The roles of inductive modelling and coherence in Bayesian statistics

TOM LEONARD*
Queen's University, Kingston, Ontario
University of Warwick

SUMMARY

The role of the inductive modelling process (IMP) seems to be of practical importance in Bayesian statistics; it is recommended that the statistician should emphasise meaningful real-life considerations rather than more formal aspects such as the axioms of coherence. It is argued that whilst axiomatics provide some motivation for the Bayesian philosophy, the real strength of Bayesianism lies in its practical advantages and in its plausible representation of real-life processes. A number of standard procedures, e.g. validation of results, choosing between different models, predictive distributions, the linear model, sufficiency, tail area behaviour of sampling distributions, and hierarchical models are reconsidered in the light of the IMP philosophy, with a variety of conclusions. For example, whilst mathematical theory and Bayesian methodology are thought to prove invaluable techniques at many local points in a statician's IMP, a global theoretical solution might restrict the statistician's inductive thought processes. The linear statistical model is open to improvement in a number of medical and socio-economic situations; a simple Bayesian alternative related to logistic discrimination analysis often leads to better conclusions for the inductive modeller.

Keywords: COHERENCE; INDUCTIVE MODELLING; AXIOMS; BAYES FACTOR REGRESSION; DISCRIMINATION; SKEWED-NORMAL; MULTI-PARAMETER ESTIMATION

1. THE RÔLE OF BAYESIANISM IN THE REAL WORLD

An overwhelming majority of practical statistical problems fall into a particularly general category. The statistician S is frequently required to investigate a real-life process R_ℓ and to extract some meaningful conclusions from his investigation. He might for example be faced with a large-scale set of medical data, and a team of medical experts, and might wish to assist in the diagnosis of the main causes of a particular disease. Alternatively, he may be

* Now at University of Wisconsin-Madison

concerned with a production process, e.g. for synthetic fibres, and be required to either forecast future output or to help detect ways in which the process can be improved. As a third example, he may be working in an educational testing environment, with the task of identifying students who could usefully attend particular colleges.

The Bayesian philosophy provides an excellent conceptual background for S's investigation of R_ℓ. As each fresh piece of information about R_ℓ becomes available to S, he is able to use it to refine his overall appreciation of R_ℓ. Whilst he might try to do this in a completely intuitive way, Bayesianism will frequently assist him in crystallising his complex thought processes, and in keeping his ideas on a sensible track.

It is one of the main themes of this paper that, whilst mathematical theory and Bayesian methodology play valuable *local* rôles in helping to clarify S's thought processes at a variety of points in his investigation of R_ℓ, they should not be expected to lead to a meaningful *global* solution to the problem of how S should approach his overall investigation of R_ℓ.

Even if it were technically possible to construct a feasible 'global' theory, we feel that such a solution would be inevitably restricted by the boundaries of its own assumptions, and could serve to constrict the inductive reasoning which is so vital to our understanding of the real world, and which no deductive theory can properly represent. For example, it is frequently the appearance of something completely unexpected which leads to new discoveries and important innovations. If our theory were insufficiently innovative to incorporate the possibility of all unexpected occurences in advance, then it might merely serve to disguise the potential discovery in a manner contrary to the general principles of science.

Similarly, if S wishes to develop a mathematical model as a device for extracting real-life conclusions from the data, then theory on its own would need to assume an enormously superhuman capacity to always select an inductively sensible model from a set of alternatives specified in advance. By examining the data, getting a good feel for its properties and its background, and interacting between the data, the client, tentative models and analyses, and possible real-life conclusions, S will often be able to use his inductive thought processes to help him to extract rich and meaningful conclusions from the data, which might well have remained undiscovered if he had followed a more formal philosophy.

In this *inductive modelling* process (IMP) which should be viewed as the basis of statistical practice. Whilst mathematical theory and Bayesian methodology will provide invaluable assistance at many local points of IMP, a more global concentration on these aspects may well lead S to either work in a theoretical vacuum or to become restricted by theoretical formalisms.

2. FORMAL AND INFORMAL JUSTIFICATIONS OF BAYESIANISM

The statistician S will typically need to convince his client of the possible benefits of Bayesian procedures when compared with other e.g. frequentist procedures. How should he seek to do this? It seems to us that S should simply try to convince his client that (a) Bayesianism often leads to a much more reasonable conceptual representation of aspects of R_ℓ and that (b) when applied to local problems, Bayesian methodology frequently leads to superior practical results (e.g. (i) multi-parameter estimation, (ii) problems involving nuisance parameters).

A number of authors (e.g. De Groot, 1970, pp. 71-76; Savage, 1954 and de Finetti, 1975) have devised axiom systems which, if acceptable to S, lead to the conclusion that he must act like a Bayesian, e.g. by representing his information by a probability distribution. Whilst some Bayesians might view such axiom system simply as a helpful description of the Bayesian approach, others (e.g. the Lindley-Smith-Dickey-Hill school) view such 'axioms of coherence' as compelling reasons for acting like a Bayesian and might even be tempted to employ such extremely appealing verbal arguments as 'Well, if you don't act like a Bayesian then you must be incoherent!'.

Most such axiom system seem acceptable from a formal point of view and it would appear sensible to act like a Bayesian if R_ℓ were simple enough to permit this. However, whilst many arguments in favour of Bayesianism based upon axiomatics possess substantive appeal, and whilst it would be pleasant if the axiomatic justifications turned out to possess a firm scientific basis, they may provide as convincing a justification as we might have hoped for.

