empty subsets of s such that $s_1 \cap s_2 = \emptyset$ and $s_1 \cup s_2 = s$, then the joint density f(s) can be expressed as the product $f(s) = f(s_1|s_2) \times f(s_2)$. The various likelihood functions under consideration in this discussion are of the form $f(s_1|s_2)$ for some particular choice of s_1, or are derived from $f(s_1|s_2)$ by integrating out quantities that are not of interest. The subset s_1 is always taken to contain x and usually, as in LF_{rv}, to contain other "variables" with given distributions. However, it should be emphasized that s_1 is sometimes also taken to contain components of the "parameters", as in Berger and Wolpert's (3.5.2) where η^2 is implicitly moved into s_1 and then integrated out. (In more colloquial terms, the choice of a likelihood function is essentially the choice of where to put the bar in the joint density $f(x,y,w,\xi,\eta)$. In LF_{rv} it is put between w and ξ, whereas in LF_{obs} it is put between x and y.)

Clearly, there are very many different possible choices of s_1, and the definition of the likelihood function can become very arbitrary. The fundamental idea is that in order to convey the evidence about the unknowns provided by the data, it is unnecessary to include any quantities other than x in s_1. Indeed, the possible inclusion of other quantities can only lead to confusion for the users of these likelihood functions. Thus we claim that the evidence provided by the data is conveyed most efficiently and most generally by the likelihood function that we have called LF_{obs}, as given by (9).

It should be emphasized that we are making an important distinction between the *evidence provided by* x about ξ and y, and the *information that is needed* to make inferences about ξ and y. This distinction is clear in the Bayesian approach, but less clear in the likelihood-based frequentist approach. However, even in that approach, the distinction becomes clear if LF_{obs} is always used but inferences incorporate other factors such as $f(y|\xi)$ in (3). Thus, a large variety of inferential aims can be accomplished with just LF_{obs} rather than an equally large variety of likelihood functions.

Discussion by Professor Bruce M. Hill

(University of Michigan)

I should like to congratulate Berger and Wolpert on their lucid and informative presentation of the history and substance of the likelihood principle, and their extension of the likelihood principle. Although I found their extension interesting, and hope that it may resolve some doubts concerning the status of the likelihood principle in the infinite case, my own view is that the likelihood principle really stands or falls in the finite case. The part of their article that I would like to discuss is that concerning the various examples that have been presented against the likelihood principle, where my views are perhaps different from those of Berger and Wolpert (BW), and in the course of the discussion my approach to the infinite case should become clear. Before doing so I want to preface my remarks with two comments. First, I think that we Bayesians should be grateful to Stein, Stone, Fraser and Monette, for their interesting examples, all of which have some real substance to them. Theories require good criticism in order to grow, and the lack of such criticism has been detrimental to the Bayesian theory. Secondly, I think it is essential that we keep in mind the distinction between the likelihood principle (by which I mean the formal likelihood principle of BW) and various implementations or interpretations of the likelihood principle. I shall try to demonstrate that none of the examples speak against the likelihood principle as such, but rather that they constitute frequentistic arguments against the use of specific improper (or diffuse finitely additive) prior distributions. I shall then explain why I think such arguments have no real teeth to them.

Let us begin with the example of Stein. Although it was originally presented by Stein as an argument against the likelihood principle, with an argument against lazy Bayesians tacked on at the end, I regard it as primarily an argument against Bayesians (such as myself) who use improper or finitely additive prior distributions to obtain approximate posterior distributions, and also against the theory of de Finetti (which I follow) which in principle does not rule out any finitely additive prior distribution. To begin with, the likelihood principle does not justify either (5.3.4) or (5.3.5) of BW, since it does not suggest a way of attaching probability to sets. It is true of course that some individuals who support the likelihood principle (perhaps with qualifications), such as George Barnard and A. W. F. Edwards, also sometime recommend the use of such probabilistic interpretations of the likelihood function, but that is by virtue of additional assumptions, whether explicit or implicit, and is not really part of the likelihood principle. BW apparently accept that (5.3.6) is a strong argument against the use of a uniform improper prior distribution for theta, but suggest that there is no difficulty for Bayesians because on the one hand theta is a scale parameter and so it is the logarithm of theta (if anything) that should be given a uniform prior distribution, as in Barnard's reply to Stein; and on the other hand, and more importantly, they argue that with proper prior distributions the type of interval that Stein shows has bad frequentistic properties can occur only rarely since "Y is almost certain to be enormous." Although I agree with both arguments of BW, it seems to me that the issue being raised by Stein is not whether a sensible Bayesian can avoid the intervals (5.3.5), but rather whether by virtue of carelessness or because his theory permits such intervals (as for example is true of the de Finetti theory) the unwary or even wary Bayesian will become frequentistic prey. If there really were a trap with teeth to it then Stein's example would suggest either that one stick with proper prior distributions, or else be quite careful in the choice of improper, or merely finitely additive, prior distributions, and as Stein says, this would make the "prior distribution used dependon accidental features of the decision problem." Now to me this seems to be a real and important issue. Suppose we are discussing a real-world parameter, whose existence, definition, and meaning, in no way depend upon the experiment to be performed. (The existence of such parameters may be far less common than is usually assumed, but presumably we could all agree that at least some such parameters exists, or at any rate are worth discussing, and confine attention to these. When the "parameter" depends upon the experiment for its existence and meaning, then, of course, the likelihood principle does not apply.) In the subjective Bayesian theory of de Finetti and L. J. Savage, the prior distribution for such a parameter would be chosen to represent one's opinions about that parameter, and whether the measurement is to be made according to the normal model or according to (5.3.3) should not in any way affect the prior distribution. If for some reason I thought that a uniform improper prior for theta was appropriate as an approximation under the normal model, and then learned that in fact the measurement error was distributed according to (5.3.3), but with the nearly identical likelihood functions that Stein produces, then it seems to me that the uniform prior should still provide a satisfactory approximation in obtaining my posterior distribution. Furthermore, a Bayesian who would use the uniform prior for theta when the measurement error is normally distributed, but would use a uniform prior distribution for the logarithm of theta when the measurement error is distributed according to (5.3.3), is coming very close to violating the likelihood principle in this example, since he is making very different inference about theta in the two cases even though the likelihood functions are in a certain sense very "close," and theta is the same fixed quantity. See Savage (1970) for a related argument. (Some Bayesians, for example Box and Tiao, actually recommend that prior distributions be made to depend upon the sampling scheme, and so would use a different prior distribution for the parameter of a Bernoulli sequence if the experiment were of binomial form than if the experiment were of negative binomial form, even when the choice of the experiment is made by randomization and is thus uninformative, and this clearly violates the likelihood principle. Whatever else may be said of such an approach, it is certainly not a part of the ordinary subjective Bayesian theory, in which the prior distribution for a parameter of the type we are discussing does not depend upon what experiments may or may not be performed at some future time. Furthermore, if BW choose to use improper prior distributions, but only when these do not lead to bad frequentist properties, then they too are perilously close to a violation of the likelihood principle, since their choice will turn out to depend upon the sampling distribution, just as with Box and Tiao.) What the Stein example actually demonstrates is that if a Bayesian uses the uniform prior distribution for theta, then his posterior probability for the interval (5.3.5), given any y, is at least .95, while given any theta, the frequentist probability for the interval is very tiny according to (5.3.6). This is the phenomenon of nonconglomerability. Conglomerability is a property of a probability distribution, and was defined by de Finetti (1972, p. 99) as follows: if the conditional probability of an event, given each element of a partition, lies between p and q, then also the probability of the event lies between p and q. Conglomerability always holds for countably additive probability distributions and countable partitions, but need not hold for merely finitely additive distributions, and in fact, as shown recently in Hill and Lane (1983) using only elementary mathematics and verifying a conjecture of de Finetti, conglomerability and countable additivity are equivalent for countable spaces. The uniform improper prior distribution can be given a finitely additive interpretation, which is why the nonconglomerability exhibited by Stein can occur. Thus for the partition based upon the value of theta, we have (5.3.6), while for the partition based upon the value of Y, the intervals (5.3.5) have posterior probability at least .95 for all possible such Y values, when the uniform prior distribution for theta is used. The unconditional probability of the interval has not been defined, but whatever value it is given must exhibit a nonconglomerability with respect to one or the other of the two partitions. In Hill (1981) I gave some general arguments as to why nonconglomerability cannot be avoided in the subjective Bayesian framework, and as to why I believe there is really very little that is operationally meaningful in the type of superficially frightening calculation exhibited in (5.3.6). After discussing the other examples, I will return to this issue, and suggest a new argument as to why there is no way to demonstrate any undesirable consequences if one uses an improper prior distribution. Thus although I agree with BW that the Bayesian can always avoid the trap by using proper distributions, I also like to use improper prior distributions or merely finitely additive prior distributions when I think they yield a simple and satisfactory approximation to my posterior distribution, and do not accept (as BW seem to do) that there are any operationally meaningful ill consequences to so using such distributions (even for all possible values of Y in the Stein example). A general theory pertaining to the type of consequences that arise in nonconglomerable situations has been formulated and elegantly presented by Heath and Sudderth (HS) in Heath and Sudderth (1978) and by Lane and Sudderth (1984), and as we shall see later all of the examples purportedly against the likelihood principle, are in fact merely more examples of the type of incoherence discussed by HS. (See HS example 5.2 for a very simple example similar to that of Stein.) It is my opinion, however, that the HS requirement for coherence, to the extent it goes beyond the de Finetti form of coherence (which only requires avoidance of sure loss with a finite number of gambles), is too restrictive, and at least in the special case of the Monette-Fraser example, I will argue that the apparent ill consequences of violating the HS condition for coherence cannot really be made operational.

