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.