The Statistician (1999)
48, *Part* 1, *pp.* 1–40

Some statistical heresies

J. K. Lindsey

Limburgs Universitair Centrum, Diepenbeek, Belgium

[*Read before* The Royal Statistical Society on *Wednesday, July 15th, 1998*, *the President*, Professor R. N. Curnow, *in the Chair*]

Summary. Shortcomings of modern views of statistical inference have had negative effects on the image of statistics, whether through students, clients or the press. Here, I question the underlying foundations of modern inference, including the existence of 'true' models, the need for probability, whether frequentist or Bayesian, to make inference statements, the assumed continuity of observed data, the ideal of large samples and the need for procedures to be insensitive to assumptions. In the context of exploratory inferences, I consider how much can be done by using minimal assumptions related to interpreting a likelihood function. Questions addressed include the appropriate probabilistic basis of models, ways of calibrating likelihoods involving differing numbers of parameters, the roles of model selection and model checking, the precision of parameter estimates, the use of prior empirical information and the relationship of these to sample size. I compare this direct likelihood approach with classical Bayesian and frequentist methods in analysing the evolution of cases of acquired immune deficiency syndrome in the presence of reporting delays.

Keywords: Acquired immune deficiency syndrome; Akaike's information criterion; Asymptotics; Compatibility; Consistency; Discrete data; Hypothesis test; Likelihood; Likelihood principle; Model selection; Nonparametric models; Normal distribution; Poisson distribution; Robustness; Sample size; Standard error

1. Introduction

Statisticians are greatly concerned about the low public esteem for statistics. The discipline is often viewed as difficult and unnecessary, or at best as a necessary evil. Statisticians tend to blame the public for its ignorance of statistical procedures, but I believe that almost all of the fault lies with statisticians themselves and the way that they use, and teach, statistics. The two groups with whom statisticians have the most direct contact are students and clients. Unfortunately, the message that these receive is full of conflicts and contradictions. Much of this arises from the inordinate emphasis that statisticians place on a certain narrow view of the problem of inference itself, beginning with the teaching of the first introductory statistics course. Knowing a vast array of tests, or, more recently, estimation procedures, is considered to be much more important than being aware of what models are available, and studying which will be suitable for each particular question at hand and for the accompanying data analysis.

In what follows, I consider some of the shortcomings of modern views of statistical inference, Bayesian and frequentist. This is necessary because these views are increasingly being ignored in applied statistical practice, as the gap between it and theoretical statistics widens, and are having negative effects on students and clients. An adequate means of presentation is difficult to find because of the vast and internally contradictory nature of both approaches: to any critical

Address for correspondence: J. K. Lindsey, Department of Biostatistics, Limburgs Universitair Centrum, 3590 Diepenbeek, Belgium.
E-mail: jlindsey@luc.ac.be

0039–0526/99/48001

statement, some Bayesian or frequentist is certain to say ‘I am a Bayesian or frequentist but would never do that’. For example, with the growing realization of the practical difficulties of implementing a purely subjective Bayesian paradigm, some statisticians have turned to evaluating techniques derived through the Bayes formula by long run frequentist criteria; I see them as essentially frequentists. Or consider how many Bayesians use frequency probabilities for observables. For me, the quintessential frequentist evaluates inference methods according to their long run properties in repeated sampling, whereas the equivalent Bayesian depends on methods being coherent after applying the Bayes formula to personal priors elicited before observing the data.

Statistical inference has evolved in many and unforeseeable ways in the past few decades. Previously, ‘classical’ frequentist procedures seemed to hold uncontested supremacy, in spite of the major contributions of R. A. Fisher (and Neyman’s denial that inference was possible). Much of the impetus for this change has come from the computerization of statistical analysis and from the Bayesian critique. One of the major features of modern statistical inference is the central role of the likelihood function, which was introduced long before by Fisher. With the rewriting of statistical history, few modern statisticians realize that Fisher was never a frequentist for inference (although he generally used a frequency interpretation of probability), being much closer to the Bayesians, such as Jeffreys, than to the Neyman–Pearson school.

In the following text, I look at some of these dominant ideas, old and new, especially as they are currently practised in applied statistics. However, first, let me clarify the area to be discussed. I am concerned with statistical *inference*, in the Fisherian sense of obtaining the maximum information from the given data about the question at hand, without incorporating prior knowledge and information, other than that required in the construction of the model (or empirically available as likelihood functions from previous studies, but that is anticipating the argument to follow), and without concern for the way in which the conclusions will subsequently be used (although this may be reflected in how the question is posed). I thus exclude decision procedures that require some supplementary model of human behaviour.

I shall use the term knowledge to refer to available scientific theories, expressible as models, and information to refer to empirically observable data, always filtered through those theories. Thus, both of these must be distinguished from (prior) personal beliefs.

I restrict attention to model selection and estimation of precision, i.e. to exploratory rather than confirmatory inference. In current scientific research, at least in my experience, more than 95% of the inferential work of most applied statisticians is exploratory, as opposed to testing one precise model conceived before the present collection of data. Scientists usually come to the statistician with a vague model formulation to be filled out by empirical data collection, not with a specific hypothesis to test. Powerful computers and easily accessible statistical packages have made this appear to be a routine task where the statistician may not even be thought to be necessary. When scientists reach the stage of having a clear simple hypothesis to test, they often no longer need sophisticated statistics: the facts may almost seem to speak for themselves. (Phase III trials in pharmaceutical research are a major exception where statisticians are required, at least formally.) Theories of statistical inference have not adjusted to these elementary facts.

As I see it, two basic principles should underlie the choice among statistical inference procedures.

(a) In mathematics, selected axioms can be used to prove a theorem; in contrast, to draw statistical inferences, *all* the information in the available empirical data about a set of models, i.e. about the question at hand, must be used if the conclusions are not to be arbitrary.

(b) Models are rather crude simplifications of reality, with epistemological, but no ontological, status. If observables are important, inferences must be invariant to the parameterizations in which such models are specified, their application in prediction being a typical example.

Many statisticians are willing to hedge the first principle by using only the *relevant* empirical information, as may be the case when an 'appropriate' test statistic is chosen. This could leave inference open to the arbitrary: without a clear definition, the relevant can be chosen to prove a point. (Models can also be so chosen, but they are there for critics to scrutinize.) It is generally agreed that this principle means that inference must be based on the likelihood function, at least when comparing models and estimating the precision of parameters. (The problem of model checking is more complex; criteria for using all the information are not generally available.) Thus, models need not be constructed to use all the information in the data; the likelihood function determines what is relevant to the problem. In turn, this function obeys the second principle. Both principles together exclude many frequentist procedures, especially the asymptotic ones: for example, for small samples, confidence intervals based on standard errors do not use efficiently the information in the data and are not parameterization invariant. The second excludes many practical Bayesian procedures: for continuous parameters, for example, regions of highest posterior density are not preserved under non-linear transformations whereas those based on tail areas are.

Having said this, we must then ask what is left. The common criticisms of both the frequentist and the Bayesian schools are sufficiently well known and do not need to be restated. I rather consider just how much can be done using minimal assumptions related to interpreting a likelihood function. Many of the ideas presented here arose while I was doing the research for Lindsey (1996a), although most are not necessarily explicitly stated in that text. The path breaking, but largely ignored, paper of Barnard *et al*. (1962) is fundamental reading.

1.1. Example

I shall use a simple example to illustrate some of the points discussed later, although obviously it will not be appropriate for all of them. In such limited space, I can only provide a parody of a proper analysis of the scientific question under consideration.

The estimation of the incidence of acquired immune deficiency syndrome (AIDS) is an important task in modern societies. We would like to obtain a stochastic model that will describe succinctly how the epidemic has evolved globally within a given country until the present and that may be useful for prediction in relation to health service requirements. Data are obtained from doctors who must report all cases to a central office in their country. Thus, we study detection, not incidence. However, reporting may take time so we have a second stochastic process describing the arrival of the doctors' reports.

One difficult problem arises because of these delays. Some reports may take several months to arrive so, at any given time, the observations over recent periods are incomplete. Thus, the available information can be represented as a two-way contingency table with the periods of interest as rows and the reporting delays as columns. A triangle in the bottom right-hand corner is missing. For examples of such tables, reported by quarter year, see De Angelis and Gilks (1994) for England and Wales and Hay and Wolak (1994) for the USA. In such a table, the row totals give the number of detected AIDS cases per quarter but the most recent values are too small because of the missing triangle. For these two data sets, the observed marginal totals are plotted in Fig. 1 along with the fitted curves from some models that I shall discuss below. We see how the reported numbers drop dramatically in the most recent quarters because of the missing data.

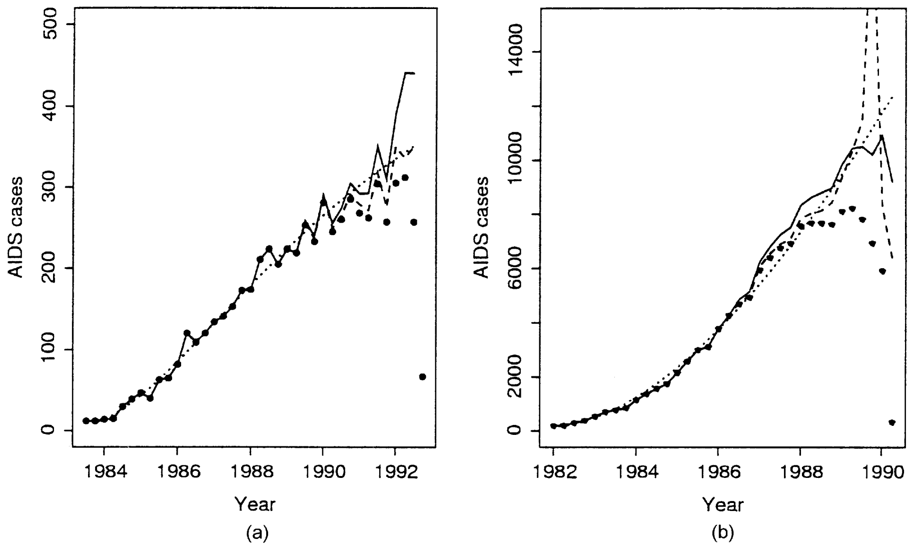

Fig. 1. Observed totals (•) and several models to estimate AIDS cases in (a) England and Wales and (b) the USA: ———, stationary nonparametric; - - - - - -, non-stationary nonparametric; ··········, non-stationary parametric

2. True models

Most statisticians admit that their models are in no sense 'true', although some argue that there is a true underlying model; many like to cite a well-known phrase of Box (1976, 1979) that all models are wrong but some are useful (which models depend on the question at hand, might I add). Nevertheless, in practice, statisticians, when making inferences, do not act as if this were so: almost all modern statistical procedures, Bayesian and frequentist, are implicitly based on the assumption that some one of the models under consideration is correct. This is a necessary condition for decision-making but antithetical to scientific inference, a first fundamental distinction between the two.

The assumption that a model is true is sometimes clearly and openly counter-intuitive; for example, a frequentist tests whether a null hypothesis is correct, generally hoping that it is not, so that the alternative hypothesis can be retained. In contrast, the Fisherian interpretation of a test is through P-values providing weight of evidence against a hypothesis; they do not require an alternative hypothesis. (Alternative hypotheses, and the accompanying power calculations, were introduced by frequentists to choose among Fisherian tests.) Asymptotic theory is more subtle; it finds its rationale in an estimated model converging to the true model as the number of observations increases: hence, for example, the importance of asymptotic consistency, discussed below, in frequentist procedures.

In a similar way, the Bayesian prior gives the probabilities that each of the various models under consideration is the true model. The likelihood principle, also discussed below, carries the assumption that the model function is true and that only the correct parameter values need to be determined.

Many will argue that these are 'as if' situations: conditional statements, not to be taken literally. But all statistical thought is permeated by this attitude which, if we take modelling seriously, has far-reaching negative implications, from teaching to applications.

Thus, much of statistical inference revolves around the idea that there is some ‘correct’ model and that the job of the statistician is to find it. But, the very possibility of making a statistical inference, and indeed all scientific work, depends on marginalization. All differences among observational units that are not thought to be pertinent for the question at hand are ignored. In other words, for generalization to be possible, the observed unique individuals must be assumed to be ‘representative’ and exchangeable within those subgroups that are being distinguished. This is always an arbitrary, but necessary, assumption, which is never true. The only ‘model’ that one (not I) might call true would be that describing the history of every subatomic particle since the beginning of the universe, and that history is unobservable because observation would modify it. All other models involve marginalization that assumes that seemingly irrelevant differences can be ignored. (See Lindsey and Lambert (1998) for further discussion.)

We should rather think in terms of models being appropriate simplifications, which are useful for detecting and understanding generalizable patterns in data, and for prediction; the type of patterns to be detected and understood, the predictions to be made and the models to be used will depend on the questions to be answered. The complexity of the model will determine how much data need to be collected. By definition, no model, in science or in statistics, is ever correct; nor does an underlying true model exist. A model is a simplification of reality, constructed to aid in understanding it; as stated in my second principle, it is an epistemological tool.

2.1. Example

In studying the evolution of AIDS, we are interested in the overall trend. An appropriate model can assume a smooth curve over time, ignoring the more ‘correct’ model that allows for variations in detection over the days of the week, with fewer cases observed on week-ends. An even ‘truer’ model would have to account for the doctors’ office hours during each day. (Of course, such information is not available from the data as presented above.)

In a ‘true’ model, various subgroups, such as homosexuals or drug users, should also be taken into account. But we must stop somewhere, before introducing sufficient characteristics to identify individuals uniquely, or we shall be unable to make any generalizations, i.e. any inferences. For the question at hand, the overall evolution of AIDS in a country, such subgroups may be ignored although modelling them and then marginalizing may produce more precise results.

3. Probability statements

Virtually all statisticians are convinced that inference statements must be made in terms of probability. It provides a metric by which such statements, made in different contexts, are thought to be comparable. However, for a probability distribution, and any model based on it, to exist, the space of *all* possible events must be defined, not just those events observed or directly under consideration. In this sense, probability is fundamentally deductive: we must know exactly what is possible before starting. This is fine for decision procedures, but not for inference, which is inductive. Why then do statisticians insist on using probability for inference?

Statisticians are not agreed about either the definition of inference probabilities or to what such statements should refer. Two widely accepted justifications can be found in the literature:

(a) the probability of errors in inference statements can be controlled in the long run and
(b) if coherent personal prior probabilities of unknowns can be elicited, before obtaining new information, and updated by the Bayes formula the posterior will also have a coherent probabilistic interpretation.

The Fisherian enumeration of all possibilities under a hypothesis, yielding a P-value indicating whether or not the observations are rare, could conceivably be used with either a frequentist or personal probability interpretation. Another approach, Fisher's fiducial probability, is not widely accepted.

The case against frequentist probability statements is well known. Among other things, they refer to and depend on unobserved outcomes that are not relevant to the information and question at hand. (This can be reasonable for model criticism, as with the Fisherian test.) The related problems with Bayesian probability statements have been little discussed, except for the objection that they are not 'objective', no empirical check being available.

The Fisherian P-value is perhaps the most widely used inference probability. Judging, in the sense of a Fisherian significance test, whether the observed data are rare under a probability-based statistical model relies on the existence of a space of observable events, the sample space. The judgment is rather weak: for given data, it is not unique because rareness can be defined in many ways. Such tests give an indication of *absolute* lack of fit of one complete model, providing no measure for comparison among models (there is no alternative hypothesis). The more observations there are, the more chance there is of detecting lack of fit of the given model, a point to which I return later. In contrast, the likelihood function only allows a comparison among models, providing means of checking only relative, and not absolute, adequacy.

Frequentist probability statements are not based on the likelihood *function* (although some involve the likelihood ratio statistic). They use goodness-of-fit criteria in an attempt to compare models; such statements are not comparable outside the context in which they arise, except through long run arguments that are not pertinent to the data at hand. This is one basis of the case against inference statements based on frequentist probability.

The fact that a frequentist test has been chosen to have power against some alternative is not an advantage over a Fisherian test in the model selection context (where the Fisherian test would not be used anyway). It is misleading because the test will also have power against many other alternatives. Developing a model in the direction of the alternative chosen generally leads off along a path to an inappropriate final model. For example, do autocorrelated residuals in a time series model indicate the wrong Markovian order, missing or too many covariates or non-linearity? Indeed, a hypothesis may even be rejected, not because it is wrong, but because the stochastic component of the model in which it is embedded is inappropriate.