In discussing ways of justifying Bayesianism, it might be useful to consider a particular set of axioms in detail. The set described by DeGroot is probably one of the easiest to follow; it is not confused by any notions of betting and its assumptions are similar in strengh to those suggested by most previous authors. They appear to have been suggested by DeGroot himself more as a description of the Bayesian approach then as a justification of it; they are related to the work of Villegas (1964).

The axioms consider a space Ω (which could for example be viewed as the space of all possible states of R_ℓ) with a sigma-field $_\alpha$ of events, where any two elements A and B of $_\alpha$ can be compared using the notation $A < B$ to indicate that S considers B to be more likely than A, $A \sim B$ to indicate his opinion that A and B are equally and $A \leq B$ to indicate that either $A > B$ or $A \sim B$; For the final axiom we require the definition

Df.: A quantity X is a *uniformly distributed random variable on the interval* [0,1] if for any two sub-intervals I_1 *and* I_2 of [0,1], $[X \in I_1] \leq [X \in I_2]$ if, only if, $\lambda(I_1) \leq \lambda(I_2)$, where $\lambda(I)$ denotes the length of the interval I.

The five 'axioms of coherence' are

Axiom 1: For any A and B, either $A < B$, or $A > B$, or $A \sim B$.

Axiom 2: For any A_1, A_2, B_1, and B_2, such than $A_1 \cap A_2 = B_1 \cap B_2 = \phi$ and $A_i \leq B_i$ for $i = 1,2$, then $A_1 \cup A_2 \leq B_1 \cup B_2$. If in addition either $A_1 < B_1$ or $A_2 < B_2$ then $A_1 \cup A_2 < B_1 \cup B_2$.

Axiom 3: For any A, $\phi \leq A$. Furthermore $\phi < \Omega$.

Axiom 4: If $A_1 > A_2 >$ is a decreasing sequence of events and B is some fixed such that $A_i \geq B$ for $i = 1,2,\ldots.$ then $\bigcap_{i=1}^{\infty} A_i \geq B$.

Axiom 5: There exists a uniformly distributed random variable X on interval [0,1].

The first three of the above axioms would probably seem reasonable to statisticians of most philosophies. Attempts should therefore be made to satisfy them, at least approximately, in local situations where an overemphasis would not detract S from the main purpose of his IMP, e.g. to induce real-life conclusions from the data. They lead to an approach described by DeGroot as 'relative likelihood', but do not in themselves give the slightest hint of a probability distribution on Ω.

The fourth axiom may be viewed as a regularity condition which ensures that the probability distribution, induced by Axiom 5, is countably additive rather than finitely additive.

The fifth axiom and its implications are of paramount importance. It introduces the notion of an auxiliary experiment (e.g. the spin of a roulette wheel) which yields an (objectively) random number X in the interval [0,1]. The statistician S is expected to be able to compare events in Ω with events on [0,1]. DeGroots's theory then leads to the construction of a unique probability distribution over Ω which represents S's feelings about elements of Ω and hence provides us with the result that S is actually acting like a Bayesian.

Implicit in DeGroot's formulation is the assumption that the first four axioms relate to any (measurable) subsets of the union of Ω and [0,1] as well as of Ω itself. It seems obvious that it is this implicit axiom (*5a*) which is primarily responsible for inducing the probability distribution on Ω since it maps subsets of Ω into the interval [0,1] in a mathematically rigorous way. It also seems that axiom *5a* is virtually as strong as the final result and that we are therefore very nearly saying "if you want to act like a Bayesian then you must act like a Bayesian"!.

Consequently, whilst axiom *5a* and the final result both possess considerable inductive appeal for Bayesians, the axioms do not in themselves appear to add anything beyond a useful interpretation of Bayesian thinking, in terms of an auxiliary experiment. The axioms should certainly never be

used as a justification for Bayesianism or as a device for convincing non-experts. It would be more reasonable to refer to the justifications discussed in (*a*) and (*b*) above.

When S is engaged in his IMP, he may find it useful to employ the ideas of coherence as a conceptual background, to help him think upon Bayesian lines. If however he sticks too closely to axiomatics then he may lose sight of the primary objective of his investigation e.g. to extract real-life conclusions from the data. He should not permit coherence to restrict his creative and innovative ideas and he should concentrate more closely on appreciating the practical situation at hand. A good inductive appreciation of R_ℓ with a background culture of Bayesian coherence is to be preferred to an over rigid approximation to coherence and a lack of appreciation of R_ℓ.

The philosophy of coherence may be viewed in similar spirit to the ideas of Birnbaum (1962), which probably comprised one of the best single contributions to theoretical statistics. Birnbaum proved that the sufficiency principle and the conditionality principle together imply the likelihood principle, a far-reaching result which enables the purist to disregard many frequentist procedures integrating across the sample space.

The conditionality principle possess similar appeal to Axiom 5*a* described above, and whilst acceptable in an idealistic sense, it is primarily responsible for Birnbaum's result that statisticians should follow the likelihood principle. When S is engaged in his IMP he may find it too restrictive to stick rigidly to the conditionality principle. For example, a responsible S would, as a general norm, obtain a good feel for his data before inducing a family of sampling distributions for his observations.

