DIFFICULTIES AND AMBIGUITIES IN THE DEFINITION OF A LIKELIHOOD FUNCTION

by

M. J. Bayarri and M. H. DeGroot
University of Valencia and Carnegie Mellon University
Purdue University

Technical Report #91-06C

Department of Statistics
Purdue University

February 1991

◇

DIFFICULTIES AND AMBIGUITIES IN THE DEFINITION OF A LIKELIHOOD FUNCTION

M. J. Bayarri
University of Valencia
and Purdue University

M. H. DeGroot
Carnegie Mellon University

Foreword

I precisely remember the day when, almost six years ago, I attended my first seminar in the United States. It was a seminar on the Likelihood Principle taught by Morris DeGroot out of Berger and Wolpert's monograph (Berger and Wolpert 1984; 2nd edition in 1988). This was the starting point of many lengthy and stimulating discussions on the subject. One recurring theme of these discussions was the rather amazing fact that, in spite of its widespread use in statistics, no general rigorous definition of the likelihood function seemed to ever have been given. We eventually concluded that such a definition could not be given, and that for any attempt to do so, examples could be found where the definition would work poorly or produce contradictions. These ideas were summarized in the paper Bayarri, DeGroot and Kadane (1988). Along the way, our work was enriched with fruitful and interesting conversations with James Berger and Robert Wolpert and we were invited to contribute to the discussion of the second edition of their monograph on the Likelihood Principle (Bayarri and DeGroot, 1988). Some few months later, Morris DeGroot was diagnosed with lung cancer and we sadly lost him on November 1989. This paper is intended as a unified review of our joint work on the subject.

1. Introduction

As with many other important statistical concepts, that of the likelihood function was introduced by Fisher (1921) and it was to play a decisive role in many approaches to Statistics, particularly the so-called "Likelihood Approach". This approach attempts to base inferences solely on an objective likelihood function and has given rise to a wide variety of so-called likelihood functions, developed to hopefully cope with an equally large variety of inferential aims. To mention just a few, and letting LF stand for likelihood function,

some of the different LF's that can be found in the statistical literature are: relative LF, marginal LF, conditional LF, partial LF, integrated LF, profile or concentrated LF, section LF, canonical LF, second order LF, several definitions of a predictive LF, several definitions also of a pseudo LF and, of course, various combinations of the above concepts. (See Bayarri, DeGroot and Kadane, 1988; and Berger and Wolpert, 1988, for references.)

At this point it should be clarified that we will not try to develop a brand new definition of a likelihood function to be added to this long list. As a matter of fact, we argue that there is not such a thing as "a" likelihood function that can be unambiguously defined in all statistical problems. Thus we will try to point out deficiencies of a "pure" likelihood approach to inference by pointing out some of the difficulties and ambiguities encountered when trying to define what a LF is. We will see that considerable subjectivity must be used in order to choose and efficiently use an "objective" LF.

In the next section we present some natural ways to define a LF and show, in a series of examples, that they can be badly inadequate. Pushing the situation to the limit leads to the conclusion that the only possible definition of a general LF would make it completely subjective: this is, of course, the Bayesian point of view. Does this render the Likelihood Principle inapplicable? In a sense, the answer to this question is "no" but only as long as we recognize that "same evidence in the data" does not mean "same inferences or decisions", so that we explicitly recognize that in many statistical problems, more complicated functions involving other factors apart from the LF may be needed in order to make inferences about quantities of interest. In this setting, a simple LF can, if it is desired, be defined so as to convey all the information in the data. We present such a LF in the final section of the paper.

2. Observables, Unobservables and Likelihoods

By now the reader may be wondering what is wrong with the usual and familiar definition of a likelihood function as being proportional to $f(x|\theta)$, considered as a function of θ for a given x. We shall argue that there is not a unique way of deciding what should be regarded as x and what should be regarded as θ. Different elections will result in different LF's and presumably in different inferences (unless a Bayesian analysis is carried out) and no election works well in every problem. The most popular choices seem to be the

following:

i) To take x to contain all the "random variables" in the experiment and to take θ to contain all the "fixed" parameters. This LF will be denoted by LF_{rv} so that

$$LF_{rv} \propto f(\text{random variables}|\text{parameters}). \tag{2.1}$$

ii) To take x to contain all the observed quantities in the given experiment and to take θ to contain unobserved quantities. We shall use the notation LF_{obs} to denote this definition of a LF. Thus