Other arguments against frequentist inference include the assumption that hypotheses are formulated before looking at the data, which is impossible in an exploratory context, and the problem of sequential testing, both discussed below. The probabilities arising from frequentist inference statements do not adequately measure uncertainty.

Let us now turn to the Bayesian approach. The argument in favour of subjective Bayesian probabilities relies crucially on the principle of coherence. Unfortunately, this has only been shown to apply to finitely additive sets of models; the extension, even to countably additive sets, has not been demonstrated; it is simply widely assumed 'for mathematical convenience' (Hill (1988) and Bernardo and Smith (1994), pages 44–45). Thus, the probabilistic content of most Bayesian inference statements is no more than an unproven assumption.

Besides coherence, the other criteria for the validity of Bayesian probabilistic statements are that the prior be personal and that it be specified before obtaining new information. Bayesians who do not have their own personal prior distribution, but who propose to use a flat prior, to try various priors to check robustness, to use some mathematically convenient diffuse prior or to choose a 'prior' after inspecting the data, do not fulfil these criteria. In no case do their posterior 'probabilities' have a personal probability interpretation, although some may be calibrated to have long run frequentist properties. Indeed, these people do not have the courage to admit that inferences can be drawn directly from the likelihood function. This is what they are, in fact, doing, at least in the first two cases. For those who are capable of formulating or eliciting a personal prior, its role should be, not to be combined with, but to be compared with the likelihood function to check their agreement, say in the discussion section of a paper; 'sceptical' and 'enthusiastic' priors can also be useful in this context (see, for example, Spiegelhalter *et al.* (1994)). Anyone who has studied closely the debate about the selection of prior distributions, which has been ably summarized by Kass and Wasserman (1996), must surely have some doubts about using Bayesian techniques for inference.

For a prior probability distribution to exist, the space, here of all possible models to be considered, must be completely defined. Any models that are not included, or that have zero prior probability, will always have zero posterior probability under the application of the Bayes formula. *No* empirical information can give unforeseen models positive posterior probability. This use of probability directly contradicts the very foundation of scientific inference, from the specific to the general, where the general can never be completely known or foreseen. Under formal Bayesian procedures, scientific *discovery* is impossible.

In contrast, in decision-making, such a complete enumeration of all possibilities is a prerequisite. This is a second fundamental distinction between scientific inference and decision-making.

At least theoretically, Bayesians can easily place personal priors on parameters within models. However, as mentioned earlier, many Bayesian conclusions, e.g. about the credibility region in which a set of parameters most probably lies, are not parameterization invariant (unless they change their prior appropriately when they change the parameterization; see Wasserman (1989)). As well, putting priors on structural subspaces (i.e. on different model functions) of the space of all possible models is extremely difficult. For example, how can models based on logistic and probit regression be directly compared? Various solutions have been proposed, the two most well known being Bayes factors that place point probability masses on the various subspaces (see Kass and Raftery (1995) for a review) and embedding in a wider class, although the latter would have to be available, with the appropriate priors, before observing the data, for the posterior to have a probabilistic interpretation. No solution is widely accepted.

We must conclude that 'probabilities' arising from almost all Bayesian inference statements in realistic applied contexts do not have a rigorous probabilistic interpretation; they cannot adequately measure uncertainty.

Frequentist conclusions, based on the sample space, are critically dependent on the choice of a model's stochastic structure. The procedures are always implicitly testing this choice, as well as the explicit hypothesis about parameters: a model may be rejected because some irrelevant part of it is misspecified. Thus, a frequentist must adopt a model that only makes assumptions that are really necessary to the questions to be answered (explaining the attractiveness of nonparametric methods, as discussed further later). In contrast, Bayesian conclusions about parameters depend only on the observed data and the prior, but always conditionally that the overall stochastic model is true. Here, the global set of model functions under consideration, that for which priors were constructed before observing the data, cannot be placed in question, whereas in the frequentist approach conclusions cannot be drawn about parameters, i.e. about the problem of interest, without simultaneously questioning the validity of the model function itself.

If we conclude that the 'probability' statements currently used in drawing statistical conclusions, whether Bayesian or frequentist, generally have no reasonable probabilistic interpretation, they cannot provide us with an appropriate foundation for scientific inference, at least in the exploratory context. We shall encounter further reasons as we proceed. We are then left with the likelihood function which always has a concrete meaning, as we shall see. It provides the means of incorporating prior knowledge and information, obtaining point estimates and precision of parameters, comparing models, including error propagation, model criticism and prediction, all without probability statements, except about the observed data (Lindsey, 1996a).

The global probability model under consideration represents that part of present theoretical knowledge which is not in question in the study while prior empirical information can be incorporated by using the likelihood functions from previous studies. Model selection and precision estimation will involve looking at ratios of likelihoods (Fisher (1959), pages 68–75, and Sprott and Kalbfleisch (1965)). Because the likelihood function only allows comparisons of models, for a given data set, it is usually normed by dividing by its maximum value, so that this value is important, not as a point estimate, but as a reference point. When the likelihood function is used in this way for inference, there is no conceptual difference between comparing the probability of the data for various parameter values within a given model function and comparing various model functions (Lindsey, 1974a). This provides a powerful sense of coherence to this approach. In contrast, except for the Akaike information criterion (AIC), no widely accepted methods for a direct comparison of non-nested models are available, within either the Bayesian or frequentist frameworks.

3.1. Example

The reporting of AIDS is a one-time affair, not a sample from some well-defined population, so any frequentist sample space can only be some arbitrary conceptual construct. The models used by De Angelis and Gilks (1994) assume that AIDS detection and reporting delays are independent and that these two processes can jump arbitrarily between time points. Their first model is equivalent to independence in the contingency table:

$$\log(\mu_{ij}) = \alpha_i + \beta_j \tag{1}$$

where μ_{ij} is the mean number of reported cases with i and j respectively indexing discrete quarters and reporting delays. It assumes that the distribution of reporting delays is stationary over time. The predicted marginal curve is plotted as the full curve in Fig. 1(a). This model has a deviance of 716.5 with 413 degrees of freedom. The Fisherian test of significance for independence has a P-value of zero to the precision of my computer, providing strong evidence against this model. This probability, although equivalent to that from a frequentist test, has no clear long run interpretation in classical frequentist inference.

De Angelis and Gilks (1994) used reasonably non-informative normal and inverse Gaussian priors for the parameters in a second independence model. They did not describe how these (personal?) priors were elicited. They only considered these two rather similar models, giving no priors for the first, so their prior probability of any other possibility, including non-stationarity, must have been 0. This contradicts the test just performed, but, given this prior belief, such a problem cannot be detected from within the Bayesian paradigm.

4. Likelihood principle

In support of basing inferences on the likelihood function, one customarily invokes the likelihood principle. One statement of this is ‘All information about θ obtainable from an experiment is contained in the likelihood function for θ given $\mathbf{x}$. Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to one another.’ (Berger and Wolpert (1988), p. 19; see also Birnbaum (1962)). The phrase ‘all information about θ’ must be taken in the very narrow sense of parameter estimation, given the model function. The principal consequence is that such inferences should only depend on the observed data and hence not on the sample space. The most common example is the identity of the binomial and negative binomial likelihood functions although the models have very different sample spaces.

If this principle is taken literally, an interpretation of the model under consideration becomes impossible. Few if any models have meaning outside their sample space. An important use of models is to predict possible observations, which is impossible without the sample space. Also, the interpretation of model parameters, an essential part of inference, depends on the sample space. For example, the mean parameter of the binomial distribution is not the same as that of the negative binomial distribution.

However, the major problem with this principle is that it assumes that the model function is known to be true. The only inference problem is then to decide on an appropriate set of values for the parameter, i.e. a point estimate and precision. It is simply not correct that 'by likelihood function we generally mean a quantity that is presumed to contain all information, *given the data*, about the problem at hand' (Bjørnstad, 1996). The likelihood principle excludes model criticism.

In fact, applying the likelihood principle almost invariably involves discarding information, in the wider sense than simply estimation, about the parameter of interest. For example, the binomial experiment can be performed, and the probability estimated, even if the only information provided by a study is the total number of successes, given the total number of trials (trials may be simultaneous). In contrast, the negative binomial experiment can only be performed, and the probability can only be estimated, when we know the order of the successes and failures, which is necessary to know when to stop. This information, about whether or not the parameter being estimated is constant over time, is thrown away when the usual likelihood function for the negative binomial model is constructed. (For the binomial model, when the ordering information is available, it obviously also must be used.) An appropriate likelihood function for a negative binomial experiment must use this information and will not be identical with that for the minimal binomial experiment.

Bayesians who wish to criticize models must either resort to frequentist methods (Box, 1980; Gelman *et al*., 1996) or embed their model of interest in some wider model (Box and Tiao, 1973). For the latter, called model elaboration, to be truly Bayesian, it must, however, be specified before looking at the data. Again, the unexpected is excluded.

Thus, the likelihood principle is hollow, of no real use in inference. The appropriate likelihood function for an experiment must take into account the sample space. Basically different experiments, such as the binomial and negative binomial experiments, are performed for different reasons and can always yield different information so the corresponding likelihoods must be different. (For further detail and discussion, see Lindsey (1997a).)

4.1. Example

The likelihood principle, applied to the stationary model above, allows us to make inferential statements about the parameters in that model, assuming that it is correct. But it provides no way of criticizing that model, in contrast with the Fisherian test used above, nor of searching for a better alternative.

5. Likelihood function

Let us now look more closely at the definition of likelihood. The likelihood function was originally defined (Fisher, 1922), and is still usually interpreted, to give the (relative) probability of the observed data as a function of the various models under consideration. Models that make the data more probable are said to be more likely.

However, in almost all mathematical statistics texts, the likelihood function is directly defined as being proportional to the assumed probability *density* of the observations **y**,

$$L(\boldsymbol{\theta};\, \mathbf{y}) \propto f(\mathbf{y};\, \boldsymbol{\theta}), \tag{2}$$

where the proportionality constant does not involve the unknown parameters $\boldsymbol{\theta}$. For continuous variables, this is simply wrong if we wish to interpret likelihood as the relative probability of the observed data. The probability of any point value of a continuous variable is 0. Thus, equation (2) does not conform to the original definition. This change of definition has had ramifications throughout the theory of statistical inference. For example, at the most benign level, it has led to many paradoxes.

One sign of the mathematization of statistics, and of its distance from applied statistics, is that inference procedures, from the empirical to the general, do not allow for the elementary fact that all observations are measured to finite precision. I am not speaking about random measurement errors, but about the fixed finite, in principle known, precision or resolution of any instrument. If this is denoted by Δ, then the correct likelihood for independent observations of a continuous variable, according to Fisher's (1922) original definition (he was very careful about it), is

$$L(\boldsymbol{\theta};\, \mathbf{y}) = \prod_i \int_{y_i-\Delta_i/2}^{y_i+\Delta_i/2} f(u_i;\, \boldsymbol{\theta})\, \mathrm{d}u_i. \tag{3}$$

In virtually all cases where the likelihood function is accused of being unsuitable for inference, including by Bayesians who claim that a prior will correct the problem, the reason is the more recent alternative definition. (The few others arise from an incorrect specification of the model or inappropriate data to study a model; see Lindsey (1996a), pages 118–120 and 139.)

A classic example of the supposed problems with the likelihood function is that it (actually the density) can take infinite values, as for the three-parameter log-normal distribution (Hill, 1963). Defined as in equation (3), the likelihood can never become infinite. Indeed, any data transformation that does not take into account Δ_i can lead to trouble.

Let us be clear; I am not arguing against *models* using continuous variables. These are essential in many applications. The problem is rather with how to draw correct inferences about them, given the available observable data. Nor am I suggesting that equation (2) should never be used; it has been a valuable numerical approximation, given the analytic complexity of equation (3), although it is rarely needed any more, with modern computing power. Thus, when problems arise, and even if they do not arise, we should never forget that this is just an approximation.

Of course, this conclusion is rather embarrassing because such basic tools as the exponential family and sufficient statistics fall by the wayside for continuous variables, except as approximations (see the discussion in Heitjan (1989)). The sufficient statistics contain only an approximation to the information in the data about a parameter in real observational conditions. Bayesian procedures can also have problems: if several indistinguishable values are recorded (which is impossible for a 'continuous' variable), the posterior may not be defined.

5.1. Example

The detection of AIDS cases is a process occurring over continuous time. Nevertheless, for reporting, it is discretized into convenient administrative intervals, quarters in our case. Because of the high rate of occurrence of events, this does not create problems in fitting models, whether in discrete or continuous time, for the question of interest. The same situation also arises in survival studies, but there the event rate is much lower so time intervals may be of more concern.

6. Normal distribution

If all observations are empirically discrete, we must re-evaluate the role of the normal distribution. That this distribution has long been central both in modelling and in inference is primarily the fault of Fisher in his conflict with Pearson. Pearson was greatly concerned with developing appropriate models to describe observed reality, but he used crude moment methods of estimation. In contrast, Fisher concentrated, not on finding the most appropriate models, but on the biasedness (from the design, not the frequentist bias discussed below) and efficiency of statistical procedures. For necessary practical reasons of tractability, that some believe to continue to hold to this day, the normal distribution was central to developing analyses of randomized experiments and, more generally, to efficient inferences (although the exponential and location–scale families also emerged from his work, almost in passing).

Thus, a key factor in the continued dominance of the normal distribution in applied statistics has been this tractability for computations. This was overthrown, in the modelling context, by the introduction of survival and generalized linear models. Developments in computer-intensive methods are leading to the same sort of revolution in inference. Analytically complex likelihood functions can now be explored numerically, although this has so far been dominated by Bayesian applications (see, however, Thompson (1994)).

Here, I propose that, just as the normal distribution is not the primary basis, although still an important one, of model construction, it should not be the foundation of inference procedures. Instead, I suggest that, if one distribution is to predominate, it should be the Poisson distribution. A Poisson process is the model of pure randomness for discrete events. This can be used to profit as a general framework within which models are constructed and for the subsequent inferences.

All *observations* are discrete, whatever the supposed underlying data generating mechanism and accompanying model. Thus, all observations, whether positive or negative, are events. Current modelling practices for continuous variables do not allow for this. The Poisson distribution provides an interface between the model for the data generating mechanism and inferences that takes into account the resolution of the measuring instrument.

A model based on *any* distribution, even the Poisson distribution, for a data generating mechanism can be fitted, to the precision of the observations, by such an approach based on the Poisson distribution for the measurement process. This distribution yields a log-linear model when the model under study is a member of the exponential family (approximating the likelihood by the density for continuous variables) and when it is a multiplicative intensity stochastic process (using a similar approximation for continuous time), perhaps indicating why these families are so widely used (Lindsey and Mersch, 1992; Lindsey, 1995a, b).

A Poisson process describes random events. One basic purpose of a model is to separate the systematic pattern of interest, the signal, from the random noise. Suppose that the combination of a distribution, whichever is chosen, and a systematic regression part can be developed in such a way that the model adequately describes the mean of a Poisson distribution for the occurrence of the events actually observed. Then, we know that the maximum information about the pattern of interest has been extracted, leaving only random noise.

A principle that has not yet failed me is that any model or inference procedure is untrustworthy if it depends on the unique properties of linearity and/or the normal distribution, i.e. if it cannot directly be generalized beyond them. If the point of view that the Poisson distribution is central to statistical inference is adopted, what we then require are goodness-of-fit procedures to judge how far from the randomness of a Poisson process is that part of the variability in the data that remains unexplained by a model.

6.1. Example

As is well known, the Poisson distribution forms the basis for estimating log-linear models for contingency tables. Thus, I implicitly used this in fitting my stationary model above and shall again for those that follow. However, as in most cases, this procedure has another interpretation (Lindsey, 1995a). In this case, I am fitting risk (hazard or intensity) functions for a pair of non-homogeneous (i.e. with rates changing over time) Poisson processes for incidence and reporting delays. Above, they were step functions described by α_i and β_j.

Below, I shall consider models that allow both AIDS detection and reporting to vary continuously over time instead of taking discrete jumps but that are, nevertheless, estimated as Poisson processes. In the context of such stochastic processes, a suitable model will be a model that is conditionally Poisson at any discretely observed point in time, given the previous history of the process (Lindsey, 1995b). That used below is a generalization of a (continuous time) bivariate Weibull distribution for the times between events (AIDS detection and reporting).