A related practical difficulty associated with Birnbaum's approach is that it is a conditional philosophy, given the truth of an underlying model for the observations. Any debate which conditions on the truth of an underlying model may be well wide of the target in the light of the philosophy "All sampling models are ultimately wrong and should simply be introduced as subjective, mathematical devices, in order to induce real-life conclusions from the data". This philosophy is an essential ingredient of our whole concept of *IMP*; it seems to provide us with one of the few sensible ways of engaging in a modelling process, and immediately detracts attention from philosophies which depend upon the truth of an underlying model.

3. JUSTIFIFYNG REAL-LIFE CONCLUSIONS

Once S had induced a real-life conclusion from the data and his appreciation of R_ℓ, he might wish to compile evidence in support of his conclusion, so that he can convince his client and other experts that it is both viable and meaningful. For example, in a paper to be published elsewhere,

(but discussed in the verbal presentation of this material), Leonard, Low and Broekhoven (1978) describe a conclusion which is not in immediate concurrence with existing medical opinion. They have found that, whilst a high risk of fetal asphyxia in babies does not in fact appear to be noticeably associated with prematurity it does appear to be strongly associated with babies who possess a much lower birthweight than might be expected, for a given degree of prematurity.

These are several possibilities open to S, for example.

(a) To test his underlying model against the data, using a conventional significance test.

(b) To informally evaluate his model and conclusions by checking them out against future observations.

(c) To informally check out his real-life conclusions against the present data set, look for patterns in the conclusions, and consider their status in connection with existing scientific knowledge on related topics.

(d) To discuss his conclusions in detail with his client, to see if they fit in sensibly with his existing views, or whether the latter can be sensibly modified to accomodate his conclusions.

(e) To refer to the level of expertise of his own inductive judgement.

I feel that (*a*) should not be regarded as completely adequate, though significance tests may be useful as intuitive devices. Firstly, situations could be envisaged where the model is inadequate, but the specific conclusions are still viable. For example, a very tentative model could be used to stimulate plausible creative ideas by S, or the real-life conclusions might only depend upon particular aspects of the model. More importantly, significance tests do not appear to possess too much formal justification. For example, Leonard (1979) shows that for large sample sizes, significance levels may be sensibly replaced by value depending on the sample size. For further discussions of significance testing see Leonard and Ord (1976), and Leonard (1977 and 1978).

The alternative (*b*) appears to provide a useful check. However, the number of future observations will typically be finite and probably never particularly large. Also, by the time they have been collected R_ℓ will probably have evolved into an updated situation, and the usefulness of any underlying model undetermined. Just as the practical viability of the theoretical concept of consistency may be critically exposed in the context of the philosophy "the greater the amount of information the greater the chance of contradiction (of

the original model)", the usefulness of predictive validation seems affected by the possible deviation of future observations from the situation currently at hand, whenever there are enough future observations to provide a case for a through validation.

Whilst (*c*) and (*d*) also provide useful checks, we feel that in the last analysis S can only refer to (*e*) and recognise that both R_ℓ and his investigation of it are basically subjective. He can only really attempt to justify his conclusions by simply indicating that he has carried out a subjective and honest investigation of R_ℓ and that his conclusions appear to be sensible.

We have thus arrive at the straightforward proposition that statistical practice is a subjective process which is highly dependent upon the expertise, honesty, and experience of the statistician, just as the practice of, say, medicine, law, psychology, economics, and indeed most branches of science, is also subjective and highly dependent upon similar qualities of experts in those areas.

In particular, the statistician will only be able to adequately complete his *IMP* if he possesses the mathematical skills and level of creativity which will carry him through the numerous local and innovative procedures which *IMP*'s typically require. People working from a "cookbook" of recipes will typically find difficulty with *IMP*'s and should therefore be discouraged from playing a leading rôle in large-scale investigations. The ultimate success of Bayesian statistics will depend upon whether we can bridge the gap between theory and practice and link theoretical innovation with practical relevance.

4. CHOOSING BETWEEN DIFFERENT SAMPLING MODELS

During his *IMP*, S may wish to use a formal Bayesian procedure to help him to measure his opinions about a finite number of sampling models. A number of authors (e.g. Dickey 1975, and Harrison and Stevens, 1976) have proposed a general approach to this problem, based upon sharp hypotheses and mixed models. However, whilst Schwarz (1978) has developed an approximate method for large sample sizes, which does not depend upon the choice of prior distribution, the general approach experiences some technical difficulties for smaller sample sizes. When more than two or three models are involved in the mixture it also appears to us to place too much emphasis on the search for a 'true' sampling model, and to be somewhat overcomplex and insufficiently motivated towards the extraction of meaningful real-life conclusions from the data. An informal consideration of alternative models in the light of real-life aspects may be more appropriate, i.e. we view the Bayesian mixed model approach as often assuming too much of a 'global' nature to provide an inductively useful service for S.