The Stone example does not directly pertain to the likelihood principle, and has been analysed by myself in Hill (1981) from a finitely additive point of view. In addition to observing that a finitely additive diffuse uniform distribution on the "length" of path yields the standard confidence result, it was also pointed out that in order to obtain the uniform distribution on the location of the treasure that Stone criticizes it is necessary to employ a diffuse finitely additive prior distribution which gives odds of nine to one in favor of paths of length j+1 versus paths of length j-1, for all j > 1, and such a prior distribution seems rather silly in this example. Nonetheless, just as in the Stein example the de Finetti theory does not rule out such prior distributions, and the question is once again whether a serious case can be made against their use. It may be noted that the posterior obtained with this prior is also incoherent in the sense of Heath and Sudderth.

Fraser in his discussion of my Valencia article Hill (1981) maintained that the Stone example also has implications with regard to the likelihood principle, and gave the example reported by BW. The example as initially presented did not seem appropriate to me, since it required that first theta, the true path to the treasure, be selected as in Stone's example, next that the observed path of the Stone experiment be given, and finally that a randomization be performed that leaves one with the same likelihood function as before. In this situation, where the second experiment consists of the first experiment together with an irrelevant randomization, the likelihood principle follows from just the sufficiency principle, and is barely worth commenting on. However, the Fraser example can be modified so that this is no longer the case, for example, one can imagine that a new experiment E* is performed as follows: first a path z from the origin is selected according to a probability distribution that depends upon theta, in such a way that z is equally likely to be any of the four paths for which theta is a one block extension or retraction of z, and we observe this z. Next, someone else who somehow or other happens to know the true theta, unobservedly retracts or extends z back (or forth) to the true theta (which therefore remains the true parameter of the experiment), and from there does the experiment with the Fraser likelihood function as presented by BW, with z = x(0). One then observes in this last experiment a path X. The likelihood function for theta based upon the data Z = z, X = z, in the experiment E*, is then identical with the likelihood function derived from the Stone experiment with the same observed path z, and so with this modification the Fraser example does meet the conditions for the likelihood principle to apply. If one adopts a Bayesian point of view, then as BW argue, one has precisely the same apriori information about theta no matter which experiment is performed, and it is certainly reasonable to draw the same inference in each experiment. Suppose, however, one imagines that it is meaningful to consider the case of no prior information (whatever this means), so that Bayesian inference is not possible. It would be interesting to know what the appropriate non-Bayesian inference about theta would be under E* as opposed to the Stone experiment. Would, for example, a non-Bayesian now treat theta as though it were equally likely to be any of the four possible paths? Rather than calling into question the likelihood principle it seems to me that this example may raise some serious problems for non-Bayesians.

Now let us turn to the new example by Monette and Fraser (MF). This example does not seem to pertain directly to the likelihood principle, since there is only one experiment under discussion. It does, like the other examples, suggest that according to frequentist standards a certain improper, or diffuse finitely additive, prior distribution is unsatisfactory, and BW, as in the Stein example, argue that for proper prior distributions, and even for the conventional improper prior distribution for something akin to a scale parameter, there is no difficulty. Although again I agree with BW that ordinarily one need only consider quite proper prior distributions, and also that the particular improper or finitely additive distributions that are being castigated may be of no special interest, I would nonetheless like to argue that as yet very little has been demonstrated against the use of such prior distributions. My argument would be much the same in all examples, but will be presented here in connection with the MF example, which is the simplest. What has been shown is that choice of an improper uniform prior distribution (or a finitely additive diffuse prior distribution) for theta would lead to a posterior distribution, such that if I were to bet in accord with it, I would be a loser in the Heath-Sudderth sense (this is closely related to a lack of extended admissibility). Since I regard the finitely additive uniform distribution as useful for approximations, and as having as much justification as any other distribution (to be given full rights, as de Finetti says), and in any case I don't think that it matters whether theta is akin to a scale parameter, so that I cannot take refuge in the BW argument unless I dispense entirely with both merely finitely additive distributions and improper priors, I am loathe to give it up so easily. So suppose I fall into the trap and agree to post odds in accord with the posterior distribution that is uniform over the three possible values for theta, given x. Let us see to what extent MF can take advantage of such foolishness as I am willing to exhibit. In order to do so they must construct a real world version of their mathematical model. So first of all they must somehow or other pick a theta, and then pick an x in accord with their model. The Heath-Sudderth gambling scenario seems to be a convenient and appropriate way of describing the operational consequences of my potential incoherency (even for those who think that they don't gamble), and if desired, can easily be translated into non-gambling terms. Thus suppose that theta is picked from amongst the positive integers by the master of ceremonies in any way he likes, and then X is selected according to the MF distribution for X, given theta. After we are all given the value x that X takes on, I then use the posterior distribution based upon the uniform prior distribution for theta to determine the odds that I, as bookie, will give for the various values of theta. Also, after observing x, MF are entitled to place any finite number of bets concerning theta they wish, and finally theta is revealed by the master of ceremonies and all bets are paid off. Suppose that MF bet a dollar on the event that theta takes on the value $\delta_1(X)$, and let G denote the final payoff from me to them. Given theta, there is at least probability 2/3 that $\delta_1(X)$ will equal theta, and so the expectation of G, given theta, is at least \$1, for all possible theta, and I am incoherent in the sense of Heath-Sudderth. However, to make the transaction operationally meaningful it is necessary to specify precisely how X will be revealed, for example, that X will be expressed to the base 10 (or in any other specified form whatsoever), and that a certain finite time limit is prescribed during which the game is to be played. Now I think that all of us could come to agreement that given the constraints of the world we live in, there is an upper bound, say N, to the value of X that can be reported to us as data in the prescribed form and in the prescribed time, for example, an N such that in the present state of technology even the fastest computer could not display an integer greater than N in the time allotted for the experiment. (To be even more realistic, the same is true with regard to theta, but for the purpose of the present argument we need not assume any constraint on the magnitude of theta, and shall follow MF in assuming that the master of ceremonies can choose any value whatsoever, and then can and does select an X in the way that MF specify. Of course theta, like X, cannot actually be reported it if exceeds N, but one might wish to consider cases where the master of ceremonies has extraordinary powers, and is entrusted to announce who wins the gamble in situations where X does not exceed N but 2 X does. This points out that there are in fact a variety of ways to make the Heath-Sudderth scenario operationally meaningful, and that our assumption that X cannot be reported if it exceeds some known N, is merely the minimal real-world constraint. This gives the present argument greater generality in that it may apply even when theta is a real-world physical parameter for which there would be no known bounds. If a bound on theta were available then of course the argument would apply all the more. However, the point is that whether or not there is such a bound on theta, there is necessarily a bound on the possible value of X that can be reported. If we do take into account known bounds on possible theta, or on possible reported values of theta, then this would lead us to proper prior distributions as in BW. However, it is not necessary to introduce such considerations in the present example since, as we shall soon see, the boundedness of the X that can be reported already destroys the frequentistic argument.) Suppose then that theta and X are selected by the master of ceremonies in accord with the MF model, without any constraint upon the magnitude of either, and that N is a known upper bound for any X that can possibly be reported as data. We do not assume that N is the least possible upper bound for a reportable X, but merely that it is an upper bound. (It is, of course, desirable that N be not too much larger than the least upper bound, but the argument does not depend upon this.) Thus our experiment now consists in precisely the MF experiment, together with the modest real world constraint that if X > N, then no value of X will be reported (since it would be impossible to do so), and hence that any bets that depend upon the value of X will be called off. In this situation the actual gamble as to whether theta is $\delta_1(X)$ is called off whenever X > N, and we are dealing with a conditional gamble in the sense of de Finetti (1974, Ch. 4). Consequently the payoff from me to MF is now as before if X is actually reported, but all gambles are called off if X > N. (There is nothing underhanded here with regard to the reduction to conditional gambles: in order that transactions can occur, so that the scenario has operational meaning, it is necessary that the bets are conditional bets, given that a value of X is reported, and hence conditional upon the event that the X selected in the MF experiment does not exceed N. If X > N then no X is reported and no gambles can be made concerning whether theta = $\delta_1(X)$. Note also that it is not necessary to assume that X must be reported if X < N, but merely that X cannot be reported if X > N, and that X must be reported if it is possible to do so in the fashion prescribed.) It is interesting now to see what becomes of the frequentist argument that showed that the conditional expectation of my loss, given theta, is at least \$1, for all possible theta. I am still using the same prior distribution as before, so that if I am actually given a value of X (necessarily $\leq$ N) then I post the same odds as before against the event that theta = $\delta_1(x)$. If theta is sufficiently small so that X both can and must be reported (hence necessarily theta < N/2), then the expectation of G, given such a theta, is the same as before, at least \$1. On the other hand, if theta > N/2, then the only values of X that can possibly be reported are X = theta/2 or (theta-1)/2, depending on whether theta is even or odd. Hence given a value of theta > N/2, and given that the gamble is not called off, it is certain that theta is not equal to $\delta_1(X)$, and so the conditional expectation of G, given such a theta and that the gamble is not called off, is -\$1, while the unconditional expectation of G, given such a theta, lies between -\$.33 and \$0. Whether in gambling terms or in coverage probability terms, it is thus seen that when a real-world constraint as to the value of X that can be reported is incorporated into the MF example, then the example breaks down, and in fact if a value of X is actually reported, then the very same $\delta_1(X)$ that appeared so desirable from the MF point of view, becomes impossible as the value of theta when theta > N/2. (A variation of this scenario would require me also to post odds on theta, given the information that X exceeds N. This would require care in obtaining the posterior distribution for a finitely additive prior distribution, but in any event the $\delta_i(X)$ are still not available, and the frequentistic argument still breaks down.)