7. Model selection by testing

One fundamental question remains open: how can ratios of likelihoods be compared for models of different complexity, i.e. when different numbers of parameters are involved? How can likelihood ratios be calibrated? The number of degrees of freedom in the probability statements of hypothesis tests has traditionally fulfilled this role. Before considering this calibration of the likelihood function by significance levels, let us first look at stepwise testing more closely. Nester (1996) has admirably described some of the problems with hypothesis testing.

Everyone agrees that, theoretically for a test to be valid, the hypothesis must be formulated before inspecting the data. No one does this. (I exaggerate; protocols of phase III trials do, but that is not a model selection problem.) In their inference course, students are taught this basic principle of hypothesis testing; in their regression course, they are taught to use testing sequentially, for stepwise selection.

Consider the case where two statisticians, independently, apply the same frequentist test to the same data. One formulates it in advance, the other does not. Both will obtain the same value. However, according to standard inference theory, only the first can reject the hypothesis at his or her stated probability level. The second must confirm the conclusion by collecting further data, and indeed the level of uncertainty reported about a given model should reflect what prior exploratory work was carried out with those data. However, if two statisticians follow formal Bayesian procedures to arrive at the same final test starting from these different initial conditions, the reason must be their differing initial priors.

Tests at a given probability level, based on some distribution such as the χ^2-distribution, are generally assumed to be comparable in the same way in any situation, even if that probability level has no meaning in itself, for example because we have previously looked at the data. However, in a sequence of tests, the degree of uncertainty is accumulating in unmeasurable ways, because of previous selections performed on the same data (even looking at the data before a first test will modify uncertainty in different ways in different contexts), so here the series of significance levels are certainly not comparable. The probability levels obtained from such tests are meaningless, even for internal comparisons.

Bayesian methods have their own problems with model selection, besides the fact that one's personal prior distribution must be altered as soon as one looks at the data. (What are the allowable effects of data cleaning and of studying descriptive statistics on one's personal prior?) With a prior probability density on a continuous parameter, the probability of the parameter being 0, and hence being eliminated from the model, is 0. Some Bayesians have argued that the question is rather if the parameter is sufficiently close to 0 in practical terms (see, for example, Berger and Sellke (1987)) but no widely accepted means for operationalizing this have been proposed. For example, what suitable intervals around this zero are parameterization invariant? Placing a point prior probability mass on such a model, as with Bayes factors, aggravates the situation, leading to Lindley's (1957) paradox: with increasing information, the model with point prior probability necessarily dominates those having continuous prior probability.

7.1. Example

A model for dependence between AIDS detection and reporting delays is equivalent to a non-stationary stochastic process for reporting delays: the shape of the delay curve is changing over time. Because of the missing triangle of observations, a saturated model cannot usefully be fitted to this contingency table. Instead, consider a slightly simpler interaction model that is still non-parametric (Lindsey, 1996b)

$$\log(\mu_{ij}) = \alpha_i + \beta_j + \delta_i t + \gamma_j u \qquad (4)$$

where t and u respectively are the quarter and the delay times. This yields a deviance of 585.4 with 364 degrees of freedom. The curve is plotted as the broken curve in Fig. 1(a). The deviance is improved by 131.1 with 49 degrees of freedom. The P-value of the corresponding test is 2.0×10^{-9}, apparently strong evidence against stationarity.

Here, I have decided on the necessity to find such a model after seeing the poor fit of the stationary model used by De Angelis and Gilks (1994). What then is the long run frequentist (as opposed to frequency) interpretation of this second P-value, even ignoring the fact that we do not have a sample and that the study can never be repeated in the same conditions?

As a Bayesian, how could I handle such a model if I had foreseen its need before obtaining the data? It contains 101 interdependent parameters for which a personal multivariate prior distribution would have had to be elicited. Should I work with multiplicative factors for the risk function or additive terms for the log-risk? The two can give quite different credibility intervals of highest posterior density for the parameters. How should I weight the prior probabilities of the two models, with and without stationarity? How should I study the possible elimination of the 49 dependence (interaction) parameters, given that the posterior probability of their being 0 would always be 0 (unless they were excluded in advance)?

8. Stepwise and simultaneous testing

I now return to the calibration problem. A change in deviance (minus twice the log-likelihood) between two models of, say, five has a different meaning depending on whether one or 10 parameters are fixed in the second model in relation to the first. It has been argued, and is commonly assumed, that a probability distribution, such as the χ^2-distribution, appropriately calibrates the likelihood function to allow comparisons under such different numbers of degrees of freedom even though the P-value has no probabilistic interpretation for the reasons discussed above. This *assumption* is not often made clear to students and clients.

Let us consider more closely a simple example involving different numbers of degrees of freedom, that comparing stepwise and simultaneous parameter elimination by testing, applicable to both frequentist and Bayesian probability statements. Notice that examining credibility or confidence intervals would yield the same conclusions as the corresponding test. Take a model for the mean of a normal distribution, having unit variance, defined by

$$\mu_{ij} = \alpha_i + \beta_j$$

with each of two explanatory variables taking values 1 and -1 in a balanced factorial design. With a likelihood ratio test (here equivalent to a Wald test), the square of the ratio of each parameter estimate to its standard error will have a χ^2-distribution with 1 degree of freedom and, because of orthogonality, the sum of their squares will have a χ^2-distribution with 2 degrees of freedom under the null hypotheses that the appropriate parameters are 0. First, we are comparing a model with one parameter at a time 0 with that with neither 0; then, both 0 with neither 0. Equivalent Bayesian procedures can be given using any desired independent priors for the two parameters.

Suppose that we perform separate tests for the parameters or construct separate probability intervals for them, and that we find that the individual test statistics have values 3 and 3.5. These indicate that neither parameter is significantly different from 0 at the 5% level, or that 0 lies in each interval. Because of orthogonality, one parameter can be eliminated, then the other, conditionally on the first being 0, in either order. However, if, instead, we performed a simultaneous test for both being 0, or constructed a simultaneous probability region at the same level as above, the statistic would have a value of 6.5 with 2 degrees of freedom, indicating, again at the 5% level, that at least one could not be eliminated. When such a situation arises in practice, it is extremely difficult to explain to a client!

I call such inference procedures *incompatible*. The criticism applies both to frequentist and to Bayesian probability statements. Frequentist multiple-testing procedures, such as a Bonferroni correction, that are so often incorrectly applied in such a context, only make matters worse. The two tests for individual elimination would be made at the 2.5% level, providing even less indication that either parameter was different from 0, whereas the simultaneous test would remain unchanged.

What exactly is the problem? Calibration of the likelihood function by fixing a probability level, Bayesian or frequentist, for varying numbers of parameters induces an increasing demand on the precision of each parameter as model complexity increases. A given global probability level necessarily tightens the precision interval around each individual parameter as the total number of parameters involved increases. For example, if one parameter is tested for elimination at a time, or individual credibility or confidence intervals used, the resulting model will generally be much simpler than if many parameters are tested for removal simultaneously, or a simultaneous region at the same level used. Such procedures are biased towards complex models.

What then is the solution? The precision of each parameter must be held constant (at least on average), instead of the global precision. To avoid incompatible inferences, it can easily be demonstrated that the likelihood function must be calibrated as proportional to a^p, where $0 < a < 1$ and p is the number of parameters estimated in the model. (This is only approximate, or on average, when parameters are not orthogonal, i.e. when the likelihood function does not factor.)

In model selection, this constant a is the likelihood ratio that we consider to be on the borderline indicating that one parameter can be set to a fixed value (usually 0). More generally, for interval estimation, it is the height of the normed likelihood that one chooses to define a likelihood interval for one parameter. Then, a^p will be the corresponding height to define a simultaneous region for p parameters that is compatible with that for one parameter.

In the process of simplifying a model, we necessarily make the data less probable (unless the parameter being removed is completely redundant). This value a is the maximum relative decreased probability of the data that we are willing to accept to simplify a model by removing one parameter. The smaller is a, the wider are the likelihood regions and the simpler will be the model chosen, so that this may be called the smoothing constant. Well-known model selection criteria, such as the AIC (Akaike, 1973) and the Bayes information criterion (BIC) (Schwarz, 1978), are defined in this way. Such methods demystify the widely believed, and taught, frequentist myth that models must be nested to be comparable.

How have such contradictory methods survived so long? In many cases of model selection, the number of degrees of freedom is small and about the same for all tests, so the differences between the two procedures will then be minimal. Most important model selection conclusions are not borderline cases so the results would be the same by both approaches. In addition, mathematical statisticians have done an excellent job of convincing applied workers that the overall error rate should be held constant, instead of the more appropriate individual parameter precision. Although the former may be necessary in a decision testing situation, especially with multiple end points, it is not useful, or relevant, in a model selection process.

I can treat one further difficult issue only briefly. After model selection has been completed, it may be desirable to make inferences about one particular parameter of interest. This raises the question of how to factor likelihood functions in the presence of non-orthogonality (Kalbfleisch and Sprott (1970), Sprott (1975a, 1980) and Lindsey (1996a), chapter 6). Parameter transformation can help but we may have to accept that there is no simple answer. The Bayesian solution of averaging by integrating out the 'nuisance' parameters is not acceptable: a non-orthogonal likelihood function is indicating that the plausible values of the parameter of interest change (and thus have different interpretation) depending on the values of the nuisance parameters. Under such conditions, an average is uninterpretable (Lindsey (1996a), pages 349–352).

8.1. Example

After some further model exploration, using some rather arbitrary transformations of time and involving a number of non-nested comparisons of models, I have found a continuous time parametric model allowing for non-stationarity (Lindsey, 1996b):

$$\begin{aligned}\log(\mu_{ij}) = \phi + \xi_1 \log(t) + \xi_2/t + \theta_1 \log(u) + \theta_2/u + \nu_1 t/u + \nu_2 t \log(u) \\ + \nu_3 \log(t)/u + \nu_4 \log(t) \log(u).\end{aligned}$$

(Fitting the log-transformation of time alone yields the Weibull model.) This model has a deviance of 773.8 with 456 degrees of freedom. It has only nine parameters compared with 101 in the nonparametric non-stationary model (1). It is a simplification of that model with a deviance 188.4 larger for 92 degrees of freedom so the P-value is 1.3×10^{-8}, seeming to indicate a very inferior model. However, its smooth curve through the data, plotted as the dotted curve in Fig. 1(a), appears to indicate a good model both for describing the evolution of the epidemic and for prediction.

With a fairly large set of 6215 observations, $a = 0.22$ might be chosen as a reasonable value for the height of the normed likelihood determining the interval of precision for one parameter; this implies that the deviance must be penalized by adding three (equal to $-2\log(0.22)$) times the number of parameters. (This a is smaller than that from the AIC: $a = 1/\mathrm{e} = 0.37$ so twice the number of parameters would be added to the deviance. It is larger than that from a χ^2-test at 5% with 1 degree of freedom, $a = 0.14$, or adding 3.84 times the number of parameters, but, with p parameters, this does not change to 0.14^p.) For the stationary model, the nonparametric non-stationary model and the parametric model, the penalized deviances are respectively 872.5, 989.4 and 800.8, indicating a strong preference for the last. (Such comparisons are only relative so penalized log-likelihoods would yield the same results.) This confirms the conclusions from our intuitive inspection of the graph.

This is certainly in no way the 'true' model for the phenomenon under study, given the arbitrary transformations of time. However, it appears to be quite adequate for the question at hand, given the available data. (See Lindsey and Jones (1998) for further model selection examples.)

9. Equivalence of inference procedures

Some statisticians claim that, in practice, the competing approaches to inference give about the same results so it is not important which is used or taught to students. One may argue that, just as models are imperfect, so are inference procedures, but that they both nevertheless do a reasonably good job. However, just as we would not stick with one model in all contexts, without trying to improve on it, so we should not attempt to apply one mode of inference in all contexts, nor refuse improvements. Unlike models, inference procedures are extremely difficult to check empirically, and there is no agreement on the appropriate criteria to do this.

Although, for many competing Bayesian and frequentist procedures, it is true that the conclusions will be similar, in our general exploratory context all methods are not alike. For small and large numbers of degrees of freedom, proper model selection criteria, as described above, give *very* different results from those based on probabilities. (The standard AIC is equivalent to a χ^2-test or interval at the 5% level when there are 7 degrees of freedom.)

For those who would still maintain that probability statements can reasonably calibrate the likelihood function, especially given their behaviour in high dimensional situations, consider the following questions.

(a) The deviance can provide a measure of the distance between a given model and the data (e.g. with log-linear models for contingency tables) and has an asymptotic χ^2-distribution. We would expect models that are more complex to be generally closer to the data. Why then does the probability of a small deviance, indicating a model close to the data, become very small as the number of degrees of freedom increases? (This was a question asked by a second-year undergraduate sociology student.)

(b) The calculation of probabilities often involves integration in multidimensional spaces. How do the surface area and volume of a p-dimensional unit hypersphere change as p increases? (This is a standard undergraduate mathematics question.)

Here then is my compromise, for model comparison inferences without probability statements. Take the ratio of likelihoods for two models of interest, the frequentist likelihood ratio or Bayes factor. Maximize each over unknown parameters, because, among other things, we have no agreed measure on parameters and integration, with whatever prior, is not parameterization invariant for making easily communicable interval statements about precision. To avoid incompatibility, combine each likelihood with a 'prior' that penalizes for complexity, being proportional to a^p, with a and p defined as in the previous section. Here, p is the number of parameters over which maximization was performed; a should be specified in the protocol for the study and must be congruent with the sample size, as discussed below. This is the 'non-informative, objective prior' for each model under consideration. The resulting posterior odds then provide suitable criteria for model comparison. All conclusions about model selection will be compatible with each other and with (simultaneous) likelihood regions for parameters in the final model.

In contrast with evaluation by frequentist long run characteristics or by Bayesian coherency among conclusions, this procedure will indicate the most appropriate set of models for the data at hand (including likelihood functions from previous studies, where available and relevant). This set of models, among those considered, makes the observed data most probable given the imposed constraint on complexity. Hence, we have a third criterion for inferential evaluation. This seems to me to provide a better guarantee of scientific generalizability from the observed data than either a long run or a personal posterior probability.

Thus, we have a hierarchy of inference procedures:

(a) if the appropriate model function and the number of parameters required are not known before data collection, and one believes that there is no true model, use the above procedure for some fixed value of a (a generalization of the AIC) to discover the set of models that most adequately describes the data for the question at hand;

(b) if the appropriate model function is not known before data collection but one believes that there is a true model (for a Bayesian, it must be contained in the set of functions to be considered), use the BIC to try to find it;

(c) if, before data collection, one has a model function that is known to contain the true model, use probability-based Bayesian or frequentist procedures derived from the likelihood function to decide on parameter estimates and their accompanying intervals of precision.

The difference between Bayesian and frequentist procedures, each based on its interpretation of inference probabilities, is less than that between either of these and model selection methods based on appropriately penalized likelihoods.

9.1. Example

We saw, in the previous section, that the penalized deviance clearly rejects the nonparametric non-stationary model compared with the other two models. However, we have also seen that Fisherian tests (incorrectly used in a frequentist manner in this model selection context) strongly reject both the stationary and the parametric models in favour of nonparametric non-stationarity. These contradictory conclusions are typical of the differing inferences produced through direct likelihood and probability-based approaches when large numbers of parameters are involved. In such circumstances, a penalized likelihood, such as the AIC, will select simpler models than classical inferences based on probabilities, whether Bayesian or frequentist.

10. Standard errors

Hypothesis testing has recently come under heavy attack in several scientific disciplines such as psychology and medicine; see, for example, Gardner and Altman (1989). In some scientific journals, P-values are being banned in favour of confidence intervals, which are felt to be more informative. (In the frequentist setting, this is not true; P-values and confidence intervals, based on standard errors, are equivalent if the point estimate is available.) Thus, the formulation, $\hat{\theta} = 5.1$ (standard error 2.1), is replacing the equivalent $\hat{\theta} = 5.1$ ($P = 0.015$). Unfortunately, many scientists believe that the only way that a statistician is capable of providing a measure of precision of a parameter estimate is through its standard error. It is often very difficult to convince even some highly trained applied statisticians that parameter precision can be specified in terms of the (log-) likelihood. Also, standard statistical packages do not supply the latter likelihood-based intervals, and these can be difficult to extract.

