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

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

$$\ell_x(y,w,\xi,\eta) = f(x|\xi,\eta)\ f(y,w|\xi,\eta). \tag{7}$$

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

$$\ell_x(y,w,\xi,\eta) = f(x|\xi,\eta), \tag{8}$$

can also be expressed because of conditional independence as

$$\ell_x(y,w,\xi,\eta) = f(x|y,w,\xi,\eta). \tag{9}$$

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 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

$$f(y,\xi|x) \propto \int\int f(x|\xi,\eta)\, f(y,w|x,\xi,\eta)\, f(\xi,\eta)\, dw\, d\eta. \tag{5}$$

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

$$\begin{aligned} \ell_x(y,\omega,\xi,\eta) &= f(x|\xi,\eta)\, f(y,w|x,\xi,\eta) \\ &= f(x,y,w|\xi,\eta). \end{aligned} \tag{6}$$

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.

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.