The above form of argument suggests why there need not be anything wrong with using the finitely additive uniform distribution in connection with experiments conducted by human beings, i.e., where the reportable observation X, if not theta itself, must be bounded, and one can with a little thought always choose a generous upper bound. More generally, when theta is not chosen by any human, but is a parameter of the real world, then one may not be able to argue for any upper bound for theta, but in my opinion neither will there by any operationally meaningful scenario in which one who chooses a finitely additive distribution can be shown to be in trouble by virtue of frequentist properties. BW suggest using proper prior distributions for theta as a way of avoiding the apparent frequentistic difficulties in the above examples. However, if BW or Heath-Sudderth wish to use improper or merely finitely additive prior distributions, and if they choose to avoid nonconglomerability and its frequentistic consequences, as in the various examples, then it seems to me that they are in fact going to violate the likelihood principle, since the particular improper or finitely additive distributions that they must rule out in order to avoid nonconglomerability will depend upon the form of the experiment, just as the prior distributions that Box and Tiao recommend depend upon the form of the experiment. (BW can avoid violating the likelihood principle by either restricting themselves to proper prior distributions, or by using improper prior distributions only when they provide an "adequate" approximation to the posterior distribution based upon some proper prior distribution. But I think it is too restrictive always to restrict oneself to proper prior distributions, and although, as mentioned earlier, I too ordinarily take the approximation point of view, I don't think the notion of what is an adequate approximation should depend upon frequentistic properties.) In the Stein example and in the Heath-Sudderth example (5.2), where according to the model (taken literally) the parameter and data are not discrete and the set of theta compatible with the data is not finite, the argument I have given above must be modified, but I think that here too, when real-world constraints are allowed for, the frequentistic argument will again break down, and I hope that my discussion of the MF example at least suggests some of the difficulties involved in trying to make the frequentistic argument operational. In my Valencia article I also suggested that as yet no serious argument for conglomerability had ever been given (since that time Lane and Sudderth (1984) have given such an argument, but I do not agree with their views concerning the appropriate gambles with which to define coherency), and suggested also that Stone's example had an implicit assumption of conglomerability for its castigation of the uniform prior. (Stone (1979) replied by asserting that Hill is like a prisoner condemned to death by guillotine who rejoices that the guillotine will be chosen from an infinite collection. I replied "Yes, Mervyn, but all your guillotines are made of butter." At a deeper level this concerns the appropriate interpretation of conditional probability, whether in terms of gambles that are called off if the conditioning event does not occur, as in de Finetti (1972, p. 81), or in the more usual way, but there is not space to go into this here.) Sir Harold Jeffreys once criticized conventional tests of significance because they reject hypotheses that may be true on the basis of data that have not occurred. Apparently some would also have us reject the use of improper prior distributions because of experiments that cannot be performed.

Finally, let me mention an important real-world problem where exactly such considerations as I have been discussing arise. Consider a balanced one-way random effects analysis of variance model, with I rows and J columns. In Hill (1980) I examined the consequences of drawing inference about the ratio of the between to the within variance, τ, using as data only the ratio of the mean square between to the mean square within. It was shown that this can in fact be justified by a fully Bayesian analysis, and is appropriate when the prior distribution of the two variance components is such that their ratio is independent of the within variance, and the overall mean is given a diffuse prior distribution. The problem then reduces to one of inference about a simple location parameter, $\gamma = \ln(1+J\tau)$, based upon data $\hat{\gamma} = \ln(MSB/MSW)$, and with the distribution of $\hat{\gamma}-\gamma$, given γ, being that of the logarithm of a random variable having the F distribution with I-1 and I(J-1) degrees of freedom. The likelihood function for γ based upon the data $\hat{\gamma}$ is then the density of this ln(F) distribution, translated so that the mode is at $\hat{\gamma}$ (and with degrees of freedom reversed), except that the density must be truncated from below at 0 because γ is nonnegative (it is convenient and harmless to think of the likelihood function as being defined for all γ, so that even if $\hat{\gamma}$ is negative, the mode is at $\hat{\gamma}$, and then to make the truncation from below at 0 stem from the prior distribution.) If one uses the uniform prior distribution for γ, with $\gamma > 0$, then one is in precisely the type of situation that the Stein, HS (5.2), and MF examples, deal with. Although there is nothing magical or mandatory about use of this particular prior distribution, and in fact there is usually a great deal of prior information about the ratio of variance components in such problems, so that I would recommend use of a proper prior distribution for γ, at the same time, I think a great deal of insight can be obtained from the improper uniform prior on γ, and do not think it should be automatically ruled out merely because it may lead to bad frequentistic risk properties. As I argue in Hill (1980), the posterior expectation based upon this improper prior yields, at least in some respects, a more plausible estimator for a multivariate mean (the realized random effects) than does the positive-part Stein estimator. For example, the posterior expectation cannot shrink all the way to the grand mean of the observations, since the weight given to the row means $\bar{y}_{i\cdot}$ decreases only to 2/(I+1) as the ratio of the between to within mean square goes to 0, whereas of course the positive-part Stein estimator can give zero weight to the row means, and this is not always sensible. The behavior of the posterior expectation stems partly from the particular form of the prior distribution for the variance components (especially the fact that the ratio of the variance components is apriori independent of the within variance component), and partly from the truncation of the posterior distribution of γ from below at 0, neither of which do non-Bayesians incorporate into their analysis. In my opinion due respect for the likelihood principle, and proper allowance for these aspects of the problem, are far more important than any frequentistic arguments against the use of improper prior distributions, while at the same time, as BW would presumably agree, a proper prior distribution for the variance components would ordinarily be reasonable, and give the best of both worlds.

DISCUSSION OF THE SECOND EDITION BY PROFESSOR HILL

Since the publication of the first edition of the monograph by Berger and Wolpert, I have written several articles pertaining to the validity of the likelihood principle, and to its role in Bayesian data-analysis. I believe that the example of Hill (1987a,b) clearly shows that the original statement of the likelihood principle by Birnbaum in terms of an abstract concept of evidence was faulty. The difficulty in the likelihood principle is easily remedied, however, and this was done in my statement of the restricted likelihood principle in those articles. In my formulation one speaks not of the evidence in some undefined abstract sense, but rather only of the evidence about the *value* of θ, and excludes from the discussion any assertion about how θ might relate to other unknowns, whether hypotheses or parameters. Thus my example can be viewed as showing that two different experiments that yield proportional likelihood functions for θ do not necessarily provide the same evidence about θ, since we can learn, for example, that θ has a different 'color' in the two experiments. The color might be an important part of the overall evidence about θ. Of course the color can be included in the parameter, but the likelihood principle, as usually formulated, does not require one to do so. It is hoped that once this point is understood, others will, like myself, become even stronger supporters of the essential part of the likelihood principle.

The basic point of my example is related to fundamental questions that arise in theories of causality, for example, concerning determinism and the possibility of independence in the real world. Such questions arise in critical discussions of quantum mechanics and relativity theory, for example, in connection with Bell's inequality, as well as in philosophy.

In Hill (1985-86, p. 223) I have given an account of how the likelihood principle must be further modified to deal with Bayesian data-analysis, where through exploration of the data, one may modify the original model. The same article, p. 202f, argues that even apart from inadmissibility, incoherence, and the failure to utilize available information, the frequentist approach breaks down completely in connection with such data-analysis, since all frequentistic assertions must be conditional not only upon the diagnostics used, but their order, and even the thoughts that cross one's mind. Such conditional probabilities are plainly both unknown and unknowable. Finally, Hill (1988) gives a very short, and partly new, proof of the stopping rule principle, i.e., that the stopping rule is irrelevant for inferential and decision-making purposes, or that "sequential analysis is a hoax," as concluded by Anscombe (1963, p. 381). Here the proof does not depend upon the likelihood principle, or even the restricted likelihood principle. Instead, it is shown that on a post-data basis, i.e., given the *realized* data, sequential analysts purport to extract information over and above that following from the corresponding fixed sample size experiment, from a *logically certain event*. In this article the important distinction between the pre-data and post-data considerations is emphasized. Once one is given the data, the primary aim must be to make intelligent and rational decisions, for which the Bayesian approach seems quite well suited. Of course sequential *design* need not necessarily be a hoax, but it appears that not very much is known about this potentially important subject, perhaps because of the confusion between pre-data and post-data considerations, as discussed in Hill (1988).

The likelihood principle is often mistakenly assumed to be largely equivalent to the Bayesian approach. The likelihood principle, as proposed by Birnbaum, in terms of an abstract and empty concept of evidence, was in fact the last gasp (intellectually speaking) of the theory of classical statistics, with its naive pretence at objectivity. Indeed, Birnbaum (1962, p. 277) quotes Jimmie Savage as follows. "Rejecting both necessary and personalistic views of probability left statisticians no choice but to work as best they could with frequentist views... The frequentist is required, therefore, to seek a concept of evidence, and of reaction to evidence, different from that of the primitive, or natural, concept that is tantamount to application of Bayes' theorem."

