Chapter 7

REPLY TO DAWID, STONE AND ZIDEK

EDWIN T. JAYNES
Washington University

Since my paper reported the beginnings of a serious attempt to turn the marginalization discovery of DSZ into a useful part of statistical theory, and marginalization opens up a large area of interesting and important new mathematical problems, I had looked forward eagerly to the comments of DSZ, thinking them the ones in the best position to contribute constructive suggestions that would help us to get on with the progress that their work has started. It is depressing to see instead a commentary which ignores all my mathematical demonstrations, recognizes no progress at all, and reaffirms all the elementary errors that my work had sought to correct.

We shall not escape from this mire of confusion until we adopt a clear notation that defines unambiguously: (1) *What specific problem are we trying to solve*? (2) *What specific calculations do the rules of probability theory prescribe for that problem*?

I will try to accomplish this by setting down, in full generality and without abbreviations, the equations defining the different problems and different calculations being confused here. But first let me show good faith by acknowledging an ambiguity in my own notation. In the following (m), $[n]$, (Rk) denote respectively the mth equation in my presentation, the nth paragraph of the DSZ commentary on it, and the kth equation of this reply.

[10]. I had thought that if a function $f(y, z)$ is found to be independent of y, then it would be permissible to express that fact by writing $f(y, z)=f(z)$, as in the DSZ eq. (1.2). But DSZ now point out, quite correctly, that this notation allows a different, unintended interpretation. So, will the reader please replace (1) by

$$\frac{\partial}{\partial y}\, p(\zeta \mid y, z, I_1)=0. \tag{R1}$$

In fact, nothing else in my presentation will be changed by this.

It is regrettable that DSZ were not equally quick to accept my suggestion, i.e. that we follow Jeffreys's example by indicating explicitly, in a posterior pro-

Bayesian Analysis in Econometrics and Statistics, edited by Arnold Zellner
© *North-Holland Publishing Company, 1980*

bability symbol, the prior information I on which it is conditional; for the notational ambiguities in $p(\zeta|x)$, $p(\zeta|z)$, by failing to make it clear that we are concerned with two posterior distributions conditional on *different* prior information, were the original cause of this "paradox".

By B_1 we mean a Bayesian who adopts a model defined by a sampling distribution $p(y, z|\eta, \zeta)$ which has the property

$$p(z|\eta, \zeta)=\int p(y, z|\eta, \zeta)\mathrm{d}y=p(z|\zeta), \tag{R2}$$

but is interested in making inferences only about ζ. He assigns a prior $\pi(\eta, \zeta|I_1)$ and his posterior distribution for ζ is then

$$p(\zeta|y, z, I_1)=\frac{\int p(y, z|\eta, \zeta)\ \pi(\eta, \zeta|I_1)\mathrm{d}\eta}{\iint p(y, z|\eta, \zeta)\ \pi(\eta, \zeta|I_1)\mathrm{d}\eta\ \mathrm{d}\zeta}\ . \tag{R3}$$

B_2 is defined to be a Bayesian who makes inferences about ζ which do not take into account the components (η, y), but use only the sampling distribution (R2). His posterior distribution is therefore

$$p(\zeta|zI_2)=\frac{p(z|\zeta)\ \pi(\zeta|I_2)}{\int p(z|\zeta)\ \pi(\zeta|I_2)\mathrm{d}\zeta}\ . \tag{R4}$$

But the *dramatis personae* are now enlarged [3] to include a third personage, whom DSZ describe merely as B_1's second possible course of action. However, such verbiage only sets the stage for another step deeper into the mire from which we are trying to escape. Owing to the need to keep separate things clearly separate, I shall take the liberty of naming this third personage B_3. He is defined to be a Bayesian who uses B_1's full model including (η, y), but is given only the data z. His posterior distribution for ζ is then

$$p(\zeta|z, I_3)=\frac{\int p(z|\eta, \zeta)\ \pi(\eta, \zeta|I_3)\mathrm{d}\eta}{\iint p(z|\eta, \zeta)\ \pi(\eta, \zeta|I_3)\mathrm{d}\eta\ \mathrm{d}\zeta}\ , \tag{R5}$$

but in view of (R2), this collapses at once to

$$p(\zeta|z, I_3)=\frac{p(z|\zeta)\cdot\int\pi(\eta, \zeta|I_3)\mathrm{d}\eta}{\iint p(z|\zeta)\ \pi(\eta, \zeta|I_3)\mathrm{d}\eta\ \mathrm{d}\zeta}\ . \tag{R6}$$

Eqs. (R3), (R4) and (R6) then define the three specific *problems* that we are trying to solve. But the ambiguity of the specific *calculations* is yet to be faced.