Suppose that S wishes to choose between a binomial sampling model with

probability θ and sample size n for a frequency x and an alternative sampling model with probability mass function $p_0(x)$. For simplicity, we suppose that $p_0(x)$ is completely specified; assume also that whenever the binomial sampling model holds, θ possesses the beta prior distributions.

$$\pi(\theta \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (0 \leq \theta \leq 1; 0 < \alpha,\beta < \infty) \tag{1}$$

Following the general approach referenced above, the posterior probability that model p_0 holds, given that either p_0 or the binomial sampling model holds, is then denoted by

$$\phi_0 = \frac{\phi R_x}{\phi R_x + 1} \tag{2}$$

where ϕ is the corresponding prior probability, and R_x is the 'Bayes factor' which satisfies

$$R_x = p_0(x)/D(\alpha,\beta) \tag{3}$$

where

$$D(\alpha,\beta) = \frac{\Gamma(n+1)\,\Gamma(\alpha+\beta)\,\Gamma(\alpha+x)\,\Gamma(\beta+n-x)}{\Gamma(x+1)\,\Gamma(n-x+1)\,\Gamma(\alpha+\beta+n)\Gamma(\alpha)\Gamma(\beta)} \tag{4}$$

Whilst (2) provides a formal and coherent Bayesian solution to this problem, it is so sensitive to the choice of prior distribution for θ that it would be viewed as impractical in many situations. Suppose, for example, that α is moderately large and is increased by a single hypothetical prior observation to $\alpha + 1$. Note from (4) that

$$\frac{D(a+1,\beta)}{D(a,\beta)} = (\varrho p + (1-\varrho)\,\xi)/\xi \tag{5}$$

where $p = x/n$, $\xi = \alpha/(\alpha + \beta)$, and $\varrho = n/(\alpha+\beta+n)$.

Therefore under our minor adjustment to the prior the Bayes factor in (3) should be divides by the quantity in (5), which will always lie between p/ξ and unity. For example, with the proportion p equal to 9/10 and the prior mean ξ equal to 1/10, the divisor could be as high as 9, radically, affecting the posterior probability in (2).

Paradoxically the sensitivity is at its greatest at $n \to \infty$, with p, α, and β

fixed, so that $p \to 1$. In this case, the Bayes factor in (3) will tend to either zero or infinity irrespective of the prior, but the rate of convergence will become particularly sensitive, as increasing α by unity is equivalent to dividing the Bayes factor by the maximum possible value of p/ξ.

The sensitivity described above is not unique to the present special case. For example, Lindley (personal communication) has informed us that there are a further sensitivity problems when investigating whether or not to take observations to be normally distributed. Other problems concerning this type of approach are discussed by Atkinson (1978).

We are drawn to the viewpoint that it may be inductively more sensible to choose a sampling model by considering various aspects of R_ℓ, and the data, and by generally following the philosophy outlined in the last paragraph of section 1 rather than by referring to a coherent Bayesian procedure with possible misleading conclusions. Note that sensitivity problems occur very generally in a number of other areas of Bayesian estimation and inference; some of these will be discussed in forthcoming publications by J.Q. Smith and J. Kadane.

5. THE RÔLE OF BAYESIAN PREDICTIVE DISTRIBUTIONS

A number of authors, e.g. Aitchison and Dunsmore (1975) view predictive distribution as playing a leading role in Bayesian methodology. It is our own view that whilst many standard predictive distributions, e.g. based upon conjugate prior distributions, play a role in idealised situations where the sampling model and prior distribution can be precisely specified, they may be of more limited importance when S is engaged in the practical details of his *IMP*. This conclusion is primarily based on the following reasons:

(a) Many predictive distributions can be as sensitive to the choice of prior as the Bayes factors discussed in section 4. For example, if (1) provides the posterior distribution for a probability θ, then the quantity $D(\alpha,\beta)$ in (4) is just the predictive probability that a binomial frequency, with probability θ and sample size n, is equal to x. Therefore if α is increased to $\alpha + 1$, this predictive probability will multiplied by a factor of up to p/ξ where p and ξ now respectively denote the predicted proportion x/n and the posterior mean $\alpha/(\alpha+\beta)$.

(b) The statistician S will typically remain uncertain about the correctness of his sampling model, and many conventional predictive distributions fail to take account in this uncertainty.

Suppose, for example, that we analyse a set of data which appear to be roughly normally distributed, that the practical situation (e.g. quality control)

requires us to predict the probability that a further observation will be negative, and that the proportion of negative observations is 0.27. We then derive a standard predictive t-distribution under normal and conjugate assumptions and find that our predictive probability, conditional on our choices of sampling and prior models is 0.15. The latter is however a highly conditional probability and it might therefore be highly misleading to quote it as a useful result. Whilst our intention might suggest that a better (subjective) predictive probability lies between 0.15 and 0.27, many formal procedures for judging it more precisely would also be highly dependent upon any assumptions made.

Our general philosophy that "all sampling models are ultimately wrong" (see the last paragraph of section 3) leads us naturally to the philosophy that "all predictive distributions based upon particular sampling models are ultimately wrong". Conclusions based upon them sould be treated with caution.

We view many conventional predictive distributions as a bit on the over-formalistic side; indeed many standard predictive distributions do not obviously lead to any further inductive understanding of $R\ell$ beyond that already provided by the sampling distributions from which they are generated. Many probabilities calculated from predictive distributions can only be considered to lead to reasonable practical predictive probabilities if these fit in closely with raw probabilities calculated from the data, or if there is some further inductive reason for using them. However, an alternative type of predictive distribution yielding greater scope to the inductive modeller will be discussed in section 7.

6. SOME PRACTICAL ADVICE ON THE LINEAR MODEL

We now discuss some practical aspects of the linear model, and consider dependent variables y_i satisfying

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \qquad (i = 1,\ldots,m) \tag{6}$$

but where, for $q \leq p < m$, $x_{i1},\ldots,x_{iq}$ are statistical observations rather than fixed constants, and where $x_{iq+1},\ldots,x_{ip}$ are functions of $x_{i1},\ldots,x_{iq}$. The y_i could denote the salaries of m individuals, and $x_{i1},\ldots,x_{iq}$ could measure socio-economic factors relating to these individuals. Alternatively, y_i could represent blood pressure, with $x_{i1},\ldots,x_{ip}$ measuring q different medical symptoms.