"Statistical theory has been dominated by the problem thus created, and its most profound and ingenious efforts have gone into the search for new meanings for the concepts of inductive inference and inductive behavior. Other parts of this lecture will at least suggest concretely how these efforts have failed, or come to a stalemate. For the moment, suffice it to say that a problem which after so many years still resists solution is suspect of being ill formulated, especially since this is a problem of conceptualization, not a technical mathematical problem like Fermat's last theorem or the four-color problem."

Birnbaum then states that "The present paper is concerned primarily with approaches to informative inference which do not depend upon the Bayesian principle of inverse probability." It would therefore appear that Birnbaum regarded his approach to evidence as meeting the objections that Savage and others had raised. However, just as the Michelson-Morely experiment spelt the death knell for classical physics (which was at least a highly successful and useful subject), one must wonder what is left of classical statistics, without even Birnbaum's likelihood principle to sustain it. All that appears to be left is the restricted likelihood principle, which is implied by the Bayesian approach, and is somewhat more general than the Bayesian approach, since it allows for versions of Bayesian data analysis such as in Hill (1988). I know of no way to demonstrate even the restricted likelihood principle, however, other than through the Bayesian approach.

I think that nowadays it will be readily understood that the pretence at objectivity in classical statistics was equivalent to taking a particular subjectivistic Bayesian view, that based upon diffuse prior distributions, and by fiat declaring that this constitutes objectivity. Such prior distributions play an important role in Bayesian statistics, via the stable estimation argument of Jimmie Savage, but do not acquire any magical status in the Bayesian theory.

The nature of "objectivity" was never seriously discussed in classical statistics, despite the fact that this was and is a notoriously difficult question in philosophy. Even in statistics, numerous examples exist showing that this pretence cannot be made, without leading to absurdities. There are many examples in which the *realized* likelihood function is nearly flat, no matter what the pre-data expected information may have been. This occurs, for example, in inference about variance components when the classical unbiased estimator of the between variance component is negative, as in Hill (1965, 1967). A more sophisticated example of the need for a subjective view occurs in deciding whether a particular observation is an "outlier," as in Hill (1974b, Section 4) and Hill (1988, Section 3). What the so-called objectivists do, as Jack Good says, is to SUTC (sweep the subjective aspects under the carpet). Probability and statistics, as related to the real world, are fundamentally subjective or personalistic. In certain situations, however, one may obtain practical objectivity by means of a consensus as to appropriate prior distributions and models. See Hill (1985-86, 1988). Also, sometimes certain "objectivistic" methods, such as the fiducial approach, can be justified Bayesianly, as for example with A(n) in Bayesian nonparametric statistics, Hill (1987c). Finally, by a delicious irony, it also turns out that the few important objective *criteria* that frequentists have recommended, such as admissibility, extended admissibility, etc., lead inevitably back to the Bayesian approach.

The distinguished philosopher and psychologist, William James (1896, p. 97) puts it quite well: "Objective evidence and certitude are doubtless very fine ideals to play with, but where on this moonlit and dream-visited planet are they to be found? I am, therefore, myself a complete empiricist so far as my theory of human knowledge goes. I live, to be sure, by the practical faith that we must go on experiencing and thinking over our experience, for only thus can our opinions grow more true; but to hold any one of them - I absolutely do not care which - as if it never could be reinterpretable or corrigible, I believe to be a tremendously mistaken attitude, and I think that the whole history of philosophy will bear me out."

James's eloquent statement can serve as a preamble to the theory and practice of Bayesian data analysis and decision-making, which is a synthesis of the empiricism-pragmatism of John Locke, David Hume, Charles Peirce, and William James, with the rationalism of Plato, Descartes, Kant, and others, and to which I believe that the next century will be devoted.

Discussion by Professor David A. Lane

(University of Minnesota)

Berger and Wolpert have done the statistics community a service by calling our attention once again to the likelihood principle (LP) and its implications. They repeat Birnbaum's(1962a) message, already admirably recapitulated by Basu (1975) and Dawid (1977): *if* you work within the classical $(X, \Theta, \{P_\theta\})$ - paradigm, you want to make inferences about "true θ" on the basis of "observed x," and you wish to respect certain fundamental principles of inference (for example, the sufficiency and weak conditionality principles), *then* your inference had better depend upon the observation x through the likelihood function that x induces on Θ. In particular, you must accept the implications of some other principles that many statisticians regard as false, never mind fundamental, like the stopping time and censoring principles.

There are several bail-out options for statisticians who choose neither to follow the LP to fully conditional analysis nor to raise adhockery to a scientific principle. They can reject the $(X, \Theta, \{P_\theta\})$ - paradigm by requiring either more structure (as do structuralists, pivoteers, and, perhaps, some "objective" Bayesians) or less (as do defenders of alternative-free significance tests and, more drastically, exploratory data analysis); or they can modify the fundamental pre-principles so that the LP and the objectionable post-principles fail to be derivable from them, as did Durbin (1970) and Kalbfleisch (1975); or they can claim that other, more fundamental, principles, like the Confidence Principle, conflict with the LP, making an ideological choice among competing principles necessary.

Since Bayesian practice is consistent with the LP, Bayesians have no need to refute Birnbaum's work. Indeed, to Berger and Wolpert, the LP is a trump card in the Bayesian salesman's hand. They argue, as did Basu (1975), that only Bayesian ideas permit the LP to be properly implemented and that Bayesian considerations unravel the "counterexamples" to the LP produced by Armitage, Stein, Fraser and others.

But even to Bayesians, consistency with Bayesian ideas should be no guarantee of foundational cogency. For example, the fact that (essentially) admissible decision rules are Bayes does not recommend Wald's formulation of decision theory to most Bayesians. So the question arises: should Bayesians promote Birnbaum's formulation and derivation of the LP as a cornerstone of the foundations of statistics? I think not, for two reasons. First, the LP is embedded in a paradigm which is not directly applicable to many, if not most, of the important real problems of statistical inference. Because of the ambiguity and limitations of this paradigm, the proof of the LP is not compelling. Second, the LP ignores what I regard as the fundamental tenet of Bayesianity: the purpose of an inference is to quantify uncertainty. When this tenet is properly taken into account, foundational arguments can be adduced that lead directly to Bayesian methods.

The next section elaborates the first of these reasons in some detail. For a development of the second, see Lane (1981) and Lane and Sudderth (1984).

$(X, \Theta, \{P_\theta\})$ and the LP

I shall discuss three problems with the LP. The first relates to the meaning, the second to the adequacy, and the third to the relevance of the $(X, \Theta, \{P_\theta\})$-paradigm. Both the first and the second of these problems call the derivation of the LP into question.

1. What do the elements of Θ represent? This question is important, since the proof of the LP requires us to consider the mixture of two different experiments with "the same Θ." There at least three possible interpretations

of the elements of Θ:

a) θ *is* the distribution P_θ;

b) Θ is an abstract set and θ merely indexes the distribution P_θ;

c) θ is a possible value for some "real" physical parameter, and P_θ is to be regarded as the distribution of the random quantity X should θ be the true value of that parameter.

Interpretations a) and b) are mathematically precise. They are defined in terms of the assumed model and do not refer to the physical reality that model is intended to represent.

Interpretation c) has an entirely different character and raises difficult philosophical issues. When - and in what sense - do "real" physical parameters exist? If I opt for interpretation c), must I believe that a coin has a propensity to come up heads θ x 100% of the time in an (infinitely) long series of repeated flips? I am inclined to believe that there may be "real" physical parameters in measurement error problems, although even here a strict operationalist construction leads to interpretation a) rather than c) for the parameter θ: the measuring process, encoded as P_θ, defines the quantity measured. In few other problems to which statistical inference is applied are there model-free physical quantities standing behind each model parameter. To decide whether or not you agree, think about your last regression or time-series analysis.

Both Berger and Wolpert (pp. 42-3) and Dawid (1977, p. 252) seem to favor interpretation c). For example, Berger and Wolpert say that the LP applies only when the elements of the two parameter sets are "the same parameter, i.e. are physically or conceptually the same quantity." Unfortunately, they neglect to tell us how we are to decide when two different experiments measure the same quantity or how to deal with model parameters that lack any natural interpretation in terms of physical quantities. Moreover, in virtually all of their examples, the set Θ is uninterpreted and merely serves to index the set of distributions $\{P_\theta\}$, which suggests that in these cases they are thinking about Θ in the sense of interpretation b).

It is hard to take the LP seriously as a foundational instrument if we must always interpret the elements of Θ as "real" physical quantities, unless we are given some guidance on what constitutes reality and how reality is tied to mathematics by the models we select.

It matters which of the three interpretations we give to the elements of Θ. They lead to very different conclusions about the validity of the derivation of the LP. Interpretation a) gives no scope for the mixture principle: only experiments whose sampling distributions are identical share "the same Θ." As such, the LP is reduced to the sufficiency principle and, for example, the stopping time principle does not follow from the LP.

Interpretation b), on the other hand, gives tremendous scope for mixing. Any two experiments with the same index set can be mixed. Consequently, if there are a pair of observations, one from each experiment, that yield the same likelihood function on the index set Θ, the LP then declares that the "evidence" or "inference" derived from the two experiments with these two observations must be identical. This is a startlingly unBayesian conclusion. For example, must my predictive inference for the next outcome in *any* sequence of Bernoulli trials in which I have so far obtained three successes and one failure be the same? But what in the mathematics of the LP proof precludes interpreting Θ purely as an index set and so deriving a version of the LP that conflicts with Bayesian practice?

The foundational status of the LP cannot be determined until Θ is interpreted. Depending on whether one adopts interpretation a), b) or c), the LP is devoid of interesting consequences, wrong, or severely and ambiguously restricted in its domain of applicability.