For non-normal models, standard errors, with the sizes of samples usually required, can be extremely misleading; at the same time, without inspecting the likelihood function, it is often very difficult to detect the specific cases, perhaps one time in 20, when asymptotic approximations are so far off that such a problem has occurred. The standard error provides a quadratic approximation to the precision obtained directly from the log-likelihood function for a given parameterization (Sprott and Kalbfleisch, 1969; Sprott, 1973, 1975b). As is well known, it is not parameterization invariant. It is less widely known just how much the precision interval depends on the parameterization and how poor an approximation it can be, especially if the estimate is near the edge of the parameter space, which is another weakness of frequentist asymptotic theory.

Standard errors are a useful tool in model selection for indicating which parameters may not be necessary. Outside linear normal models, they are not trustworthy for drawing final inference conclusions about either selection or parameter precision. Nevertheless, they are demanded by many referees and editors, even of reputable statistical journals. If they come to replace P-values, they, in turn, will have to be replaced in a few years by more appropriate intervals of precision based on the likelihood function.

10.1. Example

Because of the large number of observations, standard errors provide reasonable approximations for the AIDS data. Instead, consider a (real) case of fitting a binary logistic regression involving 1068 observations that happens to be at hand. In a model with only four parameters, the software calculates one of them to be 5.97 with standard error given to be 9.46; that would apparently yield an asymptotic χ^2-value of 0.40. However, when this parameter is removed, the deviance increases by 3.06, a rather different asymptotic χ^2-value. Of course, for a frequentist, this is a pathological example because a parameter is on the boundary of the parameter space, as someone familiar with logistic regression might guess from the size of the calculated standard error. Nevertheless, this can catch the unwary and much more subtle cases are possible whenever the likelihood function is skewed.

In contrast, changes in likelihood are not misleading (except when judged asymptotically!), even on the boundary of the parameter space; they are still simply statements about the relative probability of the observed data under different models. Thus, the value of 3.06 is the proper value to use above (the other, 0.40, not even being correctly calculated). The normed profile likelihood for this parameter is plotted in Fig. 2, showing that an AIC-based interval ($a = 1/\mathrm{e}$) excludes 0. Hand and Crowder (1996), pages 101 and 113, provided other examples of contradictions between inferences from likelihood ratios and from standard errors.

11. Sample size

Let us now consider the opposing philosophies of designing a study and drawing inferences from it. The theory of almost all statistical procedures is premised on the idea that the larger the sample the better; with enough observations, we shall know the correct model exactly. Applied statisticians, when planning a study, know that this is completely incorrect. They want the smallest sample possible that will supply the required information to answer the question(s) under consideration at the least costs. The only thing that a large sample does is to allow the detection of effects that are of no practical importance, while wasting resources.

Sample size calculations are critical before undertaking any major study. However, these same statisticians who carefully calculate a sample size in their protocol often seem to forget this important point, that samples should be as small as possible, when they proceed to apply asymptotic inference procedures to the data that they have collected.

Except perhaps in certain tightly controlled physical and chemical experiments, the complexity

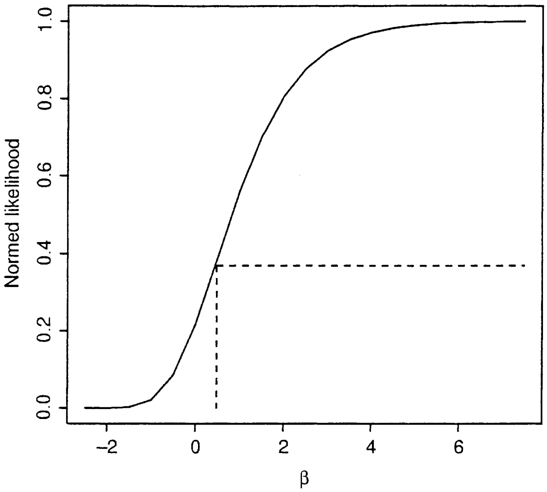

Fig. 2. Normed profile likelihood (———) for the logistic regression parameter with the standard AIC likelihood interval of $a = 1/e$ (- - - - -)

of the model required adequately to describe a data generating mechanism increases with the sample size. Consider first the case where only a response variable is observed. For samples of reasonable size, one or more simple parametric probability distributions can usually be found to fit well. But, as the sample size grows, more complicated functions will be required to account for the more complex form being revealed. The marginalization hypothesis proves increasingly inadequate. Thus, the second possibility is to turn to subgroup analysis through regression-type models. The more observations, the more explanatory variables can be introduced, and the more parameters must be estimated. At each step, the marginalization hypothesis, the foundation of the model, is modified.

Thus, this growing complexity will generally make substantively uninteresting differences statistically detectable while often hiding the very patterns that a study sets out to discover (although it may, at the same time as making the original questions impossible to answer, raise important new questions). Social scientists with contingency tables containing large frequencies are well aware of this problem when they ask statisticians how to adjust the χ^2-values, and the resulting P-values, to give sensible answers. The solution, when the sample size, for some uncontrollable reason, is too large, is to use a smaller a, the smoothing constant, in a proper model selection procedure, as described above.

The detection of practically useful parameter differences is directly related to the smoothing constant, which fixes the likelihood ratio and the two together determine the sample size. With the parameter difference of interest specified by the scientific question under study, the smoothing constant and sample size should be fixed in the protocol to have values in mutual agreement. For a given size of the substantive effect of interest, larger samples will require more smoothing, i.e. a smaller likelihood ratio comparing models with and without difference. This is the direct likelihood equivalent of frequentist power calculations (Lindsey, 1997b).

Asymptotic inference procedures, assuming that the sample size goes to ∞, are misplaced. Their only role can be in supplying certain numerical approximations to the likelihood function, or its factorization, when these can be shown to be reasonable for sufficiently small sample sizes. Asymptotic frequentist tests, if they are to be used, should be corrected to conform more closely to the (log-) likelihood function, which should never be corrected to conform to a test. Bayesians cannot take comfort in the likelihood dominating their prior asymptotically (unless they have the true model function). Unnecessarily large samples are bad. The demands by applied statisticians for accurate sample size calculations and for exact small sample inference procedures confirm this.

11.1. Example

Consider now the AIDS data for the USA for which there are 132 170 cases, compared with only 6215 for England and Wales. With so many observations, the most minute variations will be detected by using standard inference procedures. The same three models, as estimated from these data, are plotted in Fig. 1(b) (Hay and Wolak (1994) used the same nonparametric stationary model as De Angelis and Gilks (1994), shown by the full curve). Again, the parametric model appears to be most suitable for the prediction problem at hand. The nonparametric non-stationary model, the most complex of the three, fluctuates wildly near the end of the observation period. However, the difference in deviance between this and the stationary model is 2670 with 47 degrees of freedom, clearly rejecting stationarity in favour of this unstable model. The parametric model is even more clearly rejected: 5081 with 88 degrees of freedom.

With the power provided by this many observations, great detail can be detected, requiring a complex model, but irrelevant to the question at hand. Hence, I would have to penalize the likelihood very strongly to obtain a simple model. For the parametric model to be chosen over the other two, the deviance would have to be penalized by adding 60 times the number of parameters to the deviances, corresponding to $a = 10^{-13}$.

12. Consistency

The most effective way to discourage an applied statistician from using a model or method is to say that it gives asymptotically inconsistent parameter estimates. This is completely irrelevant for a fixed small sample; the interval of plausible values, not a point estimate, is essential. For a point estimate, the more pertinent Fisher consistency is applicable to finite populations. If the sample were larger, in a properly planned study, the model could be different, so the question of a parameter estimate in some fixed model converging asymptotically to a 'true' value does not arise.

However, a parameter, even a mean, has little or no valid interpretation outside the specific model from which it arises. For counts of events, the mean of a Poisson distribution has a basically different interpretation from the interpretation of the mean of its generalization to a negative binomial distribution; a given regression coefficient changes meaning depending on the other explanatory variables in the model (unless they are orthogonal). Ignoring these differences leads to contradictions and paradoxes. The choice of a new car by mean automobile fuel consumption in miles per gallon can be contradicted by that measured in litres per 100 kilometres (Hand, 1994) if the distributions to which the means refer are not specified.

Thus, if the model is held fixed as the sample size grows, it will not adequately describe the observations. But, if it is allowed to become more complex, parameters of interest will generally be estimated inconsistently, and, much more importantly, will change interpretation. Information cannot be accumulated asymptotically about a parameter.

The essential question, in a properly designed study, is whether the model describes well the *observed* data in a way that is appropriate to the question being asked. For example, a fixed effects model can be more informative than the corresponding random effects model, although the former yields inconsistent estimates. The fixed effects model allows checking of assumptions such as the form of the distribution of effects and the presence of interaction between individuals and treatments. The relative fit of the two to the given data can be compared by appropriate direct likelihood model selection criteria.

Ironically, virtually all the commonly used estimates for parameters in models with continuous response variables are inconsistent for any real data because they are based on the approximate likelihood of equation (2). These include almost all the common, in fact approximate, maximum likelihood estimates. The existence of such a lack of consistency has been known for 100 years. Sheppard's (1898) correction provides the order of the inconsistency for the variance of a normal distribution. Although Fisher was careful about this, it has been conveniently forgotten in recent asymptotic work.

Frequentists have also been much concerned with statistically unbiased estimates (as opposed to design biases that are usually far more important). The criticisms of this emphasis are well known. For example, the unbiased estimate of the variance is no longer unbiased when transformed to a standard deviation, the value most often required. Interestingly, unbiasedness adjusts the maximum likelihood estimate of the normal variance so that it becomes larger, whereas Sheppard's correction for consistency makes it smaller. For a fixed model, the biasedness of most estimates disappears with increased sample size; in contrast, the inconsistency due to the approximation of equation (2) increases in relative importance, because of the increased precision of the estimate, as do design biases.

12.1. Example

In contrast with more standard contingency table analyses, here the size of the table grows with the number of observations, as time goes by. It will also increase if the time interval is taken to be smaller than quarters. Hence, for the nonparametric models, the number of parameters is not fixed, raising questions of consistency.

For the data from the USA, the large sample size would allow one to model much more detail if that were desired. For example, seasonal fluctuations might be studied. However, by standard inference criteria, only the saturated model fits well, although it cannot provide predictions for the recent period when there is the missing triangle and hence cannot answer the questions of interest.

13. Nonparametric models

Much of modern applied statistics is dominated by the fear that the model assumptions may be wrong, ignoring the fact that all such assumptions necessarily are wrong. This is a symptom of the lack of adequate model comparison and model checking techniques, for looking at model *appropriateness*, in current statistics. This trend is fuelled by the fact, discussed above, that frequentist tests and confidence intervals cannot be used to examine parameter values without simultaneously questioning the stochastic assumptions of the model whereas Bayesian procedures must assume the overall model framework to be correct.

This fear has led to the widespread use of nonparametric and semiparametric models, especially in medical applications such as survival analysis. These, supposedly, make fewer assumptions, meaning fewer that are wrong. However, this is an illusion because the assumption that the parametric form of a model is completely unknown is just that: an assumption that will often itself actually be wrong (Fisher (1966), pages 48–49, and Sprott (1978)). We study a difference in means nonparametrically. Do we know that a stable distribution, such as the Cauchy distribution, is not involved?

A nonparametric model is generally understood to be a model where the functional form of the stochastic part is not specified so it theoretically has an infinite number of parameters; for a frequentist, this provides challenging inference problems. Paradoxically, a model in which the functional form of the systematic regression part is not completely specified, so it potentially contains an infinite number of parameters, is not called nonparametric; the frequentist can handle it routinely (by conditioning) because it does not directly influence probabilistic conclusions. Thus, a classical analysis of variance with a factor variable representing a continuous explanatory variable is parametric but a Cox (1972) proportional hazards model, with a similar representation for the base-line hazard, is semiparametric. From the point of view of the likelihood function, that only takes into account the parameters actually estimated, there is no difference between the two.

One major advantage of inferences based on parametric models is that they allow all assumptions to be checked, the available data permitting, and to be improved. In contrast, nonparametric models follow the data more closely, leaving little or no information for model checking that could indicate improvements. However, the two can be complementary: a nonparametric model can serve as a basis for judging the goodness of fit of parametric models. If none of the latter is found appropriate, one can fall back on using the former until further knowledge has accumulated.

It is commonly believed that no means exist for directly comparing parametric and nonparametric models, even for the same data. At best, diagnostics can be compared. However, from any properly specified statistical model, the probability of the observed data can be calculated; the most 'nonparametric' model simply has equal probability masses for all observations. In other words, the likelihoods of parametric and nonparametric models can be directly compared (Lindsey, 1974a, b).

This, however, leaves aside the problem of the vastly different numbers of parameters estimated in models of the two types. If the likelihood function is calibrated by means of a probability distribution and its number of degrees of freedom (and the frequentist problems, in such a context, resolved), nonparametric models will almost invariably win. Such comparisons using a probability calibration greatly favour complex models for the reasons discussed above. In contrast, with properly compatible model selection methods based on a^p, parametric models will generally be competitive and often superior, depending on the amount of scientific knowledge that is available to develop the model, as well as on the size of a chosen, i.e. the degree of smoothing required. A parametric model will be judged superior if, in the simplest case, its penalized likelihood is greater than a^{n-1}/n^n; this is the penalized nonparametric likelihood for n distinct independent observations with a point mass at each. Because of the increased information that good parametric models provide, including an indication of their possible lack of fit compared with a nonparametric model, the comparison is always worth making.

13.1. Example

Throughout the development of my example, I have been directly comparing parametric and nonparametric models. We have seen how testing led to rejection of the simpler parametric model for both data sets. In the saturated model, all n_{ij} observations reported with the same delay j for the same quarter i have probability mass $1/n_{ij}$; combining these yields its maximized likelihood. Thus, this model is the most general nonparametric model for these data. For England and Wales, there are 465 entries in the table corresponding to as many parameter estimates in the saturated model that has zero deviance. Thus the penalized deviance is 1395, which is much poorer than for the other models, as seen from the values given above.

14. Robustness

A second answer to the fear of inappropriate model assumption has been to look for robust inference procedures. Much of this search has occurred in the context of point estimation, and decision problems, that are not the subject here. Thus, in many situations, the normal distribution and the *t*-test are widely considered to be robust under moderate non-normality (Box and Tiao (1973), p. 152).

One symbol of robustness is that parameter estimates are not being unduly influenced by extreme observations. Some applied statisticians are haunted by the possibility of outliers. They believe that they must modify or eliminate them for inferences to be valid, instead of modifying the model (if the observations are not in error) or accepting them as rare occurrences. Outliers are often the most informative part of a data set, whether by telling us that the data were poorly collected or that our models are inadequate. Inference procedures should not automatically mask or adjust for such possibilities but should highlight them.

Outliers, and other diagnostic results, are defined only in terms of some model. Unfortunately, in the same way as for a test rejecting a null hypothesis, they never indicate one unique way of correcting the problem. Possible scientific ways to proceed include looking at competing models or having some more global theoretical model within which the model under consideration is embedded, in which cases direct likelihood methods are applicable. However, *ad hoc* patching of the current model is probably more common.

Many statisticians, especially frequentists for reasons developed earlier, profess a preference for results that are insensitive to any assumptions that are not directly of interest. In contrast, I claim that both a good model and a good inference procedure should be as *sensitive* as possible to the assumptions being made. Thus, a robust model will be a model that contains parameters that allows it to adapt to, and to explain, the situation (Sprott, 1982). This will then provide us with ways to detect inadequacies in our approach, hopefully giving indications of means by which the model, and our understanding of the phenomenon under study, can be further improved. However, if one's attitude is simply to answer the research question at hand, then one may argue in favour of models that are insensitive to (apparently) irrelevant incorrectness.

14.1. Example

The functional form of the parametric model developed for England and Wales also describes the data from the USA reasonably well, as seen from Fig. 1, although the parameter values are very different, especially because of the differences in the reporting-delay process. A similar, although functionally different, model can be developed specifically for these data which fits very much better. However, there is little visible difference when it is graphed, showing the robustness of the parametric models for the question at hand, at least for predictions in recently past quarters.

15. Discussion

These comments have been made in the context of (scientific) model development, the principal activity of most applied statisticians. Many are not applicable in that rarer, but not less important, process of model testing, or in decision-making. However, in spite of its name, present confirmatory inference is essentially negative: confirming that a null hypothesis can be rejected. This perhaps explains why confirmatory phase III industrial clinical trials, requiring tens of thousands of subjects while testing one end point as imposed by regulatory agencies, do not yield scientific information that is proportional to their size and cost. Owing to the inadequacy of exploratory inference in phases I and II for specifying the hypothesis to be tested, if proper goodness-of-fit criteria were ever applied to the model for the alternative hypothesis in phase III, it could be very embarrassing.