It is my practical experience, and the general experience of colleagues in a

consulting capacity, that there are a large number of practical situations where the underlying assumptions of the linear model seem appropriate, but where a modelling procedure of this nature turns out to be rather inadequate. This is particularly true of many socio-economic and medical data sets whenever there is a large amount of random fluctuation between the vectors $\mathbf{x}_i = (x_{i1},\ldots,x_{iq})$. In such circumstances it is often virtually impossible to arrive at any sensible model of the form defined in (6), whatever functional forms are chosen for $x_{iq+1},\ldots,x_{ip}$, and whatever estimation procedures (e.g. least squares, weighted least squares, or Bayes) is employed.

The data sets referred to might be viewed as possessing insufficient information to present the possibility of useful conclusions. Alternatively, a novice might feel tempted to add more and more explanatory variables in attemption to obtain a meaningful model. However, the simple Bayesian procedure described in section 7 and relating to logistic discrimination analysis very frequently leads to useful conclusions which would often be missed by the linear statistical modeller.

For a number of data set of this type, we have experienced a residual sum of squares which remains steadfastly close to the total sum of squared for virtually any model specification of the type defined in (6). This is because the $\mathbf{x}_i$ vectors are subject to so much random variation that it is almost impossible to use any set of fitted values to provide reasonable numerical predictions of the dependent variables i.e. the information content of the data is not of a predictive nature. Whilst inductive conclusions might still be available via the linear model, they will frequently be of limited scope owing to the extreme inadequacy of the model. For example, difficulties (b) involving predictive distributions, as discussed in section 5, will be highlighted in this context.

Whilst linear models present difficulties when the information content of the data is not a *predictive* nature, the same data sets often contain some very worthwhile information of a *probabilistic* nature which can be extracted via the methodology of section 7. The latter will also be enable S to model terms corresponding to $x_{iq+1},\ldots,x_{ip}$ in a direct (rather than, say, stepwise) fashion; for example it will help him to induce the presence of any complicated interaction effects without needing to engage in a long search.

Consider for simplicity the special case where $q = 1$ and $x_{i1} = x_i$. Suppose that the points (y_i, x_i) are plotted on a scatterdiagram for $i = 1,\ldots,m$. Whilst these points will seldom lie close to any particular curve for the type of data set under consideration, the frequencies of y's falling in any particular intervals will often change in a meaningful way as x increases, as long as this interval is chosen to be wide enough. Therefore, whilst fitted values under any linear model might give poor numerical predictions of the y_i, it might be possible to use the data to help predict probabilities for intervals in which,

say, a further observation y_{m+1} might lie, so that the data possess a quality of a probabilistic rather than a predictive nature. In other words, knowledge of a further explanatory variable x_{m+1} might affect S's probabilities about y_{m+1} but not provide him with enough information to be able to numerically predict y_{m+1} to any degree of accuracy.

It is my experience that data sets possessing information of this probabilistic rather than predictive nature a occur frequently in socio-economic and medical contexts, and that the linear model frequently possesses very limited scope for the analysis of such data sets. For example, many applications of the linear model to economics, sociology and medicine, might benefit from further consideration.

7. A BAYESIAN *IMP*

In the situation discussed in the previous section, where the information content is of a probabilistic rather than a predictive nature considerable headway may often be made upon categorising the dependent variable y. This will clearly lead to some loss of sampling information, but the loss need not be at all substantial, (owing to the highly random nature of the explanatory variables), as long as the dependent variable is categorised in a sensible way. For example, in the medical context of Leonard, Low and Broekhoven, three categories, referred to as 'low', 'medium' and 'high', with the boundary points based upon further medical considerations, were adequate to permit the extraction of some meaningful conclusions from the data.

If the dependent variable is split into s categories, then the vectors $x_1,\ldots,x_m$ are effectively sectioned into s subpopulations $\Lambda_1,\ldots,\Lambda_s$, where the elements of Λ_j are those $\mathbf{x}$'s for which the corresponding y lies in category j. We let $n_1,\ldots,n_s$ denote the numbers of $\mathbf{x}$'s falling in the respective subpopulations $\Lambda_1,\ldots,\Lambda_s$.

Since the $\mathbf{x}$'s are themselves vectors of statistical observations, the $\mathbf{x}$'s in each sub-population Λ_j may be viewed as comprising a random sample from a distribution, say with density $f_j(\mathbf{x})$. The form of this density may be inductively modelled by S in the light of the corresponding $\mathbf{x}$'s and his appreciation of $R\ell$. This provides a vital part of S's *IMP* in this context; he needs to model the s densities $f_1,\ldots,f_s$. Suppose now that S wishes to be able to predict probabilities for a further dependent variable y_{m+1}, given a further vector of explanatory variables $\mathbf{x}_{m+1}$. Then the probability that y_{m+1} falls into the j^{th} category, given that $\mathbf{x}_{m+1} = \mathbf{x}$ is given by

$$\text{prob}\,(\Lambda_j|\mathbf{x}) = \frac{\pi_j f_j(\mathbf{x})}{\sum_{k=1}^{s} \pi_k f_k(x)} \qquad (j=1,\ldots,s) \tag{7}$$

where π_j denotes the corresponding prior probability, However, in the

absence of knowledge of $\mathbf{x}$, S will frequently be prepared to set. $\pi_j = n_j/m$ for $j = 1,\ldots,s$, in which case we have

$$\text{prob}\,(\Lambda_j|\mathbf{x}) = \frac{n_j f_j(\mathbf{x})}{\Sigma_{k=1}^{s} n_k f_k(x)} \quad (j = 1,\ldots,s) \tag{8}$$

The formula in (8) may be applied in a simple way to data sets whose information content is of a probabilistic nature; it seems to fit in neatly with the concept of *IMP*. It provides a standard procedure for many regression problems which could be used as an alternative to analyses based upon the linear model.

