Paradigms for Statistical Inference

Murray Aitkin
University of Melbourne

August 29, 2011

Abstract

This paper expresses the different theories of statistical inference in terms of the competition between scientific paradigms as described in Kuhn (1996). The conflict among the major schools is examined in Kuhn's terms, with extensive quotes, which helps to illuminate the discussion. Critical issues among the schools are evaluated with canonical examples. The scientific and social implications of conflicting conclusions from the same data by different schools are considered, and proposals are made for extending the current Bayesian paradigm, through both reformulating model comparisons, and incorporating sample survey analysis.

1 Introduction

We use the term *paradigm* in the sense of Kuhn (1996): "universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of researchers." (The term has more than one meaning, but our use will be clear from the context.) Our concern is for the relevance of Kuhn's description of the evolution, and revolutionary change, of scientific paradigms to the current state of theories of statistical inference.

In quoting Kuhn at length, we are not endorsing his view of the evolution of paradigms over the views of other philosophers of science; we find his approach and views apposite for the current state of statistics and its theories of inference.

The term paradigm has come to be commonly used in statistics in *the Bayesian paradigm*, and less frequently, in *the frequentist paradigm*. These multiple usages of the term imply, in the Kuhnian sense, that there are separate communities of researchers following these paradigms, since the paradigms are inconsistent in their principles (axioms) for the analysis of data. We will call these separate communities *schools*. This might not matter if these schools attended to different problems in the applications of statistics, but it *does* matter since the applications of the schools overlap very substantially, and the Bayesian and frequentist schools at least claim *universality* in the breadth of their applications.

There are other paradigms (the likelihood paradigm for example), and other approaches to data analysis, which will not be discussed in this paper, as they do

not have, or claim, the generality of the Bayesian and frequentist schools. The survey sampling design-based theory is not usually called a paradigm, though there are exceptions – the "total survey error paradigm" for example (Groves and Lyberg 2010). The survey sampling school had in the past a much narrower focus, and sometimes seemed to be obsessively fixed on the estimation of the population mean alone. This has changed considerably in recent years, and the claims made for the generality of its theory have widened; we will call this approach the *survey sampling paradigm*.

There are many practising statisticians who would not regard themselves as members of only one, or any, school. Many statistical scientists are agnostic about theories/paradigms: "I'm not a Bayesian, but I use what works. I'm an empiricist – if MCMC works I use it." (Without a paradigm, how can one tell if anything works?) Statements like these make plain that statistics does not have a *single* paradigm, and the conflict among the major schools is a clear example of what Kuhn calls the *pre-paradigm* state of theories:

> (p. 177) There are schools in the sciences, communities, that is, which approach the same subject from incompatible viewpoints. But they are far rarer there than in other [non-science] fields; they are always in competition, and their competition is usually quickly ended.
>
> (p. 47) Normal science can proceed without rules only so long as the relevant scientific community accepts without question the particular problem-solutions already achieved. Rules should therefore become important and the characteristic unconcern about them should vanish whenever paradigms or models are felt to be insecure.... The pre-paradigm period, in particular, is regularly marked by frequent and deep debates over legitimate methods, problems, and standards of solution, though these serve rather to define schools than to produce agreement.
>
> (p. 178) Probably the most striking [issue of scientific community structure] ... is the transition from the pre- to the post-paradigm period in the development of the scientific field. Before it occurs, a number of schools compete for the domination of a given field. Afterward, in the wake of some notable scientific achievement, the number of schools is greatly reduced, ordinarily to one, and a more efficient mode of scientific practice begins.

The aim of this paper is to contribute to a *general* framework for statistical theory – a single extended Bayesian paradigm. We believe it is fair to say that *several* notable (statistical) scientific achievements have led to the possibility of a single paradigm. These achievements are due originally to H.O. Hartley, J.N.S. Rao and W.A. Ericson, and to A.P. Dempster. In a situation familiar from Kuhn's book, these achievements were either not recognised for their importance, lacked the necessary computing technology for practical implementation (Kuhn's "instrumentation"), or were ignored.

We will not give the histories of the current schools in any detail; these are well-documented at length elsewhere (for example, Barnett 1999, Welsh 1996). A very concise exposition can be found in Chapter 1 of Aitkin (2010). Instead we will focus sharply on fundamental points of conflict, and make proposals to resolve the conflict in a single paradigm framework. We take these points of conflict in a particular order reflecting the structure of the single paradigm.

We first describe very briefly in Section 2 the evolution of paradigms for statistical theory. We discuss the Bayesian/survey sampling conflict in Section 3 and the frequentist/Bayesian conflict in Section 4. The Bayes factor *puzzle*, or *anomaly* (there is a critical difference in Kuhn's use of these terms) is discussed in Section 4, and its resolution in Section 5. General discussion is given in Section 6. The resolutions of these conflicts have considered at some length in Aitkin (2010), and we refer to this source frequently.