Certain of the practices that I have criticized, such as the use of the normal distribution in constructing models and in developing asymptotic inference criteria, originally arose through the practical need to overcome the limitations of the technology of the times. With the advancement of technology, especially that of computing power, such restrictions are no longer necessary.

Other practices have arisen directly through the development of mathematical statistics, out of touch with the realities of scientific research and applied statistics. Based on well-reasoned arguments and deep mathematical proofs, they are the more difficult to fight. Reinforcing this, many people first encounter statistics through introductory courses presenting little other than these inference principles, taught by mathematically trained instructors and containing few hints about useful models or real applications. After a valiant struggle to master the contradictions, they come to accept these principles as the only ones possible, and that statistics is a difficult subject of little practical use, a necessary evil.

Nevertheless, applied statistics seems often to be leading mathematical statistics in the development of inference procedures. Consider, for example, how, increasingly widely, non-nested deviances are directly compared in the applied statistics literature or, preferably, the AIC is used for model comparisons, in place of tests, in spite of its known 'poor' asymptotic properties for selecting the 'true' model.

We practise statistics without ever questioning, or even seeing, these basic contradictions. Thus, for example, we tell students and clients

(a) that all models are wrong, but that only consistent estimates are acceptable,
(b) that we can calculate the minimum necessary sample size, but that we must then draw inferences as if it should ideally have been infinite,
(c) that hypotheses must be formulated before looking at the data, but that testing is appropriate as a stepwise selection procedure,
(d) that confidence intervals are probability statements about their limits, but we interpret them as probabilities about the parameters, and
(e) that inferences must obey the likelihood principle, then we use Jeffreys or reference priors.

Unnecessary underlying principles—probability-based inference statements, true models, the likelihood as probability density, asymptotic accumulation of information, consistency, model nesting, minimizing model assumptions—have wide-ranging and often unnoticed harmful effects. The role of the statistician is to question and to criticize; let us do so with our own foundations.

Acknowledgements

This is a revised version of an invited paper presented at the Symposium on the Foundations of Statistical Inference in Honour of David Sprott (University of Waterloo, October 3rd–4th, 1996).

Murray Aitkin, Mark Becker, Denis de Crombrugghe, David Firth, Dan Heitjan, David Hinkley, Colin James, Philippe Lambert, Tom Louis, Peter McCullagh, John Nelder, Franz Palm, Stephen Senn and John Whitehead, as well as five referees, including David Draper and David Hand, provided valuable comments, criticisms and suggestions.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Inference Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest: Akademiai Kiàdo.

Barnard, G. A., Jenkins, G. M. and Winsten, C. B. (1962) Likelihood inference and time series. *J. R. Statist. Soc.* A, **125**, 321–352.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. *J. Am. Statist. Ass.,* **82**, 112–139.

Berger, J. O. and Wolpert, R. L. (1988) *The Likelihood Principle: a Review, Generalizations, and Statistical Implications*. Hayward: Institute of Mathematical Statistics.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory.* New York: Wiley.

Birnbaum, A. (1962) On the foundations of statistical inference *J. Am. Statist. Ass.*, **57**, 269–306.

Bjørnstad, J. F. (1996) On the generalization of the likelihood function and the likelihood principle. *J. Am. Statist. Ass.*, **91**, 791–806.

Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.

———(1979) Robustness in the strategy of scientific model building. In *Robustness in Statistics* (eds R. L. Launer and G. N. Wilkinson), pp. 201–236. New York: Academic Press.

———(1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

De Angelis, D. and Gilks, W. R. (1994) Estimating acquired immune deficiency syndrome incidence accounting for reporting delay. *J. R. Statist. Soc.* A, **157**, 31–40.

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond.* A, **222**, 309–368.

———(1959) *Statistical Methods and Scientific Inference*, 2nd edn. Edinburgh: Oliver and Boyd.

———(1966) *Design of Experiments*. Edinburgh: Oliver and Boyd.

Gardner, M. J. and Altman, D. G. (1989) *Statistics with Confidence*. London: British Medical Journal.

Gelman, A., Meng, X. L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sin.*, **6**, 733–807.

Hand, D. J. (1994) Deconstructing statistical questions (with discussion). *J. R. Statist. Soc.* A, **157**, 317–356.

Hand, D. J. and Crowder, M. (1996) *Practical Longitudinal Data Analysis*. London: Chapman and Hall.

Hay, J. W. and Wolak, F. A. (1994) A procedure for estimating the unconditional cumulative incidence curve and its variability for the human immunodeficiency virus. *Appl. Statist.*, **43**, 599–624.

Heitjan, D. F. (1989) Inference from grouped continuous data: a review. *Statist. Sci.*, **4**, 164–183.

Hill, B. M. (1963) The three-parameter lognormal distribution and Bayesian analysis of a point-source epidemic. *J. Am. Statist. Ass.*, **58**, 72–84.

———(1988) De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference. *Bayes. Statist.*, **3**, 211–241.

Kalbfleisch, J. D. and Sprott, D. A. (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. R. Statist. Soc.* B, **32**, 175–208.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.

Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *J. Am. Statist. Ass.*, **91**, 1343–1370.

Lindley, D. V. (1957) A statistical paradox. *Biometrika*, **44**, 187–192.

Lindsey, J. K. (1974a) Comparison of probability distributions. *J. R. Statist. Soc.* B, **36**, 38–47.

———(1974b) Construction and comparison of statistical models. *J. R. Statist. Soc.* B, **36**, 418–425.

———(1995a) *Modelling Frequency and Count Data.* Oxford: Oxford University Press.

———(1995b) Fitting parametric counting processes by using log-linear models. *Appl. Statist.,* **44**, 201–212.

———(1996a) *Parametric Statistical Inference*. Oxford: Oxford University Press.

———(1996b) Fitting bivariate intensity functions, with an application to modelling delays in reporting acquired immune deficiency syndrome. *J. R. Statist. Soc.* A, **159**, 125–131.

———(1997a) Stopping rules and the likelihood function. *J. Statist. Planng Inf.*, **59**, 167–177.

———(1997b) Exact sample size calculations for exponential family models. *Statistician*, **46**, 231–237.

Lindsey, J. K. and Jones, B. (1998) Choosing among generalized linear models applied to medical data. *Statist. Med.*, **17**, 59–68.

Lindsey, J. K. and Lambert, P. (1998) On the appropriateness of marginal models for repeated measurements in clinical trials. *Statist. Med.*, **17**, 447–469.

Lindsey, J. K. and Mersch, G. (1992) Fitting and comparing probability distribution with log linear models. *Comput. Statist. Data Anal.*, **13**, 373–384.

Nester, M. R. (1996) An applied statistician's creed. *Appl. Statist.*, **45**, 401–410.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Sheppard, W. F. (1898) On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. *Proc. Lond. Math. Soc.*, **29**, 353–380.

Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. (1994) Bayesian approaches to randomized trials (with discussion). *J. R. Statist. Soc.* A, **157**, 357–416.

Sprott, D. A. (1973) Normal likelihoods and their relation to large sample theory of estimation. *Biometrika*, **60**, 457–465.

———(1975a) Marginal and conditional sufficiency. *Biometrika*, **62**, 599–605.

———(1975b) Application of maximum likelihood methods for finite samples. *Sankhya* B, **37**, 259–270.

———(1978) Robustness and non-parametric procedures are not the only or the safe alternatives to normality. *Can. J.*

Psychol., **32**, 180–185.
———(1980) Maximum likelihood in small samples: estimation in the presence of nuisance parameters. *Biometrika*, **67**, 515–523.
———(1982) Robustness and maximum likelihood estimation. *Communs Statist. Theory Meth.*, **11**, 2513–2529.
Sprott, D. A. and Kalbfleisch, J. D. (1969) Examples of likelihoods and comparison with point estimates and large sample approximations. *J. Am. Statist. Ass.*, **64**, 468–484.
Sprott, D. A. and Kalbfleisch, J. G. (1965) The use of the likelihood function in inference. *Psychol. Bull.*, **64**, 15–22.
Thompson, E. A. (1994) Monte Carlo likelihood in genetic mapping. *Statist. Sci.*, **9**, 355–366.
Wasserman, L. A. (1989) A robust Bayesian interpretation of likelihood regions. *Ann. Statist.*, **17**, 1387–1393.

Discussion on the paper by Lindsey

David J. Hand (*The Open University, Milton Keynes*)
I would like to ask for some clarification of how you see the relative roles of model and question in data analysis. In particular, it seems to me that the emphasis is often on the model, whereas it should be on the question. You argue, on p. 23, that models and inference procedures should be as sensitive as possible to the assumptions, while also saying on that page that one may argue in favour of models which are insensitive to irrelevant incorrectness. Please can you clarify your position on this?

The trick of model building is to choose a model which does not overfit the data, in the sense that the models do not fit the idiosyncrasies due to chance events, as well as fitting the underlying structure. Recent years have witnessed the development of many methods for this. Statisticians often solve the problem by choosing from a restricted family of models—linear or generalized linear models, for example—or use a penalized goodness-of-fit measure. The neural network people, building on the extreme flexibility of their models, have developed strategies for smoothing overfitted models, as have people working on rule induction and tree methods in machine learning. From the field of computational learning theory we have methods which choose a model which has minimum variance from a set of overfitted models, and so on. Many of these approaches are rather *ad hoc*. One of the things that I like about this paper is that it gives a *principled* criterion to choose between models. You suggest that the criterion of *compatibility* is attractive, and I agree. Based on this, you suggest calibrating the likelihood function in terms of a^p. By choosing a smaller a one increases one's confidence that the included effects are real. However, you also suggest choosing a smaller a so that only those effects which are practically relevant are included in the model. It seems to me that this is mixing up two uses of the a-parameter: its use as a gauge to determine how confident we want to be that an effect is real before we agree to include it in the model, and its use to force the model to be simpler for practical reasons. The first of these is an attempt to reflect the underlying truth, whereas the second forces it away from this truth. It is not clear that this is the best way to achieve the second objective. Perhaps a better way would be to build the best model that you can and then to simplify it relative to the research question. This also allows one to take the substantive importance of different effects into account.

On p. 20 you say ‘a parameter, even a mean, has little or no valid interpretation outside the specific model from which it arises’. And you illustrate this with my fuel consumption example, saying that it is resolved if the distributions to which the mean refers are specified. But this is not the case. The aspects of the model which matter here are the empirical relationships which the numbers were chosen to represent. Taking account of the Jacobian in moving between the distribution of fuel consumption in miles per gallon and its reciprocal does not solve the problem of which of the two competing empirical systems we wish to draw a conclusion about.

You say that the model you present in Section 8.1 ‘is certainly in no way the “true” model’, but surely it is, in fact, fundamentally misspecified and cannot possibly be legitimate. In the model, u is the delay time. As such it is a ratio scale and we may express it in different units. If we do this we see that an extra term, $\upsilon_5 t$ is introduced into the model. If I fit a corrected model including a term in t, I obtain a penalized deviance of 803.8, greater than the deviance of 800.8 for the incorrect model that you give, which is presumably why you rejected the former. However, if I use (fractions of) years for the delay time (approximated by dividing the given times by 4), I obtain a penalized deviance for your model of 825.9, which is substantially worse than that of my corrected model, which, of course, remains at 803.8. The misspecification of your model means that the deviance is a function of the units that you happened to use, which is absurd. In view of the elegant compatible formalism that you have introduced for comparing models, it is ironic that an unfortunate choice of model form has led to ‘incompatibility’ across a choice of units!

Of course, this is a detail which is easily corrected, and it does not detract from what is a very interesting and valuable paper. It gives me genuine pleasure to propose the vote of thanks.

David Draper (*University of Bath*)
There is much to admire and agree with in this paper; in particular the author deserves our thanks for tackling practical issues in what might be termed the *theory of applied statistics*. I especially like Lindsey's remarks on the irrelevence of

(a) asymptotic consistency and
(b) the likelihood principle when model structure is not known with certainty (i.e., in real problems, always).

However, there is also much with which to disagree, and it is on these disagreements that I will focus here, ordering my comments from mild to severe criticism of the author's positions.

(a) In Section 2 Lindsey says,

> 'Many will argue that these [points of view on the existence of a "true" model] are "as if" situations: conditional statements, not to be taken literally'.

In other words, he is saying that this is mathematics, by which I mean deductive reasoning: if I make assumption $A = \{\text{a true model exists}\}$, then conclusion C follows. But—recalling (e.g. de Finetti (1974, 1975)) that all statistical inference problems may usefully be cast predictively, with the ingredients

(i) already observed data y,
(ii) as yet unobserved data y^*,
(iii) context $\mathcal{C}$ about how the data were gathered and
(iv) assumptions $\mathcal{A}$ about how y and y^* are related

—it seems to me that the best that we can ever do (no matter what our religious views on Bayes may be) is conditional conclusions, e.g. of the form $p(y^*|y, \mathcal{C}, \mathcal{A})$: assumption-free inference is not possible. Thus to me statistical inference is inherently deductive, just like the rest of mathematics; we cannot escape our assumptions. (Even Fisher, the orginal 'likelihoodlum'—whose views receive surprisingly little attention in this paper—is vague on this point, e.g. Fisher (1955).) With his advocated methods the author certainly draws conclusions based on assumptions; so in what sense is he reasoning inductively?

(b) Lindsey's failure to address decision-making undercuts many of his criticisms. For example, model selection is best regarded as a decision problem: as I and others have noted elsewhere (e.g. Bernardo and Smith (1994) and Draper (1998a)), to choose a model you have to say to what purpose the model will be put, for how else will you know whether your model is sufficiently good? If the author took this seriously (specifying utility functions based on context) he would discover that he can dispense with the *ad hoc* nature of the choice of his quantity a.

(c) Lindsey claims in Section 8 that

> 'The Bayesian solution [to the problem of marginalization] of averaging by integrating out the "nuisance" parameters is not acceptable: a non-orthogonal likelihood function is indicating that the plausible values of the parameter of interest change (*and thus have different interpretation*) depending on the values of the nuisance parameters'

(emphasis added). This is arrant nonsense. For instance, in the location model

$$y_i = \mu + \sigma e_i \qquad e_i \overset{\text{IID}}{\sim} t_\nu,$$

it is true that σ and ν are typically correlated in the posterior distribution (e.g. with a diffuse prior), but the *meaning* of ν as an index of tail weight does not change from one σ-value to another, and the weighted average

$$p(\nu|y) = \int_0^\infty p(\nu|\sigma,\, y)\, p(\sigma|y)\, \mathrm{d}\sigma$$

of conditional distributions $p(\nu|\sigma,\, y)$ is perfectly interpretable as a summary of model uncertaintly about the need for the t-distribution instead of the simpler Gaussian distribution.

(d) The author's Bayesian is a straw man who overemphasizes coherence, relies on highest posterior density regions, has never heard of cross-validation and is well out of date on methodology and outlook. As a practising applied Bayesian I do not recognize Lindsey's caricature in myself or in (the vast majority of) my colleagues, and I commend resources such as Gatsonis *et al.* (1997) and Bernardo *et al.* (1998) to the author's attention if he wishes to form a more accurate view. Several quotes from the paper will bring this disparity into focus.

'For me, the quintessential ... Bayesian depends on methods being coherent after applying the Bayes formula to personal priors elicited before observing the data.'

What a narrow, old-fashioned view of Bayes this is! Personal priors and coherence need not be all; out-of-sample predictive calibration (see below) can help to tune the choice of prior and model structure.

'For continuous parameters, ... regions of highest posterior density are not preserved under non-linear transformations whereas those based on tail areas are.'

The full posterior is invariant; and nobody forces highest posterior density summaries on you (posterior summarization is another decision problem, for which tail areas may be an adequate working approximate solution).

'The argument in favour of subjective Bayesian probabilities relies crucially on the principle of coherence, [which unfortunately] has only been shown to apply to finitely additive sets of models.'

Coherence is overrated as a justification for Bayes (in the sense that it is necessary but not sufficient in good applied work), and the author overemphasizes it. The best that it can offer is an assurance of internal consistency in probability judgments; to be useful scientific (and decision-making) co-workers, we must aspire to more as well: to external consistency (calibration). For me the argument for the Bayesian paradigm is that *it works*: it allows the acknowledgement of an appropriately wider set of uncertainties, e.g. about model structure as well as about unknowns conditional on structure; and it makes the crucial topic of prediction central to the modelling.