2. The proof of the LP is convincing only in so far as the sufficiency and weak conditionality principles are intuitively compelling. While Bayesian practice respects both principles, only weak conditionality seems unarguable on its face. I share I. J. Good's reaction to the sufficiency principle, as reported in his discussion of Birnbaum(1962a). Despite Fisher's gift for

suggestive names (what more could you possibly need than something that is sufficient?), the fact that the distribution of X given the value of a statistic T is θ-free does not immediately impel me to base my inference only on the value of T.

Suppose, though, that the observation x is generated by first generating a value for T according to a distribution indexed by some element of the parameter set Θ and then an extraneous randomization mechanism is used to pick an x on the orbit of the observed value of T. In such a case, it is clear that T is sufficient and that inference about θ should be based only on T. (The sufficient statistic that appears in the derivation of the LP does not bear this postrandomization relation to the observation x.)

Now, for any sufficient statistic T defined on a statistical model $(X, \Theta, \{P_\theta\})$, there is no way to tell from the information encoded in $(X, \Theta, \{P_\theta\})$ whether the observation x is or is not generated from T by postrandomization. So, if you do not find the sufficiency principle compelling except in the postrandomization case, you must agree with Barnard and Fraser that not enough information is encoded in $(X, \Theta, \{P_\theta\})$ upon which to base a general principle of inference. And I believe that this conclusion is correct. After all, the information in $(X, \Theta, \{P_\theta\})$ says nothing about how the model represents reality, and it is hard to see how a principle of inference can disregard the details of this representation. Though we use models to guide the way we formulate inferences, the inferences themselves have value to us only if they yield useful statements about the world.

3. Even though "inference" is undefined in the LP formulation, the validity of the LP seems to depend on two premises about the nature of inference in the $(X, \Theta, \{P_\theta\})$ - paradigm:

a) The purpose of inference is to make some statement about the "true" value of an unobservable parameter θ on the basis of an observed quantity x;

b) θ exists independently of the "experiment" E that produces x, and information about θ can be separated into two components, one deriving just from

E (to which the LP refers) and "other information" presumably preexisting E.

I believe that these premises are rarely true in real situations to which statistical inference is applied. If I am right, the scope of the LP as a foundational instrument is narrow.

Except for measurement error problems, the real aim of inference is usually to generate a prediction about the value of some future observables; see Geisser (1971 and 1984) and Aitchison and Dunsmore (1975) for extensive discussion of this proposition and further references. This is especially true in situations where the model parameters do not represent real physical quantities, the typical case in regression and time-series analyses. Estimating model parameters is in general a "half-way house" on the way to predicting some relevant future observation, and much can be lost by focusing foundational discussion on the half-way house instead of the ultimate destination. For example, the relevant uncertainty for a patient with a particular clinical condition undergoing a particular therapy is not a confidence band for an estimated survival curve; rather, the patient and his physician should be concerned with the predictive distribution for that patient's future lifetime. The inferential question of interest to the patient is how to generate this predictive distribution.

The LP does not address this question directly. Berger and Wolpert claim that prediction can be embedded in the LP framework by including the future observable as part of the unknown parameter. But then θ appears as a nuisance parameter that is clearly not "noninformative" in the sense of Berger and Wolpert. LP ideas provide no guidance on the treatment of informative nuisance parameters. On the other hand, de Finetti's subjective Bayesian theory is directed towards the problem of predicting future observables, and the notion of coherence derived from that theory provides a foundational basis for predictive inference; see Lane and Sudderth (1984). In this theory, models may be used to help generate predictions about the future observable y based upon observed x, but the models merely provide a convenient structure and need carry no metaphysical burden of "reality" for the parameters they contain.

Premise b) cited above ignores the fact that model parameters are frequently inseparable from the "experiment" whose possible distributions they index. Especially in applications arising in nonexperimental sciences like econometrics or resource management, the model is scupltured either from data already in hand or perhaps from a realistic view of what data are potentially obtainable. In such cases, there is no way to separate what (E,x) says about θ from "prior" information about θ; in fact, θ cannot be said to exist prior to the formulation of E, even though there may be much prior information about which x might be observed. In these situations it is hard to criticize "objective" Bayesians who violate the LP by letting their "priors" depend upon the structure of the experiment E.

Discussion by Professor Lucien LeCam

(University of California at Berkeley)

Professors Berger and Wolpert are to be thanked and congratulated for giving us a closely argued view on the foundations of statistics. Their arguments in favor of the Likelihood Principle are very persuasive indeed. One may suspect, however, that some readers will be convinced and converted while some others will hold fast to their misguided beliefs, in spite of all the evidence.

I shall try here to indicate why the present writer belongs to the latter category.

There is a body of statistical theory, call it "type 1", that deals with the following kind of systems. When contemplating a particular unresolved question, one devises experiments to ascertain what the facts are. The mathematician will abstract the idea of "experiment", using an object formed by a family of probability measures on a suitable field. The consequences of using particular procedures to analyse the "experiment" are then describable in probabilistic language. One can attempt to single out procedures that have a reasonable performance in this probabilistic world. That is a bit like selecting tools: wrenches are often, but not always, successful at unscrewing bolts; paint brushes often fail in the same activity.

This kind of endeavor has given us the Neyman-Pearson theory and Wald's theory of "statistical decision functions". One can readily claim that the whole enterprise is misguided, but it does seem to have a role to play in certain endeavors, like planning experiments, settling arguments that involve several scientists and odd questions such as "is methotrexate effective in the treatment of colon cancer."

There is another body of theory, call it "type 2", that deals with axioms of coherent behavior and principles of evaluation of evidence. Some of it, and perhaps most of it, has to do with what "one" should "think" after the results of the experiments have become known. Comparatively little has been written on how "one" can transmit the "evidence" to another person, even in the Berger-Wolpert text, this communication problem takes second place to the " one should think" question.

Berger and Wolpert see evidence of contradiction between the "type 1" and "type 2" approaches. In a strictly mathematical view of the problem, there is no overlap between the two approaches because "type 1" does not have any probabilities to play with once the dice have been cast.

Consider for instance an experiment involving two containers, one with 50% red objects, the other with 25%. A coin is tossed to select a container. Then one extracts a ball from that container. It turns out to be blue. When all of that has been properly carried out there are no probabilities left since the container has been selected. It is either the first or the second and not a probabilistic mixture of both. Any assignment of probabilities at that stage requires amplification of the model, with thinking about possible repetitions of the experiments or degrees of belief, or betting strategies or whatever.

Berger and Wolpert try to convince us that in such a situation one should follow the likelihood principle. The argument is thorough. L. J. Savage's argument was also very thorough, but I have yet to find a *scientist* who would be convinced by a posterior distribution on the methotrexate and colon cancer question if the prior has been supplied by a pharmaceutical company. The point is that one can easily argue oneself into a corner.

In the present case,however, I think that the argument has one major flaw. It is based on the assumption that given an experiment E, and the result x of that experiment, there is a well defined object Ev (E,x). The nature of the object Ev (E,x) is not described explicitly. This is not the problem. What matters is the assumption that there is such an object, or more specifically a function (E,x) → Ev (E,x). Starting from such an assumption, and adding a few other "principles", one can prove that the function Ev must have certain properties.

(I am reminded here of the standard "proof" that 1 = (-1), assuming that $\sqrt{\quad}$ is a function: certainly $\sqrt{1/(-1)} = \sqrt{(-1)/1}$ and by multiplication $(\sqrt{1})^2 = (\sqrt{-1})^2$. I am also reminded of Spinoza's "Theologia more geometrico demonstrata").

The very existence of the function Ev is not clear to this writer. Even if it exists in a strictly mathematical abstraction of "experiments" and results, the relevance to practical applications is not directly evident.

Several years ago a problem of this nature was raised during the conference on the Future of Statistics held at Madison. Someone asked the panel how they would report the evidence in a clinical trial of a drug intended to suppress renal calculi. The answer, given by G. Barnard, was "report the likelihood function". That may be, but one should also report the age, ancestry, health status of the participants, the presumed mode of action of the drug, its manufacturer, ideas about whether calculi occur in clusters or bunches, their size distribution, whether their formation may be spurred or hindered by nutritional factors, etc., etc., including whether the randomization used (or unused) led to apparent disbalance.

There is no shortcut to reporting what was actually done and observed. In situations involving games of chance with definite rules, one might simplify the evidence report. It is also true that Savage could argue that anyone playing games according to Savage's rules need only report (to himself) the resulting posterior distributions.

It does not follow from such mathematical theorems that one must necessarily frame practical questions in terms of Savage's games or in terms of the Berger-Wolpert rules of evidence, even if these authors eventually argue themselves into a Bayesian framework.

Here the situation is complex because "type 1" theories have given proofs that "experiments" are characterized by the distributions of their likelihood functions. Also it is a standard result of "type 1" theories that Bayes procedures, or their limits form complete classes. A main difference is that the "type 1" theories insist that they are about risk functions, not possible interpretations of single posterior distributions.

The passage from advocation of the likelihood principle to Bayesian theory is described by Berger and Wolpert but not as a strictly logical consequence of the L. P. principle and other explicitly stated axioms. It is weak compared to the rest. However, in the process, they also demonstrate that they do not abide by their own L. P. prescriptions.

This occurs in the discussion of an example of C. Stein. The authors say

"note that it was assumed that $x = y = \sigma d$ in the above conditional analysis, and *since it can be shown that* Y *is almost certain to be enormous*...." (emphasis added).