$$LF_{obs} \propto f(\text{observed}|\text{unobserved}). \tag{2.2}$$

Through a series of examples we shall show that none of these definitions is suitable as a general definition for a LF since both of them can result in inadequacies when dealing with particular problems. Other "natural" ways of defining a LF in particular examples will also be explored. They will be seen to also be inappropriate in some problems, while in other problems they seem to provide a very sensible LF that nevertheless cannot be accommodated in any general definition of a LF. The main difficulties arise, of course, in statistical problems in which additional variables are (or can be) incorporated into the basic analysis (as is the case in prediction problems), as well as in statistical problems in which additional parameters are (or can be) present in the specification of the statistical model (as is the case with nuisance parameters). But difficulties can also be present due to the impossibility of distinguishing between "variables" and "parameters" and to the impossibility of separating the "model" from the "prior". The argument can be forced to conclude that the only general definition of a LF is a useless one that makes the prior distribution the only carrier of all the information available (see Bayarri, DeGroot and Kadane, 1988, section 7).

The difficulties in defining a LF in the presence of nuisance parameters have been considered in the statistical literature in a number of papers and several methods have been proposed to deal with them, resulting in different likelihoods such as the marginal and conditional likelihoods of Kalbfleisch and Sprott (1970, 1973) or the canonical likelihoods of

Hinde and Aitkin (1987). To eliminate nuisance parameters, traditional approaches transform the problem in terms of appropriate statistics in a way that the nuisance parameters are no longer present in the new formulation; more direct approaches, based on the original model either maximize over the nuisance parameter or integrate it out (if the nuisance parameter happens to have a "distribution" as in random effects models, or if appropriate "weight" functions can be produced). Although difficulties and ambiguities are obviously present in this problem, in all of the approaches (except in the ones that integrate out the nuisance parameter) the nature of the nuisance parameter is clearly stated as a fixed (and therefore "given") although unknown value. The ambiguities that we wish to put forward are of a more subtle nature and we shall not discuss these problems any longer in this section. For lengthy discussions and references see Berger and Wolpert (1988) and Piccinato (1987).

For simplicity, we shall assume that all distributions that appear have density with respect to some fixed σ-finite measure and we shall use the symbol f to denote an arbitrary density without any attempt to distinguish among different densities by the use of subscripts or different symbols. The nature of the difficulties we shall be dealing with are clearly shown in the following statistical situation.

2.1 Observations Subject to Error

Consider a problem in which, for each value of some parameter θ, a random variable Y has density $f(y|\theta)$. Assume also that Y cannot itself be observed, but rather what we get to observe is a random perturbation X of Y. That is, we assume that our observation is a realization of the random variable X with conditional density $f(x|y,\theta) = f(x|y)$.

What is the likelihood function in this problem? Several possibilities exist. First, we must decide whether the unknown value y of Y should be included in the likelihood function. We could argue that, since y cannot and will never be observed, it is more appropriate not to include it in the definition of a LF and thus define:

$$LF_0 \propto f(x|\theta). \tag{2.3}$$

On the other hand, since the basic formulation of the problem is in terms of y, we may think that it should enter the LF. (This would obviously be the case if we were

interested also in y.) In this case, when deriving a LF, we must decide whether y should be considered as a realization of a random variable and taken to be a component of the vector in front of the vertical bar, resulting in a LF of the form:

$$LF_{rv} \propto f(x, y|\theta), \tag{2.4}$$

or alternatively, y may be considered as an unobserved, unknown quantity and thus taken to be a component of the vector behind the vertical bar, resulting in a LF of the form:

$$LF_{obs} \propto f(x|y, \theta) = f(x|y). \tag{2.5}$$

What LF should be used is not, in principle, clear, and a subjective judgement has to be made to decide which one to use in a given problem. Since these likelihoods are usually quite different, inferences based on them will differ. For simplicity, we will stress these differences by comparing the different MLE's obtained from the various LF's under consideration (this should not be taken as a defense of the MLE as an inferential procedure on our part).