'For a prior probability distribution to exist, the space ... of all possible models to be considered, must be completely defined. Any models that are not included, or that have zero prior probability, will always have zero posterior probability under the application of the Bayes formula. No empirical information can give unforeseen models positive posterior probability. ... [Thus under] formal Bayesian procedures, scientific discovery is impossible.'

The first part of this is true, and well known to Bayesians as *Cromwell's rule* (Lindley, 1972). But, firstly, with Bayesian nonparametric methods based on Markov chain Monte Carlo computations (e.g. Walker *et al.* (1999) and Draper *et al.* (1998)) we can now put positive prior probability on the space of all (scientifically defensible) models in interesting problems of real complexity, and, secondly, we can use out-of-sample-predictive calibration (e.g. Gelfand *et al.* (1992) and Draper (1998b)) to learn about appropriate priors on model structure without using the data twice. Both of these approaches effectively take the sting out of Cromwell's rule in applied problems: we are free to find out that our initial set of model structural possibilities was not sufficiently rich and to expand it without cheating.

To summarize, the author has clearly shrugged off the shackles of frequentist inference successfully (if ever he was so shackled), and the strength of his belief in the value of the likelihood function has brought him to the very doorstep of Bayes, so—with the same sort of evangelical spirit with which Lindsey here proselytizes us about likelihood—I urge him to walk boldly through the Bayesian door: with modern computing and predictive validation to keep him calibrated, I think that he will find himself right at home.

The vote of thanks was passed by acclamation.

Chris Chatfield (*University of Bath*)

This paper is about 'heresies'. The *Oxford English Dictionary* defines a heresy as: 'an opinion contrary to orthodox or accepted doctrine'. But what is accepted doctrine in statistics? It is surely not the viewpoint of a hardline Bayesian or an ardent frequentist. Rather we need to look to the views of the 'typical applied statistician' who has to solve real problems. While agreeing that differences in philosophy exist in the narrow field of inference, the views of most statisticians are much more in harmony than might be thought from the literature, with common ground on the tenets of *good statistical practice*, such as clarifying objectives clearly and collecting good data.

The paper says much about what Bayesians and frequentists might do, but such labels can be divisive and I prefer the label 'statistician'. Most applied statisticians will use whichever inferential approach seems appropriate to solve a particular problem. Perhaps the real heretics are statisticians who cannot agree that 'we should not attempt to apply one mode of inference in all contexts'.

The biggest fiction of statistical inference is that there is a true model known *a priori*. Surely it is no longer a heresy to say this? There is a growing literature (e.g. Chatfield (1995) and Draper (1995)) on *model uncertainty* topics, such as model mixing and sensitivity analysis. With no 'true' model, most arguments about the foundations of inference cease to have much meaning and many inferential 'heresies' evaporate. For example model comparison is clearly more important than model testing if one accepts that one is looking for an adequate approximation. It may help to expand inference explicitly to include *model formulation*, but even then it would only include part of what a statistician does. The statistician's job is to *solve problems*, and not (just) to estimate parameters.

How should the paper change what I do? I would have liked more examples as the example on acquired immune deficiency syndrome was not helpful to me. In Fig. 1, I think that I can make better forecasts 'by eye' than those given by any of the models.

Overall the paper makes many good points but is less controversial than it thinks it is and I do not tell my students or clients any of the points (a)–(e) listed in Section 15. I do agree that statistics teaching, and the statistical literature, should more truthfully reflect what statisticians actually do.

Murray Aitkin (*University of Newcastle*)

I am so much in agreement with Jim's main points that I shall simply list a few points of disagreement.

The likelihood principle

I find the likelihood principle persuasive in its narrow form. The different confidence coverages of the likelihood interval for the positive and negative binomial experiments refer to reference sets of samples from different hypothetical replications of the experiment. The likelihood interval refers to the experiment actually performed. Real replications, as in Jim's predictions, require the explicit model for the new data. There is no conflict with the likelihood principle in this.

Models

The 'true model' is an oxymoron. Model specifications are generally deficient because they omit the *neglected* part. The usual normal regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),$$

fails to distinguish between normal population variation about the model mean μ_x for fixed x and the departure of the model mean from the actual population mean. The representation that we need is

$$Y_i|x_i \sim N(\mu_i, \sigma^2), \qquad \mu_i = \alpha + \beta x_i + \phi_i,$$

where ϕ_i is the neglected part. Giving ϕ an arbitrary distribution, this *overdispersed* model can be fitted by nonparametric maximum likelihood as a finite mixture, as in Aitkin (1996).

Model comparisons

How should model comparisons be made, in a pure likelihood framework? I have suggested in Aitkin (1997) a generalization of Dempster's (1974, 1997) posterior distribution of the likelihood ratio. In the simplest single-parameter case, H_0 specifies $\theta = \theta_0$, and H_1 is general. The likelihood ratio

$$\mathrm{LR} = L(\theta_0)/L(\theta)$$

has a posterior distribution derived from that for θ. We can therefore compute the posterior probability that $\mathrm{LR} < 0.1$ or any other value which would suggest strong evidence against H_0. Stone (1997) gives an example of Bernoulli trials with $n = 527135$ and $r = 106298$, with H_0: $p = 0.2$. The observed proportion $\hat{p} = 0.20165$ is 3 standard errors from H_0, with P-value 0.0027. For a uniform prior on p, the Bayes factor for H_0 to H_1 is 8 and the fractional Bayes factor is 4.88, whereas the posterior Bayes factor is 0.0156, in agreement with the P-value. The posterior density of p for the same uniform prior is $N(0.20165, 0.00055^2)$, so H_0 is far outside the 99% highest posterior density interval for p, contradicting the Bayes factor conclusion. Equivalently, the posterior probability that $\mathrm{LR} < 0.1$ is 0.964. (The posterior probability that $\mathrm{LR} < 1$ is $1 - P = 0.9973$.)

This approach is focused on likelihood ratios as measures of strength of evidence, and it expresses conclusions in posterior probability terms which require no more than conventional diffuse or reference priors. At the same time P-values can be expressed explicitly as posterior probabilities, providing a unification of Bayes, likelihood and repeated sampling inferential statements for nested model comparisons.

D. R. Cox (*University of Oxford*)
I was a little baffled by this paper. There are certainly important and constructive points to agree with and warnings to be heeded but I found it rather disconcerting that Professor Lindsey attributes views to others so firmly. In particular, his descriptions of conventional statistical thinking are not recognizable to me, as a very conventional statistician.

For example, as a keen Bayesian I was surprised to read that priors *must* be chosen before the data have been seen. Nothing in the formalism demands this. Prior does not refer to time, but to a situation, hypothetical when we have data, where we assess what our evidence would have been if we had had no data. This assessment may rationally be affected by having seen the data, although there are considerable dangers in this, rather similar to those in frequentist theory.

As an enthusiastic frequentist, again some statements seem to me not part of a sensible frequential formulation. On a more technical point, it is not true that a frequentist approach is unable to deal with non-nested hypotheses; there is an extensive literature on this over nearly 40 years. Although there are difficulties with this, there is no problem in principle.

Professor Lindsey's liking for a direct use of the likelihood is very understandable, but the difficulties of nuisance parameters are not really addressed. We need some form of calibration to avoid the difficulties of the Neyman–Scott type as well as misbehaviour of the profile likelihood in other much simpler problems.

Some of the difficulties which Professor Lindsey, and many others, have with frequentist theory as it is sometimes presented are based on a failure to distinguish between the operational hypothetical procedures that define terms like significance level and confidence interval, and advice on how to use them. The narrower the gap between these two the better but there is a clear difference. Thus, defining a significance level via a Neyman–Pearson approach requires a decision-like attitude which Neyman himself did not use when he analysed data.

Two final comments are that all careful accounts of asymptotic theory stress that the limiting notions involved are mathematical devices to generate approximations. These devices have no physical meaning and discussions that suppose that there is such a meaning are largely irrelevant.

Finally, and perhaps most importantly, the first sentence of the paper is in conflict with the empirical observation that statistical methods are now used in many fields on a scale vastly greater than even 20 years ago. Not all these applications are good and surely there are many areas where statistical methods are underused or badly used and we should try to address these. Although there are no grounds for complacency, the general note of pessimism that starts the paper seems to me misplaced.

J. A. Nelder (*Imperial College of Science, Technology and Medicine, London*)
I welcome this paper with very few reservations, but I would have preferred the title 'The likelihood school strikes back'. Many statisticians believe that inference must be either frequentist or Bayesian, which is nonsense. I hope that these statisticians will come to realize that enormous simplifications can follow from Lindsey's approach. For example, the adoption of likelihood (support) intervals for the odds ratio in 2×2 tables would lead to our discarding a large, messy and confusing literature.

I wish that the author had calibrated the log-likelihood, rather than the likelihood, function. I find it easier to deal with $\log(a) = -1$, rather than $a = 1/\mathrm{e}$. Furthermore, plots with log-likelihood (or deviance) as the ordinate are much more informative visually than those using likelihood, where the way in which the ordinate goes to zero is hard to see.

A topic missing from this paper is quasi-likelihood (QL). It is often introduced by stressing that, because it is specified by the mean and variance function only, those are the only assumptions being made: not so. From a QL we can construct an unnormalized distribution; if we normalize it we find that the normalizing constant is either exactly constant or varies only slowly with the mean. Further, the cumulants of the normalized distribution, at least up to level 4, have a pattern close to that which an exponential family would have if one existed with that variance structure. This is the assumption that we make when using QL. In the form of extended quasi-likelihood (EQL) (Nelder and Pregibon, 1987) it is identical with the unnormalized form of Efron's double-exponential families (Efron, 1986). Any EQL gives a deviance and hence Akaike information criterion and similar statistics. In summary, QL is a valuable extension to likelihood inference.

I am interested in the speaker's reaction to the idea of the h-likelihood (Lee and Nelder, 1996) as a way of generalizing the Fisher likelihood, and so likelihood inference, to models with multiple random effects. Maximization of the h-likelihood avoids the multidimensional integration that is necessary for the use of marginal likelihood, and the arbitrary specification of priors for the bottom level parameters. Markov chain Monte Carlo sampling is replaced by a generalization of iterative least squares, with computing time being reduced by orders of magnitude. I hope that the speaker will contribute to extending likelihood inference to this wider class of models.

S. M. Rizvi (*University of Westminster, London*)
I am delighted by this highly provocative and controversial paper, as it conforms to some of my own heretical views developed over the past 30 years of practising and teaching statistics. I had long suspected that Sir Roland Fisher was a clandestine Bayesian, as some time ago I derived his multivariate discriminant function from the Bayesian standpoint.

I also believe that the author's second principle (p. 2) is self-contradictory. It is necessary to appreciate that epistemology deals with questions of perception, questions of inference and questions about truth. Thus models which best conform with data—best according to one's light—are a simplification of the reality, and hence that of ontological truth. I see no contradiction in that. Further, the author's insistence that inference should be invariant to the parameterizations in which models are cast is not supported by experience, as the author himself later indicates. Once you accept a model, your inference takes a deductive form. Any inference is subservient to the parameterization, and not invariant to it. I reject his second principle.

There is no such thing as a true model or a true theory. There are only models and theories that work or do not work. Once we accept a model based on given data, there is no escape from the conclusions, at least with respect to the sample concerned. We need to accept that models are tentative and provisional. This is wholly consistent with scientific inference. That judgments of science are provisional is not appreciated even by some Nobel prize-winners.

I question the introduction of an 'intuitive or pseudo-intuitive' argument, as the author appears to do when he castigates true 'models' as counter-intuitive. We need to appreciate that intuition is an aid to perception, but it is not a ground for a discursive argument. Russell was to put it more eloquently when he equated intuition with 'animal inference'.

P-values are fine, but they do not provide us with a clear dividing line for accepting or rejecting a hypothesis. A significance test does.

I take issue with the author's insistence that inference should not be backed by probabilities. That assertions should be based on probabilities has respectable advocates. It was the philosopher John Locke who was the first to assert that the degree of assent we give to a proposition should be based on its probability.

Finally, confidence intervals are fine if you are dealing with a two-sided alternative hypothesis; they are misleading when confronted with a one-sided alternative.

Andrew Ehrenberg (*South Bank University, London*)
I am sure that we all laud Professor Lindsey's effort to face up to the many contradictions in the theoretical treatment of statistical inference.

These contradictions are, as he says, one reason why statistics is so often viewed as both difficult and unnecessary.

But even if statisticians do face up to and resolve the problems which Lindsey rightly raises it would not improve 'the low public esteem for statistics'. Statistics would still be seen as an 'unnecessary evil', because students and clients hardly ever actually *use* the tools of statistical inference which they are being offered. Nor is there any sound reason why they *should* use them.

As I see it, statistical or probabilistic inference is *at best* a relatively minor technical issue. It arises when one has just a single rather small sample, which has been properly selected from a well-defined population. That is rare.

Instead, applied statisticians and clients usually face

(a) *either* quite a large data set, or even a *very* large one,

(b) *or* data from many different but broadly comparable populations, and hence samples which are at least *cumulatively* large, or very large,

(c) *or* a small data set which is not an explicit *sample* from any defined larger population anyway (e.g. all the acquired immune deficiency syndrome (AIDS) cases in St Thomas's Hospital in the autumn of 1992).

Such data have problems of analysis, interpretation and, broadly speaking, inference. But these are not addressed either by theoretical statistics in general or in Professor Lindsey's paper in particular.

To finish with a specific point: Professor Lindsey says that in his experience more than 95% of the inferential work which is taken to statisticians for advice is *exploratory*, i.e. concerned with model development.

But what applied statisticians and clients mostly do on their own is to use *existing* models (formal or informal), rather than always to develop new ones. For example, any analysis of AIDS data should now use the prior knowledge (or 'verbal model') that there are delays in the reporting of cases of AIDS, and also in the recording of them.

Using a model is very different from first *developing* it.

G. A. Barnard (*University of Essex, Colchester*)
Lindsey's paper is very much to be welcomed for its emphasis on the fact that statistics, like mathematical physics and fluid mechanics, is a branch of *applied* mathematics in which results may be very valuable despite being only approximate. And his stress on the relevance of new computing possibilities to the logic of our field is most timely.

My only serious query arises from his remark (p. 2) that Fisher was never a frequentist for inference purposes. True, he made clear in his 1925 classic that inferences were to be expressed in terms of likelihood, not in terms of probability. And he was much closer to Jeffreys than to the Neyman–Pearson school. But likelihood ratios have a long run frequency interpretation, in terms of odds, though not in terms of probabilities. If E has been observed, the likelihood ratio for H versus H' is

$$L(E) = \Pr(E|H)/\Pr(E|H').$$

If we choose H only when $L(E)$ exceeds W, and H' only when $1/L(E)$ exceeds W, making no choice otherwise, then whenever either H or H' is true the long run frequency ratio of right choices to wrong choices will exceed W.

Limitations on computer power meant that Fisher could not implement this view in his lifetime. But now desk computers with spreadsheet facilities can exhibit various likelihood functions fully conditioned on the given data on various assumed models. We only need to worry about such differences when the spreadsheets differ seriously in relation to the questions of interest. Such cases are much rarer than unconditional approaches might lead us to think. The Cushny–Peebles data quoted by Student in 1908 and by Fisher in 1925 and more accurately by Senn (1993) illustrate this point. Cushny and Peebles were interested in the evidence for or against chirality with the drugs used. Without chirality the data distribution would be symmetrical about 0. With chirality it would be symmetrical about some non-zero value. Assuming exact normality and no chirality the odds on t falling short of rather than exceeding 4.03 are exactly 698 to 1. But if we replace the false assumption of normality with the equally false assumption of a Cauchy density we find that *for the data to hand* the Cauchy spreadsheet differs little from the normal spreadsheet. And little change results from taking proper note of the fact that the patients would not have been allowed to sleep for more than 24 hours. To relate our spreadsheets to the question at issue we may need to make assumptions about the irrelevant parameters, but we can see from the spreadsheet how much, or how little, these assumptions matter.