That seems to be a very direct appeal to a frequency evaluation of the situation, and not even a conditional one at that. Such an appeal does not fit with the logic of the rest of the paper.

There are other matters that should be discussed, but it would take too much space. One of them has to do with approximation. Assuming that the function Ev exists and that if (E_1, x_1) and (E_2, x_2) give the same likelihood function, then the evidences are the same, is one entitled to presume that if the likelihood functions are 'approximately' the same then the evidences are also "approximately" the same?

Here we have two "approximately" with undefined, but perhaps definable meaning in the first instance and an apparently undefinable meaning in the second occurrence since Ev (E,x) itself is an undefined object.

For instance it is a classical result that if one takes a very large sample $(x_1,\ldots,x_n)$ from the standard Cauchy $\{\pi[1+(x-\theta)^2]\}^{-1}$, for "most" samples the likelihood function will be "close to" one obtainable from a single observation y from $\mathcal{N}(\theta, \frac{2}{n})$. Does that have any "evidential" meaning for the L.P.? Must one necessarily interpret it only through a computation of posterior distributions? If so, for what priors?

To summarize Berger and Wolpert have given us a valiant defense of the L.P. However it does depend on a basic assumption of existence of the evidence function Ev. This function, if it exists, does not conform to the tradition of reports in scientific journals. The theory does not actually conflict with the so-called "classical" one because their domains of existence are separate and their aims different.

This author presumes that there is some value in some of "classical statistics" and also in the likelihood principle, but feels that one cannot support the practical application of either (or of other theories) on purely mathematical grounds. One should keep an open mind and be a bit "unprincipled".

DISCUSSION OF THE SECOND EDITION BY PROFESSOR LE CAM

In order to avoid any misunderstanding, let me repeat two of the criticisms made above: Using "evidence" as a function of a pair (E, x) and using "approximate" likelihood functions. The two points are highly interconnected.

For simple experiments Professors Berger and Wolpert use a mathematical entity E that consists of a set Θ and a family $\{P_\theta, \theta \in \Theta\}$ of probability measures on a given σ-field.

"Evidence" is undefined, but it is supposed to be a function of the pair (E, x) where x is the observed value. This may seem innocuous, but it is definitely not. It rejects a part of the data usually considered as part of the evidence in the common language use of the term, namely the thinking that went into the selection of the mathematical entity called E.

Except perhaps in those cases where the randomization is man-made on purpose and perhaps also in the Poisson formulas for radioactive decay, the selection of the "model" E is based on rather loose arguments that are not themselves representable by pairs (E', x'). Here, by "model" I do not mean something like "linear models" but more a mathematical construct that attempts to catch the important features of a physical or biological phenomenon. I have, of course, no objection to a theory of "evidence" based on a function of pairs (E, x). It just fails to connect properly with my own intuitive notion of what evidence is. Therefore I do not feel bound in practice by the theorems derived from such a theory.

Even if one tries very hard to put the information in the form (E, x) one will almost always put in E certain formulas for the sake of convenience, simplicity or plain laziness. Thus (E, x) will only be our approximation to a "better" (E', x'). It is my feeling that, if one wants to take into account the fact that such approximations are the rule, one must also explain what differences they may make in the use of the undefined "evidence" Ev (E,x). I am not too sure that this can be accomplished without introducing in the system a variety of concepts that go beyond pairs (E, x).

In summary I remain opposed to the apparent normative aspect of a theory that says that I *must* abide by the LP when I am unable to put my emotions and various bits of knowledge, or lack of knowledge, into it.

Reply To The Discussion

We are very grateful to the discussants for their stimulating comments. Besides describing interestingly different perspectives, the comments serve to highlight a number of important issues we inadequately discussed in the text.

REPLY TO PROFESSORS BAYARRI AND DEGROOT

It is indeed a pleasure to thank Professors Bayarri and Degroot for their careful reading of our manuscript and the deep insight reflected in their discussion. In the manuscript we tried to explore the implications of the LP and the issues it raises without endorsing any particular mode of inference (until the final chapter); in particular we tried hard not to let our Bayesian point of view color the basic arguments enough to make them unpersuasive to followers of the frequentist tradition. Thus our emphasis was not on "what is the likelihood function?" Rather, we took the likelihood function as given, and argued that the LP would follow no matter what reasonable definition of the likelihood function is used. The definitions in (3.5.1) and (3.5.2) are both reasonable, and serve different purposes.

But we are Bayesians, and are in essentially complete agreement with the basic issues raised by Bayarri and DeGroot. We agree that there is no clear distinction between "parameters" and "variables", and that definition of the likelihood function is ambiguous. As Bayarri and DeGroot observe, any partition of the parameters and variables into two disjoint sets s_1 and s_2, with s_1 containing the observed quantity x, leads to an acceptable likelihood function $\ell_x(s_2) = f(s_1|s_2)$ (providing this function is accepted as "known"). As long as one also keeps track of all known marginal and conditional information about the variables and parameters, any such partition leads to a likelihood function which contains all evidence from the experiment (at least to a

Bayesian). But the need to keep track of this marginal and conditional information, and to treat unknowns in s_1 differently from unknowns in s_2, should be sources of concern to non-Bayesians.

Bayarri and DeGroot suggest that the choice s_1 = "observed" and s_2 = "unobserved" (which they and Professor Kadane call LF_{obs} in BDK) has logical preeminence as a definition of likelihood; then $\ell_x(s_2)$ represents precisely what was learned from the observation of x, unconfounded by any given information about s_2. We again agree; it was only the sociological concerns mentioned in our first paragraph above that kept us from so defining the likelihood function in general.

Further repetition of the insights of Bayarri and DeGroot would be unnecessarily duplicative. Suffice it to say that we agree that non-Bayesians can have a very difficult time defining and interpreting the likelihood function; and once they pass this hurdle, they still must contend with the LP.

REPLY TO PROFESSOR HILL

It would seem rather foolish of us to question Professor Hill's interesting discussion at all, because he seems to feel that we do not go far *enough* in our support of the LP. First we would like to clear up that misimpression (we are fully as enthusiastic as he is concerning the applicability of the LP), and then proceed to the deep issue he raises concerning use of improper, or proper but finitely additive, priors.

From Professor Hill's comments (and also those of Professor Le Cam) it is clear that we did not express ourselves clearly in the Monette-Fraser, Stone, and Stein examples, with regard to the role of frequentist measures and our own conditional perspective of statistical analysis. (This lack of clear expression was primarily due to our concentration on using the examples to indicate the necessity for some type of Bayesian processing of likelihood functions.) Our discussion of frequentist measures was motivated partly by the fact that the examples were historically developed in that fashion, and partly to indicate that conditional (proper) Bayesians will naturally overcome the difficulties involved, even from the frequentist perspective. We also do believe that there can be value in frequentist measures as possible warning signals that care must be taken in the Bayesian analysis. However, we by no means meant to imply that (because of a Bayesian-frequentist conflict) one must *necessarily* change the Bayesian analysis.

These points are, perhaps, best illustrated by the Stein example, in that it is well recognized that (for data from univariate normal models) the uniform improper prior is typically quite satisfactory. It is typically satisfactory, however, *only* because σ is usually small enough that true prior beliefs can be approximated by the uniform prior. The bad frequentist performance of the uniform prior in the model (5.3.3) should be a warning that the adequacy of the uniform approximation to prior beliefs should be investigated, and indeed such an investigation would usually indicate that the approximation was *bad*; this would be the conclusion unless y was very small. The frequentist measure, here, is actually superfluous, however. The conditional Bayesian would naturally use a uniform prior (as a good approximation) when y/d was very small, and would recognize (if he had any prior information *whatsoever*) its inadequacy for typically large y/d, simply because the likelihood function would then be much more diffuse than even very vague prior information. No knowledge of frequentist properties, or of differing properties of scale and location parameters, is necessary to behave sensibly. Also, we in no sense recommend changing the prior as the model changes. If one's prior opinions truly are diffuse over the range of the likelihood function, by all means use the uniform prior, no matter what the model. We simply do not feel that this will be the case for the model in (5.3.3), however, unless y happens to be exceptionally small. (Likewise, we judge that the uniform prior will usually be inappropriate for normal models which have monstrously large variances.) There is no incompatibility with the likelihood principle here, since the "adequacy of the approximation" can be judged simply by looking at the likelihood function.

The major issue raised by Professor Hill concerns the need for conglomerability, or alternatively the concern that need be felt when the frequentist answer completely contradicts the posterior Bayesian answer. Specifically, in the Monette-Fraser example Professor Hill argues that the uniformly bad frequentist performance of analysis based on the finitely additive uniform prior is not operationally meaningful, because the sample space is, in reality, always bounded. The issue here is not directly related to the likelihood principle, but is another aspect of the possible problems caused by the use of infinite models to approximate reality. If the sample space in the Monette-Fraser example is bounded by N, then certainly the uniform prior becomes permissible, since one can actually simply choose the proper discrete uniform prior on 0 to 2N. We do feel, however, that the subjective assessment of uniformity on 0 to 2N would rarely be reasonable in practice, precisely because the use of the infinite model as an approximation would typically be due to the belief that no X, so large as to be unmeasurable, would actually occur; this *implies* a prior belief that θ could not be extremely large. In general, we would view a uniform conflict between frequentist and Bayesian measures as an indication that either the approximation of an infinite model was inappropriate, *or* the use of the finitely additive prior was inappropriate.