Suppose, for example, that the distribution of Y given θ is exponential and that the distribution of X given y is also exponential, that is:

$$f(y|\theta) = \theta e^{-\theta y} \text{ for } y > 0,\ \theta > 0, \tag{2.6}$$

and

$$f(x|y) = y e^{-yx} \text{ for } x > 0,\ y > 0. \tag{2.7}$$

Then, the different likelihoods proposed above would be

$$\begin{aligned} LF_0 &\propto \int_0^\infty f(x|y) f(y|\theta) dy = \frac{\theta}{(\theta + x)^2}, \\ LF_{rv} &\propto y\ \theta e^{-y(\theta+x)}, \\ LF_{obs} &\propto y e^{-yx}. \end{aligned} \tag{2.8}$$

In this example, LF_0 provides the MLE $\hat{\theta} = x$, which can be a sensible estimator for θ, but it is totally uninformative about the unknown value y of Y. LF_{obs} gives the estimator $\hat{y} = 1/x$, which might be regarded as a sensible estimator for y, but it is totally

uninformative about θ and completely loses the relationship between y and θ reflected in $f(y|\theta)$. Finally, LF_{rv} does contain in its formulation both y and θ, but yields the useless MLE's $\hat{\theta} = \infty$ and $\hat{y} = 0$.

The Bayesian approach is based on the specification of the joint density $f(x, y, \theta)$. From there, all that has to be done is to condition on the observed value x and to integrate out θ or y or none if we are interested in making inferences about y or θ or both, respectively. It is thus completely irrelevant which of the possible factors that may be considered to form the joint $f(x, y, \theta)$ are taken to form a LF, and inferences will, in any case, be identical (the rest of the factors will then, of course, be called "prior"). Thus, if we factor $f(x, y, \theta)$ as follows:

$$f(x, y, \theta) = f(x|y)\ f(y|\theta)\ f(\theta), \tag{2.9}$$

then LF_{obs} would be formed by taking just the first factor in the right hand side of (2.9), while LF_{rv} would also include the second factor. LF_0 is nothing but a further elaboration of LF_{rv} in which y gets integrated out.

In other words, the basic input for a Bayesian analysis of this problem is the joint $f(x, y, \theta)$, while the one for a likelihood-based analysis would be either $f(x|y, \theta)$ (as in LF_{obs}) or $f(x, y|\theta)$ (as in LF_{rv} or LF_0). Therefore, the election of a LF can be reduced to the election of where to put the vertical bar in the joint density. We will continuously turn back to this point, which will be explicitly taken in the final section. If we were to provide an entertaining title to this paper, it could have been called: "Where is the bar?"; Bayesians would have been then considered the healthiest ones, since for them the bar is lacking!

2.2 Prediction Problems

Prediction problems have always been especially difficult to handle from a likelihood point of view. Sophisticated methods, some of which cannot be implemented in all prediction problems, have been developed to deal with them. The interested reader is referred to the (different) likelihood methods of Lauritzen (1974), Hinkley (1976) and Butler (1986), as well as the discussions and references in Berger and Wolpert (1988). Butler (1988) gave what seems to be the most general definition of a LF, and this is the one proposed by Berger and Wolpert (1988, sec. 3.5.3) as the "practical" definition of likelihood in predic-

tion problems. It is indeed no wonder that it does generally yield sensible LF's since, in fact, it is the definition closest to the full joint density on which a Bayesian analysis would be based. Loosely speaking, what Butler (1988) proposes is to move as many quantities as possible in front of the vertical bar; quantities that are not of interest get integrated out. In so doing, some "parameters" are in some problems also put in front of the vertical bar, namely all parameters whose distributions are "known". According to Butler's definition, the likelihood function may well change as the inferential aim changes, which can complicate the analysis. Also, proper utilization of this LF is difficult since, depending on which distributions are given, it can vary from being a fully "traditional" type of LF to an integrated LF or even to a full posterior distribution. Finally, it leaves unanswered the problem of handling nuisance parameters whose distributions are not known.

We will not pursue these approaches further. Instead we will continue to point out the difficulties that also arise in prediction problems when we try to give a general definition of LF. Consider thus, a problem in which we observe a random variable X with density $f(x|\theta)$ and we are interested in predicting the value of some random variable Y with density $f(y|\theta)$, as well as in estimating θ. Here again we can consider three possible LF's. We could simply use $LF_0 \propto f(x|\theta)$ to produce an estimate $\hat{\theta}$ of θ and then predict Y from the "likelihood function" $f(y|\hat{\theta})$. This procedure is unsatisfactory because it does not take into account the uncertainty about θ in the prediction of Y. Alternatively, we could jointly estimate θ and predict Y from either

$$LF_{rv} \propto f(x, y|\theta), \tag{2.10}$$

or

$$LF_{obs} \propto f(x|y, \theta). \tag{2.11}$$

Each of the likelihoods in (2.10) and (2.11) has shortcomings. Thus if, as is most commonly the case, X and Y are conditionally independent given θ, then LF_{obs} reduces to $LF_0 \propto f(x|\theta)$ and Y disappears entirely. On the other hand, if we use LF_{rv}, then the MLE of θ can depend on which random variables we want to predict, as in the following example.