Stephen Senn (*University College London*)
This stimulating paper makes many excellent points not least of which is that frequentist and Bayesian approaches both require prespecified models but that this requirement is commonly ignored. However, I am not convinced that the author's implicit distinction between parameter estimation and model choice is fundamental. For example, he objects to 'incorporating prior knowledge and information, other than that required in the construction of the model', but in practice I think that there will be many cases where the applied statistician will be unsure what this Lindseyian prescription permits and forbids. Take the AB/BA crossover design (Grieve and Senn, 1998). We can consider models with and without carry-over. For this choice it seems that the author would allow prior knowledge to play *some* role, but the latter model is merely a special case of the former with λ, the carry-over effect, equal to 0. However, the

model with carry-over effects included is useless unless coupled with an informative prior that λ is likely to be small which, however, it seems his prescription forbids us to assume.

Also the problem (Section 8) of what is called *consonance* in the technical literature on multiple testing (Bauer, 1991) is not unique to statistics and there is not the slightest difficulty in explaining it to non-statisticians. Two squabbling children are separated by their mother. She knows that at least one of them is guilty of aggressive behaviour but she will often be incapable of establishing definitely that any given child is. Is this *incompatible* with human life?

Finally, a small matter: contrary to Section 15, it is the exception rather than the norm for drug development programmes, let alone individual clinical trials, to involve tens of thousands of subjects. A typical programme will involve 5000 or fewer subjects and individual trials are usually measured in the hundreds. The author's point about lost opportunities in measurement is, none-the-less, well taken.

D. V. Lindley (*Minehead*)

My comments are mostly on statements made in the paper about Bayesian ideas.

(a) 'Inference statements must be made in terms of probability' because of arguments by de Finetti and others. These take self-evident principles and from them prove that uncertainty satisfies the probability calculus. It is wrong to say that the Bayesian method is 'internally contradictory' when it is based on coherence—the avoidance of contradiction.

(b) Countable additivity is not taken 'for mathematical convenience'. There are enough examples to show that some consequences of finite additivity are unacceptable. In any case, conglomerability is self-evident for most.

(c) Likelihood is not, alone, adequate for inference because it violates one of the self-evident principles in (a). It can happen that, with sets of exclusive events $\{A_i\}$, $\{B_i\}$, A_i is more likely than B_i for all i, yet the union of the A_i is less likely than that of the B_i.

(d) Likelihood is not additive, so there is no general way of eliminating nuisance parameters. 13 varieties of likelihood testify to the struggle to overcome this handicap.

(e) The old example in Section 8 is not incompatible in the Bayesian framework. The joint distribution of α, β has margins for α and β separately. Is the suggestion that the latter are incompatible with the former?

(f) Probability is invariant. What may not be invariant are summary statistics, like the mean or highest posterior density intervals. But this is a criticism of the summary, not of the distribution. Similar comments apply to the comparison of the negative and positive binomials.

(g) Probability is always conditional on what is known, or assumed, when the uncertainty statement is made. It is not therefore true that '*all* possible events must be defined'.

(h) All of us are continually making informal prior judgments after seeing the data. Just before writing this, I read of some data on lobbyists. A reaction was to think of my prior (to the data) on lobbyists, and to recall the 'cash-for-questions' affair in the House of Commons.

(i) A model is your description of a small world in terms of probability. Like probability, it is not objectively true, but expresses a relationship between you and that world.

(j) The statement about the paradox is wrong. The model with point prior does not dominate if the alternative is true.

(k) Many years ago Savage correctly said that a model should be as big as an elephant. Until recently we have not been able to act on the advice for computational reasons. It now looks as though Markov chain Monte Carlo sampling will overcome the difficulty, but there still remains the unsolved problem of the assessment and understanding of multivariate distributions.

Nick Longford (*De Montfort University, Leicester*)

Jim Lindsey should be congratulated on a thoughtful critique of many conventions deeply ingrained in statistical practice. Many of them do have a rational basis, but only in a specific context. Our failure is that we apply them universally. I have specific comments on Section 11.

I do not see the conflict between 'the larger the sample the better' and the needs of an applied statistician. All other design factors being equal, larger sample sizes are associated with more information. The purpose of the sample size calculations is to save resources and to ensure precision that is *at least as high* as a preset standard. When a model of certain complexity (smoothness) is required it is a fault of the model selection routine that larger data sets tend to yield more complex models. No model selection procedure is very good if it functions well only for 'good' sample sizes. A more constructive approach than that implied by Lindsey's discussion is to remove, or smooth out, from the model selected by a formal *statistical* procedure any effects (differences) that are substantively unimportant. Then we can address, if need be, the issue of precision greater than the minimum acceptable.

Vaguely speaking, by the sample size calculations we aim at matching statistical significance with substantive importance. The product, the sample size n, may be way off the mark because the formula used involves residual variance and other parameters, some of which are only guessed, often with only modest precision. Also, in multipurpose studies, or studies that will by analysed by a wide constituency of secondary users, the sample sizes will be 'too large' for some inferences and too small for others. It is unfortunate with the latter, but no problem with the former. If everything else fails, simply analyse a simple random subsample of the cases.

T. A. Louis (*University of Minnesota, Minneapolis*)
Lindsey makes many excellent and many provocative points. Most of my comments derive from the view that the likelihood is key to developing inferences and that no model is correct. Though I agree on the likelihood's central role, effective designs and analyses also depend on inferential goals and criteria (see Shen and Louis (1998) on the poor performance of histograms and ranks based on posterior means).

Lindsey supports goal- and criteria-driven procedures in his consideration of Bayesian highest posterior density (HPD) regions. An HPD region has the smallest volume for a given probability content but is not transformation equivariant. Posterior tail area intervals for a scalar parameter are equivariant, and Lindsey promotes their use. However, transform equivariance is only one criterion and I shall use the HPD region when minimizing volume is important, for a skewed posterior, for a multimodal posterior or one without a regular mode and for a multivariate parameter.

Lindsey is correct that no model is correct. Unfortunately, he concludes that model complexity must increase as the sample size increases and so asymptotic analyses based on an assumed true model are irrelevant. Believing that some parameters retain their meaning with increasing sample size, I strongly disagree that asymptotic properties such as consistency are irrelevant. A large sample size *allows* us to increase complexity but should not *require* us to. Also, asymptotics can reveal important properties and justify modern methods of finite sample inference such as the bootstrap.

In choosing between models, a^p likelihood calibration effectively structures the trade-off between complexity and degrees of freedom. Selecting a is vital and more guidance is needed (what is the basis for $a = 0.22$ in Section 8.1?). However, goals and prior knowledge can be instrumental in choosing between candidate models (see Foster and George (1994)). And, selection procedures such as cross-validation have the advantages of dealing directly with predictive goals and of being relatively free of assumptions.

Lindsey goes overboard in criticizing continuous probability models. Continuous distributions provide only an approximation to (discrete) reality and have their quirks, but they can be effective in studying properties and communicating insights. Furthermore, replacing the Gaussian by the Poisson distribution will generate as many problems as it eliminates.

Lindsey criticizes the use of standard errors (SEs). His criticism is old but worth repeating. SEs should be replaced by a set of likelihood regions and other informative summaries of the likelihood (or sampling distribution or posterior distribution).

Finally, are the 'heresies' in the title current statistical practices or the author's views?

Peter McCullagh (*University of Chicago*)
In statistics, as in fashion, a model is an idealization of reality. Joe Public knows this all too well, and, although he may grumble silently, he does not ordinarily feel obliged to engage publically in incoherent Quixotic ramblings when his wife or girlfriend does not live up to the image portrayed in glamour magazines and romantic novels. Scientists know this also. To a large degree, science is the search for pattern, a search for models as an aid for understanding. Mendel's laws are good science, but it does not follow that they are universally obeyed in nature. Just as Mendel's laws do not cover all of genetics, so too neither the Bayesian nor the frequentist approach covers every facet of statistical practice. I do not see this as a bad thing for genetics or statistics; it is evidence of the richness and diversity of both subjects.

Current statistical practice presents ample opportunity for valid criticism. But sweeping unsubstantiated claims, undue emphasis on the picayune and plain errors of fact do not add up to a strong case. Despite the author's sincere sermonizing I see no internal contradictions in the Bayesian approach.

The opportunity to discuss and to reassess foundational issues is ordinarily welcome. While this paper is passionately provocative on minor points such as likelihood (3), the discussion of major issues is irritating in tone and confused on the facts. Despite this, the data analysis seems reasonably sensible, a triumph of common sense over unsound dogma. In the discussion of foundational matters, however, the parts of the paper that are true are not new, and parts that are new are not true.

R. Allan Reese (*Hull University*)
Although agreeing that the lack of shared credo and approach contributes to the poor public perception of statistics and statisticians, it is extreme to place the blame entirely on statisticians and teachers. Users of statistical methods themselves approach the subject with fear and reluctance, treating it as somewhere between abstruse mathematics and witchcraft. An inability to relate quantitative data to reality may start well before students meet statistics (Greer, 1997). From that starting point, and invariably against tight deadlines, it is difficult to instil good principles of data analysis.

There are shining examples of good practice and excellent software, and I have taught using Genstat, GLIM, SPSS and Stata. However, I have known a student attend a whole course on GLIM only to ask ‘but how do I do regression like in the book?’. Mead’s (1988) classic work contains an excellent chapter on ‘Designing useful experiments’, but this is tucked away at the back because ‘the standard format of books on design is inevitable, and therefore correct’. I wish that he had challenged that view!

Underlying the paper appears to be a philosophical question. We make assumptions and hence propose mathematical models which reflect a theory that is external to the data. We can test the models and find those which fit best. However, we can never test or verify the assumptions. Each real-life data set is unique. Even when repeating a well-established protocol under controlled conditions, we need to allow for the possibility of outliers. Hence the argument is forever circular—we trust the model because it fits the data, and we believe the data because they follow the model. In that sense, ‘all’ statistics is heuristic and all relies on inference.

This dilemma is illustrated in Fig. 1(a). A Health Minister seeing that graph would ask, ‘Is the incidence of acquired immune deficiency syndrome now going up or down?’. Because of delays in reporting, we can ‘only’ answer that question by relying on a model. The model must include assumptions about the pattern of reporting. Choose your assumptions, and you can give whichever answer you think the Minister would prefer: lies, damned lies, and . . .?

The following contributions were received in writing after the meeting.

P. M. E. Altham (*University of Cambridge*)
Professor Lindsey has given us a discussion paper which is scholarly, thoughtful and thought provoking. I have just a small point to make.

The example about acquired immune deficiency syndrome corresponds to a large and *sparse* (triangular) contingency table. The sum forming the deviance of 716.5, with 413 degrees of freedom (DF) quoted on p. 8, and the subsequent deviances for the same data set include a high proportion of cells with fitted values around 0 or 1. This must surely have a severe effect on the reliability of the χ^2-approximation. A traditional remedy is to combine some of the cells. For example, as a quick solution we combine the last eight columns of the data set of De Angelis and Gilks (1994). Then the deviance for model (1) is now 514.54 with 254 DF. Using the negative binomial fitting routine supplied in Venables and Ripley’s (1997) library, I found that the corresponding negative binomial model has deviance 337.79 with 254 DF. Considering that the total sample size is 6160, this does not seem such a severe departure from model (1) and it allows some degree of overdispersion relative to the Poisson model.

F. P. A. Coolen (*University of Durham*)
I shall mention two approaches to statistical inference and refer to comments by Lindsey. I strongly support the subjective approach to inference whenever practical situations require the use of knowledge other than information that is available in experimental or historical data. de Finetti (1974) used prevision (or ‘expectation’) instead of probability as the main tool for inference (Lindsey, Section 3), and indeed prevision for observable random quantities provides a natural concept for dealing with uncertainty. This line of work has been extended successfully by Goldstein; see the overview and further references in Goldstein (1998). One of the nice foundational aspects of this work is that a Bayes posterior is the expectation of what one’s beliefs will be after obtaining the additional information, but not necessarily equal to those beliefs (Goldstein, 1997). Of course, scientific discovery (Lindsey, Section 3) is important to a subjectivist. The fact that actual beliefs might be different from a Bayes posterior is perfectly acceptable. The subjective approach allows the use of non-observable random quantities, or parameters, but well interpretable quantities are needed for assessment. Problems with regard to

inference on uninterpretable parameters (Lindsey, Section 8) are avoided if the actual problems are expressed in terms of well interpretable random quantities.

Notwithstanding my support for the subjective approach to inference, I feel the need for a non-subjective method for induction. I was delighted to see Lindsey's reference to Hill (1988) but sad that no further attention was paid to Hill's assumption $A_{(n)}$ and related inference. In many situations (but not for example in the case of time series), both Bayesians and frequentists are happy with a finite exchangeability assumption on future real-valued observations, meaning that all possible orderings are considered as equally likely. For nonparametric methods this assumption of equally likely orderings should still hold after some data have been observed. This post-data assumption is Hill's $A_{(n)}$, which is not sufficient to derive precise probabilistic inferences in many situations but does provide bounds (Coolen, 1998) which are imprecise probabilities (Walley, 1991). I suggest that $A_{(n)}$ provides an attractive, purely predictive, method for statistical inference that can be used as a 'base-line method' (Lindsey, Section 13). Moreover, $A_{(n)}$ can deal with data described by Zipf's law (Hill, 1988), which represents more real world data than any other known law, including the Gaussian law, and I doubt whether this is true for Poisson process models (Lindsey, Section 6). The main challenge for $A_{(n)}$-based statistical inference is generalization to multivariate data (Hill, 1993).

James S. Hodges (*University of Minnesota, Minneapolis*)

Professor Lindsey's paper is most welcome and great fun. Hear, hear: 'Foundations of statistics' is dead because it is irrelevant to practice; the needs of applied statisticians should drive the new foundational debate. (Hodges (1996) takes a similar tack.) Lindsey's paper was also delightful on many specifics, e.g. Bayes factors.

None-the-less, two of Professor Lindsey's key premises are unsatisfying. The first is that 'model development [is] the principal activity of most applied statisticians'. In 15 years of highly varied statistical practice, my principal activity has been helping colleagues to answer substantive questions. *Contra* Lindsey ('Scientists usually come ... with a vague model formulation to be filled out by empirical data collection, not with a specific hypothesis to test'), my colleagues almost always have a specific purpose; though it is not always formulated as testing some $\mu = 0$, Professor Lindsey's description is still inaccurate. Moreover, selecting a model rarely matters even to me. Instead, the main issue is whether the data give the same answer to the substantive question in alternative analyses, accounting for the uncertainty that is internal to each analysis. (Oddly, Fig. 1 shows no uncertainty bounds.) A minority of such alternative analyses are motivated by statistical obsessions like outliers. More often, they involve excluding large subsets of the data, variant definitions of variables and other problems that are awkward to represent in likelihood functions and not worth enshrining in model form. The answer to the substantive question usually does not change across the alternative analyses but, when it does, in my experience no mere model selection criterion will convince a sceptical colleague or editor on a particular analysis.

This difficulty is compounded by Professor Lindsey's preferred criterion, for sceptics must buy the magic number a. But where does a come from? Sometimes the Akaike or Bayes information criterion is acceptable, but otherwise we must pick a. In Section 8.1, $a = 0.22$—not 0.21, not 0.23—is presented as 'reasonable'. Why should we prefer Professor Lindsey's view of reasonable to anyone else's? The rationale in Section 11.1 is better: we need $a = 10^{-13}$ to choose the parametric model over the others. This moves in a potentially useful direction—what does it take to change the answer?—but is still quite removed from the questions for which someone obtains money to collect data to bring to a statistician.

D. A. Sprott (*University of Waterloo*)

This is an interesting, thoughtful and well-written paper. I have only a few minor points of disagreement or emphasis.

It would be nice to believe that the criticisms of the emphasis on unbiased estimates are well known. But journals still seem intent on publishing papers correcting the maximum likelihood estimate for bias and estimating the variance of the resulting estimates even though the resulting estimating intervals contain highly implausible, if not impossible, values of the parameter.

But my main comment is about the position relegated to confirmatory or non-exploratory inference. Perhaps 95% of the work of statisticians is exploratory. But I think that the advance of science also rests on what happens after the exploratory stage—the process of demonstrating repeatability and of the subsequent accumulation of evidence. This is given short shrift by the comment that when this stage is reached sophisticated statistics is no longer needed, the facts speaking for themselves. This is undoubtedly the final result, as discussed by Fisher (1953a) and by Barnard (1972). But much, possibly sophisticated, statistics precedes it. Fisher (1953b) was himself an example of this. Perhaps the scientists who originate the problems, not statisticians, will be those who are involved in this process, as suggested by MacGregor (1997). But this would merely confirm the obvious low esteem adverted to in the introductory sentence of the paper, although the concern of statisticians about this low esteem seems less obvious.