Note that the expression on the right hand side of (8) plays the rôle of a regression function. We for example have

$$\log\,[\text{prob}\,(\Lambda_j|x)/\,\text{prob}\,(\Lambda_k|\mathbf{x})] = \log\,(n_j/n_k) + \log f_j(\mathbf{x})/f_k/(\mathbf{x}) \tag{9}$$

This result is employed in logistic discriminent analysis. For example, Anderson (1974) mentions that multivariate normal assumptions for the f_j lead to a quadratic discriminant of regression function on the right hand side of (9).

Under our general *IMP*, S is expected to simply induce $f_1,\ldots,f_s$ from the $\mathbf{x}$'s and R_ℓ. Our point is that no further modelling will then be required because appropriate substitutions in (8) will complete the specifications of the predictive probabilities. During this process, S will need to interact between scatterdiagrams of the $\mathbf{x}$'s in the different sub-populations and his other experience and he will therefore be able to take full account of the probabilistic-type information content of the data. This inductive modelling will enable him to obtain predictive probabilities via (8), By considering graphical plots of the latter against different explanatory variables he is then in a position to extract real-life conclusions from the data.

Note that the above *IMP* automatically models the form of the regression function and hence the presence of any interaction effects, even if these are of a complex nature. As a simple example, multivariate normal assumptions for the f_j lead to cross-product terms on the right hand side of (9), which may be viewed as the interaction terms in a logistic regression. They now become completely determined upon identification of the f_j, providing a much more straightforward modelling procedure, then, say, standard stepwise procedures for the linear model. For non-normal f_j the interactions can assume a much more complex nature, but S has a very straightforward way of inducing them.

We recommend replacing any unknown parameters in the f_j by suitable points estimates (e.g. maximum likelihood or Bayesian). This should be frequently superior to the coherent Bayesian procedure of integrating each f_i in (8) with respect to the corresponding prior distributions of the parameters, since the latter will suffer similar sensitivity problems to those discussed in sections 4 and 5.

There are a number of ways of checking the probabilities in (8) against the data set. For example, boundaries on $\mathbf{x}$ could be determined for each j such that prob $(\Lambda_j | \mathbf{x})$ is greater than a specified value. Then the proportions of actual $\mathbf{x}$'s falling inside these boundaries could be enumerated, and they will all ideally be greater than the specified lower bound for the predictive probability. Added credibility will also be given to the *IMP* if the curves of prob $(\Lambda_j | x)$ against $\mathbf{x}$ evolve in a sensible way for increasing j.

The above approach has been found to yield practical conclusions in a variety of different situations, than would appear possible under a standard linear model approach. Similar methodology was employed by Leonard, Low and Broekhoven in their medical context.

8. THE SKEWED-NORMAL DISTRIBUTION

The statistical modeller is frequently faced with data with both a positive and negative tail, and which indicate a definite skewness. There are surprisingly few probability distributions in the literature for adequately modelling skew data when the latter are scattered on the whole real line. The following properties would however seem to be desirable for a family of two-tailed distributions which provide skew alternatives to say, the normal or t-distribution:

(i) A meaningful set of at least three parameters, with convenient functions of the parameters representing location, spread and skewness.

(ii) A useful symmetric distribution as a special case.

(iii) The property that whilst the two tails can be different they should be 'similar in nature', in the sense that different functional forms assumed for the tails might suggest a difference which was not exhibited by the data.

(iv) The form of the likelihood function, given n observations, should not permit the observations in one tail to unduly influence the estimated thickness of the other tail.

(v) Straightforward *ad hoc* and Bayesian estimation procedures for the parameters.

(vi) Easily tabulated interval probailities.

(vii) Reasonable regularity conditions for the density e.g. a continuous first derivative at all points.

All the above properties are satisfied by the *skewed-normal distribution*, with parameters μ, σ_1^2, σ_2^2, and density

$$p(x|\mu,\sigma_1^2,\sigma_2^2) = \begin{cases} \sqrt{(2/\pi)}(\sigma_1+\sigma_2)^{-1}\exp[-\tfrac{1}{2}\sigma_1^{-2}(x-\mu)^2] \text{ for } x \leq \mu \\ \sqrt{(2/\pi)}(\sigma_1+\sigma_2)^{-1}\exp[-\tfrac{1}{2}\sigma_2^{-2}(x-\mu)^2] \text{ for } x \geq \mu \end{cases} \tag{10}$$

This distribution possess mode μ and probabilities $\sigma_1/(\sigma_1 + \sigma_2)$ and $\sigma_2/(\sigma_1 + \sigma_2)$ either side of the mode. Its technical properties, including a Bayesian analysis, will be reported in more detail elsewhere.

9. SUFFICIENCY, OUTLIERS AND COHERENCE

In many statistical problems, the existence of a sufficient statistic of small dimensions implies in effect that the sampling distributions is a member of the exponential family. Therefore any discussion of the inductive reasonability of the concept of sufficiency must be closely related to a debate on the adequacy of the exponential family of distributions.