Suppose that initially all the priors are proper and all three agree on the prior

for ζ:

$$\int \pi(\eta, \zeta | I_1) d\eta = \pi(\zeta | I_2) = \int \pi(\eta, \zeta | I_3) d\eta = \pi(\zeta). \tag{R7}$$

Then, evidently, B_2 and B_3 are necessarily in agreement. i.e. to withhold the data y from B_3 has the same effect on his inference as if he had stricken both (η, y) from his model from the start. But in general B_1 will disagree with them. It appears to me so clear as to be beyond the possibility of dispute that, at this stage, any difference in the conclusions of B_1 and his colleagues is due *solely* to the fact that B_1, in using both (η, y), is taking into account relevant information that they are *not* considering.

Presumably, no statistician worthy of the name would hold an inference which takes into account *more* information to be inferior to one based on *less* information. Presumably, nobody would, at this stage, attack B_1 for exhibiting any symptoms of "unBayesianity" or "impropriety". Clearly, B_2 (if his neglect is willful) is the guilty one.

Then what happens to these formulae in the case of improper priors? Part of the answer given by DSZ [11] is that B_3's posterior distribution has no unique form because of the "difficulty" of defining his marginal prior density of ζ (the integral over η diverges). But surely, there can be no such difficulty if B_3 assigns independent priors to (ζ, η) with proper $\pi(\zeta | I_3)$. Indeed, DSZ assumed such a prior in their example 1, although they now hold it to be meaningless when I do the same in (21) and (26). But that was only for convenience in formulating the problem that I wanted to consider. We can equally well consider general, nonindependent priors; for the "difficulty" that DSZ feared does not arise if we do our calculations correctly. This brings us to the second ambiguity.

The rules of probability theory tell us that, for any proper prior, the posterior distributions of B_1, B_2, and B_3 are given unambiguously by the calculations indicated in (R3), (R4) and (R6). But in general they cannot tell us anything at all if we try to insert an improper prior directly into formulae (R3) or (R6), which then become meaningless.

Yet the mathematical situation is no different from what arises if we ask: "What is the value of $f(x) \equiv \sin x/x$ at $x = 0$?" The rules of algebra cannot answer this by direct substitution into the formula; yet we give the answer unhesitatingly, with no thought of paradox or impropriety. By $f(0)$ we could mean only the *limit* of $f(x)$ as $x \to 0$.

It is exactly the same here. In general, by the posterior distribution for an improper prior, we could mean only the *limit* of the posterior distribution for a sequence $\{\pi_i\}$ of proper priors. But then the rules of probability theory again tell us unambiguously: in the above formulae, as in $(\sin x/x)$, we are to take the limit of the ratio. Which is, of course, just what I did in (24). The "difficulty" feared by DSZ would arise only if one tried, erroneously, to take instead the ratio of the limits.

It is astonishing that it should be necessary to point out such things here; and not only because the point was demonstrated and emphasized in my presentation. Indeed, unless we see clearly the distinction between the limit of a ratio and the ratio of the limits, we cannot even grasp the concept of a derivative (dy/dx).

Now, what happens if we consider a sequence $\{\pi_i\}$ of proper priors satisfying (R7), and the corresponding sequences of posterior distributions (R3) and (R6), such that the limit of $\{\pi_i\}$ is improper? The answer is, of course, that if we do our mathematics correctly, we shall find that B_2 and B_3, being in agreement for each member π_i of the sequence, will remain in agreement in the limit. And we should not be surprised to find, as (24) shows, that B_1, being in a superior position for each member π_i of the sequence, will still be in a superior position in the limit, because his extra prior information remains relevant.