Suppose that $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_n)$, where X and Y are conditionally independent random samples from an exponential distribution with parameter θ,

as given in (2.6). If we let

$$s = \sum_{i=1}^{m} x_i, \quad \text{and} \quad t = \sum_{j=1}^{n} y_i, \tag{2.12}$$

then

$$LF_{rv} \propto \theta^{m+n} \ e^{-\theta(s+t)}. \tag{2.13}$$

It is easily found that the MLE's of t and θ are

$$\hat{t} = 0, \qquad \hat{\theta} = \frac{m+n}{s}. \tag{2.14}$$

The estimators in (2.14) are most unsuitable: the value $\hat{\theta}$ depends on the *dimension* n of the vector Y that we want to predict even though the prediction $\hat{t}$ itself does not at all depend on the data.

A Bayesian analysis of the problem would, as always proceed from the joint density

$$f(x, y, \theta) = f(x|\theta) \ f(y|\theta) \ f(\theta), \tag{2.15}$$

by completing whatever factors are not included in the specific LF that is chosen, so that choice becomes irrelevant. Butler (1988) suggests using the first factor in the right hand side of (2.15) for estimating θ and the two first factors for predicting Y (what to do then with θ is not very clear). The first approach that we mentioned used the first factor to produce an estimate $\hat{\theta}$ of θ which was then substituted for θ in the second factor thus producing a prediction of Y; (interestingly enough, this procedure will result in a prediction of Y which is identical to the one obtained with Hinkley's predictive LF.) LF_{obs} is formed by taking the first factor in (2.15), and LF_{rv} the first two factors in (2.15).

2.3 Separating the Model from the Prior

In the previous examples we have seen that, when expressing the joint density as "the likelihood times the prior", different factors can be taken to form different "likelihoods" and the rest of them would then form what could be called different "priors". In this section we will stress the fact that the distinction between model and prior is not at all clear and therefore it is not clear which factors should be incorporated to form the LF.

Consider a discrete-time Markov Process $X_1, X_2, \ldots, X_n$ with joint conditional density $f(x_n|x_{n-1},\theta)\cdots f(x_2|x_1,\theta)$ given $X_1 = x_1$ and some parameter θ. Here X_1 is taken to be the earliest state in which we are interested (it does not have to be the "initial" state of the process). Suppose that we observe the state $X_n = x_n$ and are interested in estimating $\phi = (X_1, \ldots, X_{n-1})$. We can again consider three different LF's. First consider

$$LF_{rv} \propto f(x_n, \phi|\theta) \propto f(x_n|x_{n-1},\theta)\cdots f(x_2|x_1,\theta)\ f(x_1|\theta). \tag{2.16}$$

It should be noted that this LF includes the factor $f(x_1|\theta)$ which is traditionally considered as part of the prior distribution for the unknown value of the state X_1 and the parameter θ.

$$LF_{obs} \propto f(x_n|\phi,\theta). \tag{2.17}$$

This LF reduces to $f(x_n|x_{n-1},\theta)$ and, therefore, the rest of the factors that appear in (2.16), most of which are traditionally taken to be part of the "model", should be here considered as part of the prior.

Finally, consider

$$LF_0 \propto f(x_n, \ldots, x_2|x_1,\theta). \tag{2.18}$$

This LF seems to be the one that is most commonly regarded as the appropriate LF for this problem. It should be noted, however, that it does not conform to any general definition of a LF since it contains "random variables" as well as unobservable on *both* sides of the vertical bar.

Similar problems arise when hierarchical models or different reparametrizations are considered (Bayarri, DeGroot and Kadane, 1988). As a final comment, consider all the statistical problems in which the sample size n is not fixed in advance and does carry information about θ. In these problems, a factor of the form $f(n|\theta)$ is usually incorporated into the LF. Whereas in some problems there would be certain agreement on the form of $f(n|\theta)$ there are a wide variety of problems in which $f(n|\theta)$ would be considered to be highly subjective and thus part of the prior distribution. Notice, nevertheless, that it looks like an ordinary "model" since n gets observed.