We do, of course, feel that all sample spaces are actually finite, and that (for virtually any problem) one could actually provide a (perhaps overly large) finite sample space. Do examples of the type we are discussing exist for finite sample spaces? If so, such would seem to provide a counterexample to Professor Hill's argument. If not, one could indeed not object, philosophically, to the use of finitely additive measures. There would, however, remain pragmatic questions concerning the practicality of using finitely additive priors (as opposed to countably additive priors) to approximate prior beliefs, but that is an issue for another time and place.

We have long been admirers of Professor Hill's careful treatment of the random effects analysis of variance model, and do not really disagree with his comments here. If we observed a likelihood function, over the range of which our prior was very diffuse, we would have no qualms about using the uniform improper prior. If, however, the uniform prior leads to a *procedure* with bad frequentist properties, we would infer that the uniform prior was a *poor* approximation to our prior beliefs for most of the likelihood functions that could be encountered, and would be loathe to implement it in, say, a "routine" computer package.

Our view on this matter is partly tied to the discussions surrounding Examples 16 and 37. Good frequentist performance will often give some assurance that a type of conditional Bayesian analysis is moderately robust, while bad frequentist performance of such an analysis is often an indication of nonrobustness. Such implications are by no means certain, and use of frequentist verification may often be an inefficient way of investigating robustness, but we should not dismiss any available aids. In this we also perceive at least partial agreement with Professor Hill, as witnessed by his numerous papers on the matter (referenced and discussed in Berger (1984e)).

REPLY TO PROFESSOR LANE

Before considering the two deep issues raised by Professor Lane, we would raise one minor quibble. His second paragraph consists of a listing of "bail-out options" for statisticians who choose not to follow the LP. A major purpose of the monograph was to argue the inadequacy of such bail-outs. Professor Lane does not make his views on such bail-outs clear, although presumably, as a Bayesian, he does not accept their validity (for perhaps reasons other than those given in the text).

The two main issues raised by Professor Lane are (i) the adequacy of the model paradigm and usefulness of the LP within it, and (ii) the fact that the LP ignores the Basic Tenet of Bayesianity, namely that inference should consist of the quantification of uncertainty. In our analyses of these issues it is particularly important to realize that we perceive little, if any, disagreement between us and Professor Lane concerning the correctness of the Bayesian pardigm for statistics. We do differ, however, in our opinions concerning the most convincing and practically useful way in which this paradigm should be presented. We emphasize the basic agreement because, all too often, non-Bayesians use these rather mild disputes between Bayesians to reject the entire Bayesian paradigm.

Professor Lane only briefly mentions the second issue, that quantification of uncertainty should be the goal of inference. This tenet seems almost self-evident (even though it is not accepted by the bulk of the advanced statistical community), and indeed the LP does not directly incorporate it. An alternate phrasing of the tenet, however, is that statisticians should treat known quantities as fixed and treat unknown quantities probabilistically. The LP does deal with the first half of this phrasing, treating the known data, x, as fixed for inference, while at least treating θ as variable (if not as a random quantity). Hence the LP embodies a major part of the Basic Tenet of Bayesianity.

We have several reasons for approaching the Bayesian paradigm through the LP, rather than through acknowledgment of the Basic Tenet. The first is sociological, and is partly due to the current state of statistics. Two prevelant notions in this "current state" are that the frequentist paradigm provides a satisfactory underpinning for statistics, and that Bayesian analysis is unacceptable because of its prior inputs. It is because of these notions that the majority of statisticians would reject the Basic Tenet, and that direct arguments for Bayesianity often make little headway. Note, however, that the LP directly impunes the first notion, while avoiding the biases of the second notion.

Of course one can argue that transient sociological concerns should not be the basis for judgement, but even from a strictly scientific perspective there is some doubt as to the correct route to take to the Bayesian paradigm. Direct arguments for the Basic Tenet involve some variety of coherency arguments, based on axioms of rational behavior. Such axioms are by no means above criticism. For instance, the arguments listed in Section 3.7, that have been raised against the common "betting scenario," are not easy to dismiss (see also Le Cam (1977)). Also, even if all such axioms are accepted, the fact that the only "rational" analyses are those compatible with some Bayesian analysis does not logically imply that the only acceptable way to do statistics is to write down a prior distribution (which can never be more than an approximation to prior beliefs) and perform a Bayesian analysis.

Although the LP is also subject to axiomatic and operational criticism, it has several advantages in these regards. The first is that its axiomatic basis is compelling to most people. The WCP is compelling to almost everyone, and the SP is an integral part of most existing statistical paradigms. We do acknowledge that the SP is not really "obvious," and indeed went to considerable effort in Sections 3.6 and 3.7 to justify the principle (and not just for the "betting scenario"). The simple fact remains, however, that very few statisticians will reject either of these axioms, while most seem unmoved by the coherency axioms.

As to the operational criticism, the LP would again seem to have an edge, precisely because it does *not* provide a final answer and can hence be more specific in its *partial* answer. The coherency approach provides only the vague general requirement that substantial inconsistency with some Bayesian analysis should be avoided. The LP is, on the other hand, specific in its recommendation to utilize only the observed likelihood function, even though it does not address the question of how this is to be done. And from a purely pragmatic viewpoint, this first step may well be the most important step of all. The reason is that, in practice, one often spends the greatest effort in model selection and verification; the knowledge that one need only consider the *observed* likelihood function can simplify this task enormously. Indeed, it is not unusual for the choice of a prior on model parameters to be of such secondary importance that one never gets beyond "playing with likelihoods."

Professor Lane does point out that the "standard" decomposition into model and prior is often artificial, and so should not be a part of statistical foundations. While sympathetic, we view such decompositions as essential practical simplifying devices, necessary to achieve progress. One usually progresses on a hard problem by identifying simple components that can first be analyzed separately, and then combined together. Although we agree that such decompositions are not always appropriate, their pragmatically central role to statistics is hard to deny. And the fact that the LP provides so much insight into what is probably the most crucial component of the decomposition, gives it considerable appeal.

In summary on this issue, the coherency approach to the Bayesian paradigm has many admirers (ourselves among them) and can perhaps claim a logical ascendancy over the LP approach, but (for the reasons mentioned above) we feel that the LP approach has had, or at least can have, a larger impact. The quotation on p. 2 from L. J. Savage is revealing in this regard, coming from an ardent admirer of coherency. (More complete discussions of this issue can be found in Berger (1984b) and Berger (1984e).)

It was perhaps unfair to spend so much time on this issue, given that Professor Lane only briefly mentions coherency. However, we feel that it is important to view Professor Lane's objections to the LP in the larger perspective of alternative approaches to the Bayesian paradigm.

Let us now turn to Professor Lane's second issue, specifically the criticisms about the "model" paradigm and the applicability of the LP within it. The first issue Professor Lane raises is that of the interpretation of θ, and the question of applicability of the LP unless θ is a "real" physical parameter. We wanted to avoid the philosophical problems inherent in any discussion of the meaning of parameters, but in retrospect should have spent more time on the issue. The reason is that, while of course the LP will apply if θ is a real physical parameter (in some sense), it also applies in the much more common situation where θ is only defined by some aspect of the experiment. For instance, a very large part of statistics deals with situations involving a series of (approximately) i.i.d. observations X_i. The parameter θ is often implicitely defined by the assumed density (say), $f_\theta(x_i)$, for the observations, and is *not*, as Professor Lane implies, necessarily identified with the overall $\{P_\theta\}$, which could also involve other aspects of the experiment such as the stopping rule, possible censoring, and so forth. We could consider any number of experiments with the same implicitely defined θ, but with different $\{P_\theta\}$, and apply the WCP and SP to deduce the LP. Indeed a major purpose of the LP is to show that features of $\{P_\theta\}$ which are *irrelevant* to the implicit definition of θ are ignorable in the analysis. Professor Lane's three interpretations of θ do not cover this case, which we would call the case of major practical interest.

Of course, we did not mean the LP to apply in Professor Lane's case b), where Θ is just an index set, and specifically warned against this on several occasions. (The entire mixing setup makes no sense if the parameters in the two experiments can differ.) Our failure to carefully define θ in examples was admittedly sloppy, but was based on the desire to avoid complex philosophical issues that are of uncertain practical import. (Convincing a practicing statistician, who routinely uses models, to base his analysis on the observed likelihood function is a significant practical step. Informing him that his model parameters really have no meaning is unlikely to cause much improvement in his statistical practice.)

Professor Lane next questions the value of inference about model parameters, arguing that predictive inference about future observables is of most concern. We do not dispute this point, but, as Professor Lane acknowledges, we do handle predictive inference by incorporating the future observable in θ. The complaint that the LP does not then say how to eliminate θ is one of the arguments we use for Bayesian implementation of the LP, but the complaint in no way limits or casts doubt on the LP. Professor Lane may prefer the de Finetti approach, which allows direct dealing with predictive inference, but, as discussed earlier, we feel the model-based "half-way house" is generally a pragmatic necessity. It is enormously difficult to attempt directly to ascertain such complicated things as predictive distributions. Even inventing crude models and artificially creating model-prior separations will, we feel, serve predictive statisticians best in the long run. Professor Lane does raise the valid point that our emphasis throughout the text on model parameter inference, itself, may be misleading. Our only defense is the essential impossibility of sensibly discussing predictive inference outside a Bayesian framework, combined with our desire to minimize Bayesian involvement (for already mentioned sociological reasons).