The **author** replied later, in writing, as follows.

I am pleased to see the high level of constructive criticism by most discussants. This topic can engender strong emotions! Many question that I say anything heretical. Yet, after one earlier presentation, a theoretical statistician asked me 'Do you actually teach your students this stuff?'. One referee's conclusion was 'This is a shallow, incomprehensible, unscholarly offering that ought not to grace our Society's journals'. Heresy for the theoretician is often everyday practice for the applied statistician.

In exploratory scientific inference, as in much of applied statistics, the goal is to convince your public, here the scientific community, that the empirical data support your scientific theory better than others available. Replicability is key. In contrast, if a scientist applies independent replication to the statistical analysis process, consulting several statisticians, he or she risks receiving quite different answers!

One great weakness of statisticians is that we often have one favourite model, and inference procedure, that we tend to apply everywhere possible. In contrast, the likelihood approach is restricted to exploratory scientific model selection. It has nothing to offer in court decisions, industrial quality control or phase III clinical trial drug acceptance.

Several discussants object that they do not recognize real statisticians in my caricatures of Bayesians and frequentists, but these descriptions refer to foundation principles, not to people. Should I have replaced 'Bayesians or frequentists do' each time by 'an applied statistician strictly following Bayesian or frequentist principles at this particular moment does'? (Most theoreticians are less flexible.) If the procedures of these two schools are meant to describe good practice, why then do certain argue ('Neyman himself did not use [it] when he analysed data') that they are meant to be ignored in application? Should we not expect the same high standards from our profession as we do of scientists who consult us? Will not our public otherwise be unconvinced, because of the arbitrariness?

Virtually all the discussants seem agreed that no model is true. Few, however, address my more fundamental argument, that probability is inappropriate for exploratory inference. Comparability of likelihood contour levels with changing parameter dimension requires some function $g(p)$ for calibration: $L(\boldsymbol{\theta};\, \mathbf{y}) \propto g(p)$. For likelihood inference, it is

$$L(\boldsymbol{\theta};\, \mathbf{y}) \propto a^p; \tag{5}$$

for the frequentist approach, using the deviance,

$$L(\boldsymbol{\theta};\, \mathbf{y}) \propto \exp(-\chi^2_p/2) \tag{6}$$

whereas for the Bayesian approach it is more complex,

$$L(\boldsymbol{\theta};\, \mathbf{y}) = \frac{h\{\Pr(\boldsymbol{\theta}|\mathbf{y})\}\, f(\mathbf{y})}{\Pr(\boldsymbol{\theta})}. \tag{7}$$

Although equations (5) and (6) superficially appear most similar, they are not. Rather equations (6) and (7) both require integration, whereas equation (5) is a point function.

Several discussants ask for more details about the choice of a. As with a confidence or credibility level, it specifies the level of precision desired: how much less probable can a model make the data, compared with your best one, before you conclude that it is implausible? Along with the specification of a scientifically useful effect in planning a study, this then determines the sample size. Just as 95% intervals have been found reasonable in many practical situations, so has $a = 1/\mathrm{e}$.

I agree virtually completely with many of the discussants (George Barnard, Chris Chatfield, Andrew Ehrenberg, David Hand, James Hodges, Nick Longford, John Nelder, Allan Reese, Stephen Senn and David Sprott) and shall only reply to some of then, very briefly.

The danger of exploratory inference is that, if enough different models are considered for a given data set, some well fitting models will always be found. Thus, I agree with David Hand that the question is primary: the restricted group of models to be considered, whether before or, if necessary, after seeing the data, must be sensitive to the question but insensitive to points that are irrelevant to the question.

However, I disagree that forcing a model to be simple drives it away from the truth: the art is to use simplicity to highlight, and to convince your audience of, the real effects.

I also maintain my solution to the fuel consumption problem: the question determines which units to use, but a unique model must apply for the two units or the answers to the two questions will be fundamentally, not just superficially, contradictory. As to the problem with units in my own model (John Nelder first pointed this error out to me), I can only plead guilty: the model must be hierarchical—criticism by the scientific community in operation.

Frequentist procedures are notorious for being *ad hoc*; David Draper brings us up to date on recent patches to Bayesian theory. It will make its breakthrough in applied *confirmatory* statistics when phase III clinical trial protocols specify prior distributions. The beauty of the likelihood approach is that it yields simple clear interpretations, requiring no such arbitrariness: models are judged by how probable they make the observed data.

To convince your audience, probabilistic inference statements require prior (to data collection) guarantees: a fixed model (hypothesis) and sample space for the frequentist, a fixed set of models with their prior probabilities for the Bayesian (Senn, 1997). In this sense, they are deductive, making scientific discovery difficult or impossible. Likelihood inference allows freedom to consider models suggested by the data, with the same interpretable measures of uncertainty, although they will be less convincing than those stated beforehand, hence requiring repeatedness within the scientific community. Scientific inference is not a (necessarily individual) decision problem.

David Draper downplays the importance of coherence. But how can a posterior Bayesian calculation have a probabilistic interpretation without coherence, except in the formal mathematical sense that any set of non-negative numbers summing to 1 has?

Murray Aitkin has not yet convinced me that posterior Bayes factors can help, compared with the simple interpretation of likelihood ratios, in communicating to students and clients.

Nowhere do I claim that frequentists are unable to deal with non-nested hypotheses. As David Cox points out, methods are available, although, as I stated, they are not widely accepted. Perhaps the reason is that the underlying models are unreasonable in most contexts. Most standard courses (and all texts that I have been able to consult) do not cover these methods. Students are only taught the nested approach and infer (if they are not told!) that it is the only one possible.

The question of nuisance parameters arises *after* selecting a model, when we make inferences about specific parameters of interest; this is not the topic of the paper. Nuisance parameters create no problem for likelihood-based exploratory inference. Profiles, as summaries of a complex likelihood surface, are always interpretable in the likelihood approach (far more than marginal posteriors). The judgment that they misbehave is asymptotic (e.g. consistency). Thus asymptotic theory does not just generate approximations but permeates criteria for choosing procedures. To be acceptable (publishable), a procedure must yield consistent, asymptotically normal estimates not on the parameter space boundary, with calculable standard errors, superior asymptotic relative efficiency, and so on, all irrelevant to likelihood inference, but need not give the probability of the data (estimating equations).

I am more pessimistic than David Cox about our image. Statistics is more widely used through compulsion, not the active desire of many actual users; wide use does not imply respectability. (If you want to see the expression of real disgust, without revealing that you are a statistician talk to any laboratory assistant or junior scientist who once suffered through introductory statistics.) The main impulsion for this expansion appears to be the need for good design, not inference.

I entirely agree with John Nelder that the *log*-likelihood should be calibrated. I do that in the example. However, in didactic communication with students and clients, speaking of the probability of the observed data under various models, rather than its logarithm, is often clearer.

With modern computing power, quasi-likelihoods can easily be normalized numerically so this approach is now amenable to full likelihood inference. I have not had the occasion to work with the *h*-likelihood, perhaps because I have not yet encountered consulting problems requiring it. The large number of parameters must make complete inferences (and not just model selection) complex.

I agree with Andrew Ehrenberg that most exploratory work involves comparing existing models; I did not mean to imply that the 95% referred to developing new models.

Being a frequentist for inference purposes, as I defined it, is not the same as using a frequency interpretation of probability in inference, something that Fisher obviously did. Indeed, I discuss Fisherian *P*-values—George Barnard provides another example—and we should not forget fiducial probability. But I can find no reference where Fisher required long run guarantees of his procedures; he rather argued that significance tests provide measures of rarity of the *observed* data under given hypotheses.

If Stephen Senn's model with carry-over is useless without an informative prior, then what exactly can the data tell us about it? Can we calculate the probability of the data under this model and is the result distinct from that for the model without carry-over? If not, and we believe that a small carry-over effect is present, and important, perhaps the design should be changed to obtain the required information.

Incompatibility means that the mother would either discover, individually, that neither of the children or, collectively, that at least one is guilty of aggressive behaviour. The point is that the criterion of individual aggressiveness (precision) differs when they are judged separately and together.

Unfortunately, Dennis Lindley's 'self-evident principles' are not evident to me. He agrees that the alternative model must be *true* to avoid the paradox of point prior probabilities. This is fundamental to my rejection of this Bayesian approach to exploratory model selection. Assessment and understanding of multivariate distributions of the data are a sufficiently big problem for me without the added complication of probabilities on the parameters.

I would be interested to know how Nick Longford proposes to remove, or smooth out, from a model effects that are substantively unimportant. This seems to be developing a new model, a further step in the exploratory process.

Larger sample sizes are rarely better. Besides costs and time, problems of homogeneity, quality control, missingness, and so on, increase. I strongly oppose analysing a random subsample of the data; if they are available, use them all.

I agree with Tom Louis that a large sample perhaps should not require us to increase complexity but show me a statistical procedure that does not have this characteristic for fixed precision level, unless there is a true model.

I do not criticize continuous probability models, but rather likelihoods based on their *densities*. A likelihood function, the probability of the *observed* data, must be based on discrete probabilities even when the model is continuous.

I make no claim to newness in the paper; indeed I stated the opposite in my presentation. It would have been more helpful if Peter McCullagh had pointed out what was not true rather than making strong unsubstantiated accusations. Nor have I claimed any internal contradictions in the formal Bayesian approach based on coherent probabilities. The problem is rather when we try to implement it in a context for which it was not designed: exploratory scientific inference. As illustrated by David Draper's comments, it is now in a state of arbitrariness rivalling frequentist theory.

As Pat Altham points out, sparse contingency tables are of concern for inferences using asymptotic approximations, such as the χ^2-distribution. However, I am arguing against the latter in the context of exploratory work. Such tables do not pose a problem for likelihood inference.

Overdispersion models are meant to handle counts with dependent events, usually on the same individual unit. That is not the case here where an assumption of independently reported acquired immune deficiency syndrome events should be reasonable. Hence, I would oppose correcting for overdispersion.

I certainly agree with Jim Hodges that scientists come with a specific purpose (question), but to me the role of a statistician is to formulate it in terms of models giving the probability of the data and to select the most appropriate set. Nowhere do I argue for keeping only one model. In likelihood-based model selection, the procedure for selecting among model functions is identical with that for choosing among parameter values. Just as a likelihood region contains models with different parameter values supported by the data, so you can have an envelope of plausible model functions.

Dave Sprott is absolutely right about the importance of repeatability to scientific inference. My concern here is with *discovering* something worth attempting a demonstration of repeatability.

Resort to emotional accusations, as with one Bayesian-inclined referee and one frequentist-inclined discussant, often means that fundamental principles are threatened and no reasoned reply is available. In current statistics, the fundamental divide is not between frequentists and Bayesians but between applied statisticians and the theoreticians dominating many teaching institutions and journals. As one very eminent statistician told my son after the meeting, applied statistics is a peculiar profession where you have to unlearn much of what you were taught as a student.

References in the discussion

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statist. Comput.*, **6**, 251–262.

———(1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood (with discussion). *Statist. Comput.*, **7**, 253–272.

Barnard, G. A. (1972) The unity of statistics. *J. R. Statist. Soc.* A, **135**, 1–15.

Bauer, P. (1991) Multiple testing in clinical trials. *Statist. Med.*, **10**, 871–890.

Bernardo, J. M., Berger, J., Dawid, A. P. and Smith, A. F. M. (eds) (1998) *Bayesian Statistics 6*. Oxford: Oxford University Press. To be published.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference (with discussion). *J. R. Statist. Soc.* A, **158**, 419–466.

Coolen, F. P. A. (1998) Low structure imprecise predictive inference for Bayes' problem. *Statist. Probab. Lett.*, **36**, 349–357.

De Angelis, D. and Gilks, W. R. (1994) Estimating acquired immune deficiency syndrome incidence accounting for reporting delay. *J. R. Statist. Soc.* A, **157**, 31–40.

Dempster, A. P. (1974) The direct use of likelihood for significance testing. In *Proc. Conf. Foundational Questions in Statistical Inference* (eds O. Barndorff-Nielsen, P. Blaesild and G. Sihon), pp. 335–352. Aarhus: University of Aarhus.

———(1997) The direct use of likelihood for significance testing. *Statist. Comput.*, **7**, 247–252.

Draper, D. (1995) Assessment and propagation of uncertainty (with discussion). *J. R. Statist. Soc.* B, **57**, 45–97.

———(1998a) Discussion of "Decision models in screening for breast cancer" by G Parmigiani. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press. To be published.

———(1998b) 3CV: Bayesian model choice via out-of-sample predictive calibration. *Technical Report*. Statistics Group, Department of Mathematical Sciences, University of Bath, Bath.

Draper, D., Cheal, R. and Sinclair, J. (1998) Fixing the broken bootstrap: Bayesian non-parametric inference with skewed and long-tailed data. *Technical Report*. Statistics Group, Department of Mathematical Sciences, University of Bath, Bath.

Efron, B. (1986) Double exponential families and their use in generalized linear regression. *J. Am. Statist. Ass.*, **81**, 709–721.

de Finetti, B. (1974) *Theory of Probability*, vol. 1. Chichester: Wiley.

———(1975) *Theory of Probability*, vol. 2. Chichester: Wiley.

Fisher, R. A. (1953a) The expansion of statistics. *J. R. Statist. Soc.* A, **116**, 1–6.

———(1953b) Dispersion on a sphere. *Proc. R. Soc.* A, **217**, 295–305.

———(1955) Statistical methods and scientific induction. *J. R. Statist. Soc.* B, **17**, 69–78.

Foster, D. P. and George, E. I. (1994) The risk inflation criterion for multiple regression. *Ann. Statist.*, **22**, 1947–1975.

Gatsonis, C., Hodges, J. S., Kass, R. E. and McCulloch, R. E. (1997) *Case Studies in Bayesian Statistics*, vol. 3. New York: Springer.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. Oxford: Oxford University Press.

Goldstein, M. (1997) Prior inferences for posterior judgements. In *Proc. 10th Int. Congr. Logic, Methodology and Philosophy of Science* (eds M. L. D. Chiara, K. Doets, D. Mundici and J. van Benthem). Dordrecht: Kluwer.

———(1998) Bayes linear analysis. In *Encyclopaedia of Statistical Sciences*, update vol. 3. New York: Wiley. To be published.

Greer, B. (1997) Modelling reality in mathematics classrooms: the case of word problems. *Learnng Instructn*, **7**, 293–307.

Grieve, A. and Senn, S. J. (1998) Estimating treatment effects in clinical cross-over trials. *J. Biopharm. Statist.*, **8**, 191–233.

Hill, B. M. (1988) De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 211–241. Oxford: Oxford University Press.

———(1993) Parametric models for A_n: splitting processes and mixtures. *J. R. Statist. Soc.* B, **55**, 423–433.

Hodges, J. S. (1996) Statistical practice as argumentation: a sketch of a theory of applied statistics. In *Modeling and Prediction Honoring Seymour Geisser* (eds J. C. Lee, A. Zellner and W. O. Johnson), pp. 19–45. New York: Springer.

Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models. *J. R. Statist. Soc.* B, **58**, 619–656.

Lindley, D. V. (1972) *Bayesian Statistics, a Review*. Philadelphia: Society for Industrial and Applied Mathematics.

MacGregor, J. F. (1997) Using on-line process data to improve quality: challenges for statisticians. *Int. Statist. Rev.*, **65**, 309–323.

Mead, R. (1988) *The Design of Experiments: Statistical Principles for Practical Applications*, 2nd edn. Cambridge: Cambridge University Press.

Nelder, J. A. and Pregibon, D. (1987) An extended quasi-likelihood function. *Biometrika*, **74**, 221–232.

Senn, S. (1993) *Cross-over Trials in Clinical Research*. Chichester: Wiley.

———(1997) Present remembrance of priors past is not the same as a true prior. *Br. Med. J.*, **314**, 73.

Shen, W. and Louis, T. A. (1998) Triple-goal estimates in two-stage hierarchical models. *J. R. Statist. Soc.* B, **60**, 455–471.

Stone, M. (1997) Discussion of papers by Dempster and Aitkin. *Statist. Comput.*, **7**, 263–264.

Venables, W. N. and Ripley, B. D. (1997) *Modern Applied Statistics with S-Plus*. New York: Springer.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1999) Bayesian nonparametric inference for random distributions and related functions (with discussion). *J. R. Statist. Soc.* B, **61**, in the press.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.