2.4 Unobserved Variables

The ambiguity in the definition of the LF that we have been discussing becomes especially noticeable when we consider unobserved variables that may be present in any given problem. One example was given in subsection 2.2, where the MLE of the parameter θ (derived from LF_{rv}) in an exponential distribution depended on how many future observations we wanted to predict. The situation can be even worse when we introduce unobserved variables in which we are not primarily interested ("nuisance" variables) and very wild examples can be built in which both LF_{obs} and LF_{rv} work very poorly in the sense that the inferences may change dramatically by *not* observing random variables that we can obviously consider in any problem (see Bayarri, DeGroot and Kadane, 1988, sec. 5). Besides, it is important to realize that these difficulties cannot be removed since there are *always* a wide variety of unobserved variables that can be introduced. In fact, it is obvious that in any given problem there are many more unobserved variables than observed variables.

3. The Evidence Provided by the Data

The difficulties and ambiguities discussed above derive mainly from the (hopeless?) attempt of basing a statistical analysis solely on an "objectively" defined LF with no formal explicit role for subjectivity in the problem. We have tried to argue that no such definition is possible. On the other hand, and taking as the basic line of reasoning the Bayesian argument, if one is prepared to recognize that other factors might have to be added to the LF in order to form a more complicated function in which inferences will be based, then a very simple LF can be argued to convey all the information contained in the data.

In complete generality, we can consider in a statistical problem the observation x, and a pair of random variables (or random vectors) Y, Z that are not observed in the given experiment where Y is of interest and Z nuisance. Similarly, we can consider the vector parameter as formed by components θ and ω, where θ is of interest and ω is nuisance. The basic purpose of a LF 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. In trying to derive a simple and meaningful LF to serve this

basic purpose, it is enlightening to consider the Bayesian approach to the learning process, which we shall openly adopt in the rest of the paper.

From a Bayesian point of view, all the relevant information about the quantities of interest is contained in $f(y, \theta|x)$. The way in which $f(y, \theta|x)$ is derived and whether or not Bayes theorem is at all used is irrelevant. As a matter of fact, if the design of the experiment is not under consideration and if we were perfectly trained and coherent in our learning process we could simply wait until x is observed and then assess the density $f(y, \theta|x)$ 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 $f(y, \theta|x)$ in the form:

$$f(y, \theta|x) \propto f(y|x, \theta)\ f(x|\theta)\ f(\theta). \tag{3.1}$$

In some problems, there could be general agreement about the form of both $f(x|\theta)$ and $f(y|x, \theta)$. This general agreement can then be taken to mean that the form of these two densities is "given" or "known" so that it would make sense to define a LF as being proportional to their product, that is:

$$LF = LF_{rv} \propto f(x|\theta)\ f(y|x, \theta) = f(x, y|\theta). \tag{3.2}$$

This is just one of the general definitions of a LF attempted in the last section, namely the one in (2.1) derived from the conditional density of the "variables" given the "parameters".

As we have argued in the preceding sections, we do not believe there is a clear-cut distinction between unobserved variables and parameters. Hence, we regard the LF in (3.2) as unsuitable as a general definition. Besides, the form of (3.2) relies on the density $f(x, y|\theta)$ being given or agreed upon. If we rewrite it as

$$f(x, y|\theta) = f(x|y, \theta)\ f(y|\theta), \tag{3.3}$$

this agreement implies that there must be agreement about both factors on the right-hand side of (3.3), or, otherwise stated, that both factors could be considered as "given". It nevertheless often occurs that there is general agreement about the form of $f(x|y, \theta)$, while the form of $f(y|\theta)$ is considered as highly subjective. Would this be the case, a LF for y and θ would simply be given by

$$LF \propto f(x|y, \theta). \tag{3.4}$$

Notice that both (3.3) and (3.4) are functions of x, y and θ and that there is no way to tell from the function alone whether the factor $f(y|\theta)$ has or has not been included in the LF. Hence, in order to be able to use the LF to make inferences (or calculate posterior distributions) we must know not only the function itself but also which factors have been used to derive it, in clear contradiction with the basic role of a LF as expressed at the beginning of the section.