We have tried to describe accurately the reasons for our preference for the LP approach to Bayesianity. Admittedly this preference may be due to our traditional probabilistic and statistical background (with its model orientation), but, on the other hand, the alternative developments have not managed to produce any broadly useful new practical methodology. There is real danger in letting philosophical games obscure the *practical* realities of the situation. (For instance, the coherency game of "betting" serves to give various sound meanings to probabilities, but it seems completely backwards in its application: people decide how to bet by *first* determining probabilities, usually through some comparative likelihood method.) A philosophical game that can be played to support the LP, foundationally, is the "finite sample space" game (see Section 3.6.1). Reality always has a finite sample space, and the LP always applies to the implied "model." This formulation has little operational significance, however, and so we do not view it as a serious argument for the LP approach.

In conclusion, although we certainly support, and indeed find philosophically enlightening, approaches to Bayesianity based on coherency, our own preference is for the LP approach.

REPLY TO PROFESSOR LE CAM

The major and probably most important point made in Professor LeCam's interesting discussion is that we should be "a bit unprincipled." He sees value in both classical methods and the LP. As "tools" in the statistician's toolkit, we agree that there is possible value in classical methods, although we would tend to prefer Bayesian tools, if available. The choice of a tool is, however, not really the question addressed in the monograph. The purpose of the monograph was to attempt to clarify the more fundamental question: What should the statistician be using his tools for? We believe the vast majority of statistical users want to know "the evidence about θ from E and x," and indeed will likely be unable to assign any other meaning to a statistical conclusion. Because of the demonstrated conflict between this goal and the frequentist goal of procedure performance (except, of course, for the various discussśd exceptions, such as experimental design), we feel that this question of *purpose* can not be ignored. And while a variety of tools may be useful in reaching our stated goal of the determination of conditional evidence (even frequentist tools may be useful - see Section 5.4 and also Berger (1984b) and Berger (1984e)), we would argue that the value of the tool must be related to this ultimate goal. The big stumbling block in the long-running controversy in statistics has been the lack of separation of purpose and method.

A recurring theme in Professor LeCam's discussion is the issue of communication of statistical evidence. Indeed, because we briefly indicate in Chapter 5 that we feel it necessary to be Bayesians (and hence produce priors and posteriors), Professor LeCam intimates that we have "argued...into a corner." Our interpretation, however, is that, even if communication of evidence through Bayesian measures is deemed unappealing, it is a *scientific necessity*, unless one is willing to sacrifice the goal of communicating the actual evidence obtained about θ. Of course, the Bayesian situation (as regards scientific communication) is not nearly as bad as many non-Bayesians think; the spectre of being forced to accept someone's unreasonable prior distribution is not really an issue. Good Bayesian reporting can be done with a variety of strategems involving the presentation of the conclusion for a wide variety of priors (c.f. Dickey (1973)). And simply presenting likelihood functions or, perhaps somewhat better, posterior distributions for noninformative priors can be viewed as a reasonable conditional communication device. Of course, such are not traditional in scientific journals, but we all know of a number of "traditions" concerning statistical reporting in scientific journals that we would all gladly retire.

As to Professor LeCam's feeling that one should report all possibly relevant data about an experiment, no subjectivist would think of disagreeing. After all, a subjectivist is (theoretically) responsible for producing his likelihoods (as well as priors), and *all* data about the experiment could be relevant to this enterprise. Of course, the LP does say that in processing all this information the conditional viewpoint should have primacy.

Professor LeCam feels that a major flaw in the LP axiomatics is the assumption that Ev(E,x) exists. Since we allowed Ev(E,x) to be *anything*, any collection of conclusions or reports,we are unclear as to the exact objection. (One surely must make some report.) All the axiomatics say is that if one processes information in violation of the LP, perhaps by reporting frequentist error probabilities, then one is behaving in violation of either the WCP or SP or both. It is, perhaps, conceivable that, for each experiment, one could process information in a completely new way, so that one's Ev(E,x) would be continually changing, and so that no violation of the WCP or SP could be established. This, however, is not realistic: as statisticians we are bound to standardize many of our analyses, or at least parts of many of our analyses. The text argues that any such standardized methods of processing information should be in accord with the LP.

Professor LeCam is certainly correct in his comment that our passage from the LP to Bayesianity is much weaker than the argument for the LP. We felt little need to rigorously justify this final step, mainly because we feel that it is belief in the LP that is the major hurdle; it is hard to avoid becoming a Bayesian after fully accepting the LP.

Professor LeCam feels that we make a direct appeal to frequentist ideas in our attempt to resolve the Stein example. We clearly did a bad job in the example, of explaining our position, because Professor Hill likewise sees us as resorting to frequentist reasoning. The passage to which Professor LeCam refers was an attempt to explain *to frequentists* why, as

Bayesian conditionalists, we would not suffer from a frequentist perspective. The conditional Bayesian analysis we discussed in no way depended on frequentist evaluations, however. For a more lengthy discussion of this point, along with a brief description of the role conditional Bayesians can ascribe to frequentist measures, see our reply to Professor Hill.

The final issue raised by Professor LeCam is that of application of the LP when only "approximate likelihoods" are available. We have seen no evidence to indicate that the need to use approximations with the conditional approach causes any more problems than the use of approximations with any other approach. In the example of n independent Cauchy observations, we would of course prefer use of the exact observed likelihood function, but if n were enormous and we had technical problems in calculating and using the exact likelihood, we would certainly consider using the $\mathcal{N}(\theta, \frac{2}{n})$ approximation. But we would use the *observed* likelihood function from this approximation as the experimental input to evidence, not frequentist measures calculated by averages over the normal approximation. Without knowing the specific problem one cannot safely recommend specific priors. When n is large, however, prior information will typically be vague compared with the likelihood function, so use of the noninformative uniform prior would be a reasonable first approximation.

REPLY TO PROFESSOR LE CAM'S SECOND EDITION COMMENTS

We are sympathetic to Professor LeCam's position, that attempting to summarize a complex situation by the pair (E, x) may omit much that is relevant. Thus we have always been interested in attempts to depart from the usual statistical framework of probabilistically-modelled experiments (though we have yet to see an alternative framework that works better). Note, however, that virtually all of classical statistics is based on considering particular notions of Ev (E, x). Thus Professor LeCam's observations would seem to apply equally well to all standard statistical concepts. He does mention the possible need for "introducing in the system a variety of concepts that go beyond pairs (E, x)"; the argument to abandon (or at least extend) the usual statistical framework is too big for us.

Perhaps Professor LeCam is making the smaller logical point that principles (e.g., the LP) that are deduced within a too narrow framework are not necessarily valid in the correct framework. The constructive side of the LP ($\ell_x(\theta)$ summarizes what is needed from (E, x)) may thus be questioned; but the destructive side of the LP, that measures based on (E, x) which are incompatible with the LP (such as frequentist measures) are contraindicated, seems intact. After all, a frequentist measure based on (E, x) should certainly be able to pass an evaluation in its own domain. If it fails there, it is hard to imagine that it would be good in an enlarged domain.

We have been a bit overly dogmatic to emphasize our basic views. At the same time, our position, stated in Sections 5.4 and 5.5, bears a certain similarity to LeCam's, in that we also do not feel that all our actions "*must* abide by the LP." Our own summary position, however, is that abiding by the LP is a generally good guideline, and that major deviations from the LP are highly suspect.

Additional References In The Discussion

Listed here are those references in the discussion that are not in the regular bibliographies (pp. 143-159 and 160-160.2).

DICKEY, J. M. (1973). Scientific reporting. *J. Roy. Statist. Soc. B 35*, 285-305.

GEISSER, S. (1984). On the prediction of observables: a selective update. In *Bayesian Statistics II*, J. M. Bernardo, et. al. (eds.). North-Holland.

HILL, B. (1965). Inference about variance components in the one-way model. *J. Amer. Statist. Assoc. 58*, 918-932.

HILL, B. (1967). Correlated errors in the random model. *J. Amer. Statist. Assoc. 62*, 1387-1400.

HILL, B. (1980). Robust analysis of the random model and weighted least squares regression. In *Evaluation of Econometric Models*, J. Kmenta and J. Ramsey (eds.). Academic Press, New York.

HILL, B. (1985-86). Some subjective Bayesian considerations in the selection of models (with discussion). *Econometric Reviews 4*, 191-288.

HILL, B. (1987c). DeFinetti's theorem, induction, and A(n), or Bayesian nonparametric predictive inference. To appear in *Bayesian Statistics III*, J. M. Bernardo, et. al. (eds.). Oxford University Press, Oxford.

HILL, B. (1988). Bayesian data analysis. Technical Report, Departments of Statistics and Management Science, University of Michigan.

HILL, B. and LANE, D. (1983). Conglomerability and countable additivity. To appear in *Essays in Honor of Bruno de Finetti*, P. Goel and A. Zellner (eds.). North-Holland, Amsterdam.

JAMES, W. (1896). The will to believe. Reprinted in *Essays in Pragmatism*, Alburey Castell (ed.). The Hafner Library of Classics, Number 7 (1966).

LANE, D. (1981). Coherence and prediction. *Proceedings of the 43rd Session of the ISI*, 81-96.

LANE, D. and SUDDERTH, W. (1984). Coherent predictive inference. *Sankhya.*

SAVAGE, L. J. (1962). *The Foundations of Statistical Inference, A Discussion.* Methuen and Co., London.

STONE, M. (1979). Review and analysis of some inconsistencies related to improper priors and finite additivity. In *Proceedings of the 6th International Congress on Logic, Methodology, and Philosophy of Science*, L. J. Cohen et. al. (eds.). North-Holland, Amsterdam.