The general concept of sufficiency could be criticised on the grounds that a sufficient statistic typically reduces the number of pieces of information we can extract from the data, i.e. from the sample size to the dimension of the sufficient statistic. The data are therefore reduced to a form where they can, say, only describe one or two aspects of the sampling distribution, e.g. location and spread, but may tell us nothing about, or even disguise, other important aspects of the sampling distributions, e.g. possible bimodality or thicker tails than might be experienced with the exponential family.

Consequently, in situations where we might wish a formal analysis to tell us as much as possible about the sampling distribution, the concepts of sufficiency and the exponential family of distributions do not seem to be completely adequate. The formal Bayesian could, for example, be tempted to refer to the interesting approach of O'Hagan (1979) and employ outlier-prone and outlier-resistant sampling distributions in an attempt to cope with outliers.

On the other hand, sampling distributions yielding sufficient statistics typically possess meaningful characteristics and meaningful parameters. They seem to fit in well with the concept of *IMP* since *S* should always examine the data carefully and get a good feel for its properties before inducing a sampling distribution. He could for example investigate bimodality and outliers

intuitively rather than referring to the formalisms of a more complicated sampling model.

The statistician would probably do best to compromise between these two extremes. He could start off by referring to meaningful sampling distributions, with simple sufficient statistics, and to practical judgements of the data, with the objective of concentrating on the extraction of real-life conclusions from the data. However, he will sometimes find that his induction is unable to provide him with a clear enough picture. In this case slightly more complicated sampling distributions and an analysis taking formal account of further aspects of the data would sometimes be very useful.

As an example of the above approach, the skewed normal distribution in (10) is frequently applicable to (clearly unimodal) data with two tails. It can be employed as a useful device for locating the mode of the underlying distribution and for investigating its skewness. Its parameters are meaningful in this context; it provides a simple modification of a member of the exponential family. For example, when μ is known, statistics of the form

$$\sum_{i:x_i<\mu} (\mathbf{x}_i-\mu)^2 \text{ and } \sum_{i:x_i>\mu} (x_i-\mu)^2$$

are jointly sufficient for σ_1^2 and σ_2^2.

The skewed-normal distribution would clearly be inferior in a formal sense to a distribution with 't-type' tails if there we enough outliers in the data to suggest that its tails might be too thin. However, an adherent of *IMP* could still start off with the skewed-normal distribution and interact betwen tentative analysis based upon it, and the data, to see if the outliers affected the important real-life conclusions which could be induced from the data. For example, *S* could firstly try an analysis withouth the outliers, and then compare it with a further analysis with outliers present. Only if he convinces himself inductively that the outliers actually make a real difference should he consider a more formal (local) analysis based upon a complicated distribution with thicker tails. He is in this way able to increase his chances of extracting conclusions which might otherwise become confused by over complications.

The procedure outlined above is not obviously formally coherent, but we seem to have described a good example of a situation where a strict demand for formal coherence would appear to be inductively inappropriate.

10. MULTI-PARAMETER PROBLEMS AND PRIOR STRUCTURES

Consider next a general formulation where *S*'s n x 1 observation vector $\mathbf{x}$ is thought to possess a sampling distribution $f(\mathbf{x}|\theta)$ depending upon a q x 1

vector $\theta = (\theta_1,\ldots,\theta_q)^T$ of unknown parameters. In such multi-parameter situations, S might be concerned about Stein-type effects and lack of smoothness of the maximum likelihood estimates, and might therefore wish to employ shrinkage estimates for the θ_j. (See, for example, a method proposed by Leonard, 1973, for smoothing the probabilities in a histogram).

Following a general procedure discussed by Leonard (1972), S might seek a q x 1 vector $\alpha = (\alpha_1,\ldots,\alpha_q)^T$ of transformed parameters such that he is prepared to take the prior distribution of α to be multivariate normal, say with mean vector μ and covariance matrix $\mathbf{C}$. When μ and $\mathbf{C}$ are known a Bayesian shrinkage estimate for α is given by the posterior mode vector α, wich satisfies the equation

$$\left.\frac{\delta \log f(\mathbf{x}\,|\,\alpha)}{\delta\,\alpha}\right|_{\alpha\,=\,\tilde{\alpha}} = \mathbf{C}^{-1}(\tilde{\alpha} - \mu) \tag{11}$$

For example, when all the elements of μ are equal to a scalar μ, and $\mathbf{C}$ is a scalar multiple of the identity matrix, the elements of α will be a priori *exchangeable* and (11) will roughly speaking provide Stein-type shrinkages of their maximum likelihood estimates towards a common value μ.

However, S is typically faced with the problems of choosing suitable special forms for μ and $\mathbf{C}$ and evaluating any hyperparameters appearing in these special forms (these forms may be referred to as *prior structures*). The situation will often be far too complex for S untangle if he confines himself to strictly coherent Bayesian procedures. We recommend that he should instead assess his prior structures by interacting between his prior feelings, possible special forms for μ and $\mathbf{C}$, tentative estimates obtained from (11), any real-life conclusions he can induce from these estimates, his overall experience of R_ℓ, and cooperation with his client.