Let's pursue further the basic steps in the statistical reasoning behind the parametric model building approach. So far, all we have included in our formulation are the observation x and the quantities of interest Y and θ. In many problems, however, the densities $f(x|\theta)$ and $f(y|x,\theta)$ can still be difficult to specify or can still be judged to be 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 θ and the "sample space" of y. Thus, a "nuisance parameter" ω is introduced so that $f(x|\theta,\omega)$ is easier to assess and/or interpret, or reaches a wider agreement among others than $f(x|\theta)$ does, or both. In a similar fashion and with similar goals, a "nuisance" variable Z may be conveniently introduced so that the assessment of $f(y|x,\theta)$ is derived from the assessment of $f(y,z|x,\theta,\omega)$. As a result, (3.1) now becomes

$$f(y,\theta|x) \propto \int\int f(y,z|x,\theta,\omega)\ f(x|\theta,\omega)\ f(\theta,\omega)d\omega\ dz. \tag{3.5}$$

This formulation emphasizes the fact that the "nuisances" z and ω are not to be considered as quantities that are regretfully present in our models, which we have to somehow get rid of. On the contrary, they are very convenient quantities that we have carefully selected so as to help us to build models and to achieve agreement about those models. The traditional term "nuisance" for ω is most unfortunate and does clearly not convey this idea; a more appropriate name might be auxiliary parameter (and auxiliary variable for Z, if such a distinction is at all wished).

If we have been successful in our selection of the auxiliary ω and z, then there will be a general agreement among others on the form of $f(x|\theta,\omega)$ and $f(y,z|x,\theta,\omega)$. As it is customary in statistics, when agreement is reached, the densities are regarded as given, so that it makes sense to consider (as in Berger and Wolpert, 1988) a LF in these problems

to be their product, that is:

$$\begin{aligned} LF &\propto f(y,z|x,\theta,\omega)\ f(x|\theta,\omega) \\ &= f(x,y,z|\theta,\omega). \end{aligned} \tag{3.6}$$

When x is a vector of observations, a typical way in which a convenient choice of the auxiliary parameter ω can simplify the density $f(x|\theta,\omega)$ is making the components of x conditionally independent (usually, i.i.d.). More importantly, a convenient choice of ω may make y and z conditionally independent of x given θ and ω. In this case, $f(y,z|x,\theta,\omega)$ reduces to $f(y,z|\theta,\omega)$, and the LF in (3.6) adopts the familiar form given by

$$LF \propto f(x|\theta,\omega)\ f(y,z|\theta,\omega). \tag{3.7}$$

Notice that, regardless of whether the density $f(y,z|\theta,\omega)$ is "given" or "subjective" it does *not* at all involve the data x, and thus all the *evidence in* x about the unknowns is contained in the first factor $f(x|\theta,\omega)$ in the right-hand side of (3.7). Thus, we believe that it should be the *only* factor to be included in the LF. The inclusion of other functions of the unknowns, such as $f(y,z|\theta,\omega)$ or the prior $f(\theta,\omega)$ seems artificial. Therefore, in this situation, we recommend the use of the simplest LF, namely

$$LF \propto f(x|\theta,\omega), \tag{3.8}$$

which, because of conditional independence, can be also expressed as:

$$LF \propto f(x|y,z,\theta,\omega). \tag{3.8}$$

(3.8) can be recognized as the LF introduced in (2.2) and denoted by LF_{obs} because it is derived from the conditional density of the observations given the unobserved quantities (notice that, unlike variables and parameters, observed and unobserved can always be distinguished from each other). This likelihood function is in basic agreement with the familiar "$f(x|\theta)$" when θ is taken to consist "of all unknown variables and parameters that are relevant to the statistical problem" (Berger and Wolpert, 1988, sec. 3.1).

More generally, every Bayesian analysis proceeds from a specification of the joint density $f(x,y,z,\theta,\omega)$. From this, all a Bayesian has to do is to condition on whatever

is observed and to integrate out whatever is not of interest. If we let s denote the set $\{x, y, z, \theta, w\}$ 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 $s_1 \cap s_2 = \varnothing$ and $s_1 \cup s_2 = s$, then $f(s) = f(s_1|s_2)f(s_2)$, the various likelihood functions discussed 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. Still other likelihood functions can also be shown to correspond to a $f(s_1|s_2)$ but eliminating the quantities that are not of interest by maximizing over them (instead of integrating). Thus, the choice of a LF corresponds basically to the choice of s_1 (together with the decision of whether to integrate out or maximize over the nuisances), or, in more colloquial terms, where to put the vertical bar in the joint density $f(x, y, z, \theta, \omega)$: in LF_{rv} it is put between z and θ, whereas in LF_{obs} it is put between x and y. The subset s_1 is always taken to contain x and usually, as in LF_{rv}, to contain other "variables" with given distributions, but it is not infrequent for "parameters" to also appear in front of the bar (see Butler, 1988). We claim that, in order to convey the evidence about the unknowns provided by the data, it is unnecessary to even include 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 conclude that the evidence in the data is conveyed most efficiently and most clearly by LF_{obs} as given by (3.8).