Yet DSZ insist that B_1, in the limit, becomes guilty of gross inconsistencies, while B_2 and B_3 become guiltless! Then, at what point of the sequence $\{\pi_i\}$ does this reversal of status occur? As I showed by explicit mathematical demonstration (24), we can find a member of the sequence at which all priors are still proper, and B_1's different conclusions *are*, obviously, due entirely to his greater information; but at which all his conclusions are within one part in 10^{10} of their limiting values. Evidently, then, DSZ must believe that in that final one part in 10^{10} change in his estimate of ζ, there occurs a sudden *qualitative* change in his status as a consistent Bayesian.

Viewed in this way, it seems that DSZ can maintain their position only if they now deplore, not only the use of improper priors, but also proper priors that are in some sense close to them. Indeed, this is just what they seem to be doing in [9], where they note, quite correctly, that the inequality $t \ll yQ$ has a low prior probability. But of course given any prior and any data set, we can always find an inequality, satisfied by the data, which has a low prior probability; by that criterion any data set can be made to appear "exceptional". The point of my discussion was that *if* $t \ll yQ$, then in (24) it is surely a valid simplifying approximation – and not an impropriety – to set $t = 0$.

But we are still mystified: how can DSZ believe so strongly in the guilt of B_1 that explicit mathematical demonstrations to the contrary have no effect on that belief? What is the argument that, for them, carries such overwhelming weight? In the DSZ commentary we can locate only two sentences, at the end of [3], that address this point, and what they say is in part contradicted by later statements in [7], [8], and [10]. But if we look only at [3], we see the following argument: *if* the data y are irrelevant, then each of the two procedures (R3) and (R6) appear "unobjectionable", and there is an "inconsistency" if they do not yield the same result. This is, in essence, a statement of an intuitive *ad hoc* principle, similar to the Reduction Principle.

But against that intuitive judgment, the mathematical rules of probability theory tell us that: (1) the data y are irrelevant only for an improper prior (this was stressed also by Fraser in the discussion following the DSZ paper); (2) for an

improper prior the posterior distribution (R3) is not in general – and (R6) is never – *defined* except as the aforementioned limit of a sequence; and (3) the result of that mathematical limiting operation shows, rigorously and unambiguously, that B_1 and B_3 will *not* agree because B_1's prior information about η remains relevant. This is just what I stated in the opening sentences of my presentation; so, as in the *Ring des Nibelungen*, after all this drama we manage to end up right back at our starting point.

For DSZ, then, their intuitive judgment must carry more weight than the mathematical rules of probability theory. For the rest of us, the issue is: what shall we believe, the intuition or the theorem? On this note we can close our discussion of marginalization by quoting the words of DSZ on a different issue (DSZ, p. 231): "... if there is a clash, it *could* be that the intuition needs sharpening".

Venturing away from marginalization, DSZ [5] then point to a similar example of "inconsistent behavior". If the data D consist of n observations from $N(\mu, \sigma)$, the maximum likelihood estimate of σ^2 based on D is the statistic S; while the MLE based on S alone is $nS/(n-1)$. The reason for this is clear, since the MLE is the mode of a posterior distribution with uniform prior. Indeed, for any prior $\pi(\mu, \sigma|I)=\pi(\sigma)$ the modes of $p(\mu, \sigma|DI)$ and $p(\sigma|SI)=p(\sigma|DI)$ occur at different σ, as one expects when two quantities do not have independent posterior distributions. What they do *not* note is that the Bayes estimator σ^* of σ based on any loss function $L(\sigma, \sigma^*)$ is the same for $p(\mu, \sigma|DI)$ and $p(\sigma|SI)$.

As we see, this "inconsistent behavior" is indeed similar to the marginalization paradox; for both are self-inflicted, and both are cured at once, not by confronting one *ad hoc* principle (maximum likelihood) by another (Reduction), but simply by correct use of Bayesian methods – which is to say, by avoiding all gratuitous *ad hoc* principles and drawing only the inferences that follow by rigorous mathematics from the product rule $p(AB|C)=p(A|BC)\ p(B|C)$ and sum rule $p(A|B)+p(\sim A|B)=1$ of probability theory. In the last analysis, that is all that Bayesian calculations amount to.

In conclusion, the progress in basic statistical theory that I still believe will emerge from all this, could not take place as long as we allowed ourselves to be distracted by the red herring of improper priors. For a practical statistician, the notion of an improper prior is often a natural and useful idealization – just as the notion of a perfect triangle is for a surveyor. For both, it is part of their professional competence to understand clearly under *which* conditions the idealization is appropriate.