S will find it difficult to assign specific values to any hyperparameters appearing in his prior structures. A typical prior structure may be expressed in the form $\mu = \mu(\lambda_1)$ and $\mathbf{C} = \mathbf{C}(\lambda_2)$, once S has induced the dependence of the mean vector and covariance matrix on hyperparameters λ_1 and λ_2. For any such prior structure under consideration S should estimate λ_1 and λ_2 from the data and any prior information which might be available. We are however rather uncertain about the existence of convenient prior information for hyperparameters in complex models like this, except in special cases or when the prior information is itself data based. It is generally much more straightforward to avoid complicated and possible confusing distributions at the second stage of the prior model, and to simply estimate λ_1 and λ_2 from the data by maximising their 'marginal likelihood'.

$$\ell(\lambda_1, \lambda_2, |\mathbf{x}) = E[f(\mathbf{x}|\alpha)] \tag{12}$$

where the expectation on the right hand side is with respect to α, given μ and $\mathbf{C}$.

In summary, S may induce the functional forms of $\mu(\lambda_1)$ and $\mathbf{C}(\lambda_2)$ by following the general philosophy of *IMP*, and may then estimate λ_1 and λ_2 via a data-based procedure. Obviously, particular practical considerations might lead to refinements of this scheme.

Whilst it would be difficult to demonstrate formal coherence of the above procedure, it seems likely to often prove useful in a real-life sense when compared with more complex coherent procedures.

11. NON-PARAMETRIC DENSITY ESTIMATION

The approach described by Leonard (1978) to the non-parametric estimation of a density fits with the philosophy of *IMP* since it enables S to allow for real-life considerations as part of theoretical local analysis. For example, a hypothesised density can be introduced as a prior estimate, then the theoretical method can be used to provide a posterior estimate which can be considered inductively by S, to see where it differs from his null hypothesis, and to consider whether these differences are due to real-life aspects. He could also try out different hypothesised densities as part of his *IMP*, and generally interact between his prior specification, his posterior results, and possibly meaningful conclusions. The approach seems to be more useful than many previous frequentist procedures based on kernel functions, since these tend to place a bit more emphasis on data-fitting, rather than on the diagnosis of meaningful conclusions.

Note that Leonard uses a prior and posterior likelihood approach rather than a strictly Bayesian approach since this avoids certain technical problems over function spaces. We in general see nothing wrong in following an alternative philosophy if it is based upon similar prior information and leads to similar conclusions.

12. DISCUSSION

The concept of coherence has played an invaluable theoretical rôle over the years by highlighting the inadequacies of many frequentist procedures. However, the Bayesian philosophy is now firmly established and accepted as one of the few viable theoretical approaches to Statistics. It should therefore now look beyond debates with other philosophies, and theoretical discussions on the foundations, and emphasise its practical viability in non-trivial contexts, e.g. large scale data sets where the client provides background information from his own discipline. When broader considerations are taken

into account the rôle of coherence no longer seems paramount, and much more emphasis should be place on the *IMP* aspects of statistics. Whilst existing coherent methodology is useful at a variety of local points of *IMP*, the theoretical structure should be kept to a level of intellectual complexity where it assists the statistician to induce real-life conclusions from the data.

ACKNOWLEDGEMENTS

The author wishes to thank Professor J.M. Dickey and L. Broekhoven for helpful advice and discussions.

REFERENCES

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

ANDERSON, J.A. (1975). Quadratic logistic discrimination. *Biometrika,* **65**, 39-48.

ATKINSON, A. (1978). Posterior probabilities for choosing a regression model, *Biometrika* **65**, 39-48.

BIRNBAUM, A. (1962). On the Foundations of Statistical Inference. *J. Amer. Statist. Assoc.*, 269-326.

DE FINETTI, B. (1975). *Theory of Probability* 1 New York: Wiley.

DEGROOT, M.H. (1970). *Optimal Statistical Decisions.* New York: McGraw Hill.

DICKEY, J.M. (1975). Bayesian alternatives to the *F*-test and least squares estimation in the normal linear model. In *Studies in Bayesian Econometrics and Statistics* (Fienberg and A. Zellner, Eds.), 515-554. Amsterdam: North Holland.

HARRISON, P.J. and STEVENS, C.F. Bayesian Forecasting (with Discussion). *J. Roy. Statist. Soc. B* **38**, 205-247.

LEONARD, T. (1972). Bayesian Methods for Binomial Data, *Biometrika,* **59**, 581-589.

— (1973). A Bayesian Method for histograms. *Biometrika,* **60**, 297-308.

— (1977). A Bayesian approach to some multinomial estimation and pretesting problems, *J. Amer. Statist. Assoc.* **72**, 867-874.

— (1978). Density estimation, stochastic processes and prior information (with Discussion) *J. Roy. Statist. Soc. B.* **40**, 113-146.

— (1979) Why do we need significance levels? *M.R.C. Tech. Report.* University of Wisconsin-Madison.

LEONARD, T., LOW, J.A., and BROEKHOVEN, L. (1978). Assessing the risk of fetal asphyxia. *STATLAB Tech. Report.* Kingston, Ontario: Queen's University.

LEONARD, T. and ORD, J.K. (1976). An investigation of the *F*-test as an estimation shortcut. *J. Roy. Statist. Soc. B* **38**, 95-98.

O'HAGAN, A. (1979). On outlier rejection phenomena in Bayes inference. *J. Roy. Statist. Soc. B* **41**. 358-367.

SAVAGE, L. (1954). *The foundations of Statistics.* New York: Wiley.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* **6**, 461-4.

VILLEGAS, C. (1964). On qualitative probability Γ-algebras. *Ann. Math. Statist.* **35**, 1787-1796.