Some readers may be wondering whether we are now ignoring all the difficulties we just put forward in the previous sections when we discarded LF_{obs} as a suitable general definition of a LF. The answer is, of course, no. In Section 2, we tried to show that no general definition of a LF function can be given if by LF is to be understood the *only function* that it is needed *to make inferences* about θ and y. In this section we claim that LF_{obs} is the only carrier of all the *evidence provided by* x about θ and y. Other factors may have to be incorporated to LF_{obs} in order to effectively make inferences about θ and y. This distinction between information provided by the data and information needed to make inferences is always 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 $f(y|\theta)$ in (3.3). In this way, a large variety of inferential aims can be accomplished with just LF_{obs} rather than an equally large variety of likelihood functions.

Of course, we are aware that we have not removed all ambiguities since, unless a Bayesian analysis is carried out, it is still to be decided which factors should be added to LF_{obs} in order to obtain a suitable function on which to base inferences. Nevertheless, the formulation we are defending has the advantage, as it is often the case with Bayesian reasoning, of clearly and openly stating all the inputs of the problem. Thus, different readers facing maybe the prediction of different variables (or other inferential aims) or maybe agreeing with some of the factors involved but disagreeing about other factors, can judge (and use) which inputs do apply to their problems and which ones do not apply. Other, maybe more sophisticated, formulations not only suffer from more ambiguities and difficulties than our simple one but also, and more importantly, they can surreptitiously introduce information and inputs that not all readers would be willing to accept if explicitly revealed.

Acknowledgements

This work was supported in part by the Spanish Ministry of Education and Science under D.G.I.C.Y.T. grant number BE91–038.

References

Bayarri, M. J. and DeGroot, M. H. (1988). Auxiliary parameters and simple likelihood functions. In *The Likelihood Principle* (by J. O. Berger, and R. L. Wolpert), 160.3–160.7. Hayward, California: Institute of Mathematical Statistics.

Bayarri, M. J., DeGroot, M. H., and Kadane, J. B. (1988). What is the Likelihood Function? In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.), 3–27 (with discussion). New York: Springer Verlag.

Berger, J. O., and Wolpert, R. L. (1984; 2nd edition in 1988). *The Likelihood Principle*. Hayward, California: Institute of Mathematical Statistics.

Butler, R. W. (1986). Predictive likelihood inference with applications. *Journal of the Royal Statistical Society B* **47**, 1–38 (with discussion).

Butler, R. W. (1988). A likely answer to "What is the Likelihood function?" Discussion of Bayarri, DeGroot and Kadane. In *Statistical Decision Theory and Related Topics*

IV (S. S. Gupta and J. O. Berger, eds.), 21–26. New York: Springer-Verlag.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1**, 3–32.

Hinde, J. and Aitkin, M. (1987). Canonical Likelihood: a new likelihood treatment of nuisance parameters. *Biometrika* **74**, 45–58.

Hinkley, D. V. (1979). Predictive Likelihood. *Annals of Statistics*, **7**, 718–728 (corig., **8**, 694).

Kalbfleisch, J. D. and Sprott, D. A. (1970). Applications of likelihood methods to models involving large number of parameters. *Journal of the Royal Statistical Society B* **32**, 175–208 (with discussion).

Kalbfleisch, J. D. and Sprott, D. A. (1973). Marginal and conditional likelihoods. *Sankhyā A* **35**, 311–328.

Lauritzen, S. L. (1974). Sufficiency, prediction and extreme models. *Scandinavian Journal of Statistics* **1**, 128–134.

Piccinato, L. (1987). Insieme di verosimiglianza e parametri di disturbo. In *Rassegna di Metodi Statistici ed Applicazioni* **5** (W. Racugno, ed.), 71–91. Bologna: Pitagora.