2 The evolution of paradigms

In Kuhn's view a current paradigm does not *evolve* into a new paradigm – there is a *revolutionary* change from the old to the new paradigm – but once adopted the paradigm is developed and articulated steadily, and defines the structure of "normal science" in the scientific field:

> (pp. 35-6) To scientists, at least, the results gained in normal research are significant because they add to the scope and precision with which the paradigm can be applied. ... Bringing a normal research problem to a conclusion is achieving the anticipated in a new way, and it requires the solution of all sorts of complex instrumental, conceptual, and mathematical puzzles. The man who succeeds proves himself an expert puzzle-solver, and the challenge of the puzzle is an important part of what usually drives him on.
>
> (p. 38) The scientific enterprise as a whole does from time to time prove useful, open up new territory, display order, and test long-accepted belief. Nevertheless, *the individual* engaged on a normal research problem *is almost never doing any of these things.* Once engaged, his motivation is of a rather different sort. What then challenges him is the conviction that, if only he is skilful enough, he will succeed in solving a puzzle that no one before has solved or solved so well. Many of the greatest scientific minds have devoted all of their professional attention to demanding puzzles of this sort. On most occasions any particular field of specialization offers nothing else to do, a fact that makes it no less fascinating to the proper sort of addict. (author's emphasis)

As the paradigm is articulated further, difficulties begin to arise – the problems become more difficult to solve, and for some problems it is not clear that a solution can be found:

> (p. 83) The early attacks on the resistant problem will have followed the paradigm rules quite closely. But with continuing resistance, more and more of the attacks upon it will have involved some minor or not so minor articulations of the paradigm ... Through this proliferation of divergent articulations, (...more and more frequently described as *ad hoc* adjustments), the rules of normal science become increasingly blurred. Though there is still a paradigm, few practitioners prove to be entirely agreed about what it is.

2.1 The frequentist paradigm

Bayesian theory began as the theory of inverse probability, proposed by Thomas Bayes (posthumously) in 1763. Remarkably, Bayes's theorem remains the centrepiece of the current Bayesian paradigm – its extreme simplicity, as the simple inversion of conditional probability, is unshakeable. What became shakeable was Bayes's postulate of the uniform prior for a parameter as a general principle, in the absence of information to the contrary. By itself this was not a sufficient criterion for opposition to the theory – there was no other.

Fisher's definition of the likelihood (though not by that name) in his first 1912 paper gave a firm basis for an alternative paradigm, though it was not till his 1922 paper that an explicit theory was put forward, with some of its properties, extended in his 1925 paper. The 1928-33 work on hypothesis testing by Neyman and Pearson extended the Fisher paradigm in a different direction, inconsistent with Fisher's own views, and the current version of the paradigm is sometimes called the Fisher-Neyman-Pearson, or generally the frequentist paradigm.

This was a truly revolutionary paradigm, in the Kuhnian sense. It provided a theoretical foundation for statistical analysis that did not depend on a prior specification for the model parameter, yet obtained exact or asymptotically exact results (in the sense of random variables with exact sampling distributions) for a very wide range of important probability and statistical models. The benefits of this paradigm have been enormous, and it is still taught as the standard paradigm in most statistics departments, though the Bayesian paradigm is taught increasingly as an alternative, or (much less commonly) as *the* paradigm.

In his 1934 paper, Neyman overthrew the then-current theory of purposive sampling and established the theory of random stratified cluster sampling, the foundation of the survey sampling paradigm. This led to a very substantial development of the paradigm.

Its principles were different from those of the frequentist school: though repeated sampling gave the probabilistic argument, as in the frequentist school, its basis was the repeated sampling of the *indicator variables* which define whether a population member was included in the sample or not. As there was no population model for the variable of interest, there was no likelihood in the parameters of the population. The principal question of interest was how to estimate the population mean or total of a variable taking into account the design of the survey.

To repeat Kuhn, "... [the paradigms,] for a time, provide model problems and solutions for a community of researchers." We argue below that both the frequentist and the survey sampling paradigms have stalled. They no longer provide the solutions to their problems – they have exhausted their scientific contributions – and need to be replaced by a Bayesian paradigm which does provide the solutions to its, and their, problems. This paradigm however has a major unsolved puzzle, or anomaly, which endangers its scientific contribution, and its practitioners have not been able to provide the solution to this critical problem.

We discuss these claims below, and give a resolution of the puzzle/anomaly which provides an extended Bayesian paradigm which includes the survey sampling field and is free of the anomaly.

3 Bayesian and survey sampling paradigms

In his review of the foundations of survey sampling, Smith (1976) concluded (p. 193):

> The basic question to ask is why should finite population inference be different from inferences made in the rest of statistics? I have yet to find a satisfactory answer. My view is that survey statisticians should accept their responsibility for providing stochastic models for finite populations in the same way as statisticians in the experimental sciences. These models can then be treated within the framework of conventional theories of inference. The problems with the Neyman approach then disappear to be replaced by disputes between frequentists, Bayesians, empirical Bayesians, fiducialists and so on. But at least these disputes are common to all braches of statistics and sample surveys are no longer seen as an outlier.

Survey sampling, in its recent formulation (see for example Särndal, Swensson and Wretman 1992), makes use of model-based estimators of population parameters, but without reliance on the correctness of the model, and therefore on the model likelihood. (The usage of the term "model" in "model-guided" or "model-assisted" analysis varies; in Särndal, Swensson and Wretman it is used for the *structural* part of a stochastic model, that is the mean and variance, and any independence, specification. They explicitly dismiss the complete specification of a stochastic model, which is mentioned in only four pages in their §14.5, where analysis based on this use of a model is called *model-dependent*.)

Without a formal model, inference is based on the repeated sampling distribution of the sample selection indicators, not of the population values themselves. From the viewpoint of model-based likelihood or Bayesian theory, this approach is unsatisfactory. The argument is clear if we construct the likelihood as the probability of *all* observed data. We denote by I the population index number, by Y_I^* the I-th population value of the response variable Y, and by Z_I the indicator variable which denotes selection into the sample ($Z_I = 1$) or

not ($Z_I = 0$). The data are *both* the sample selection indicators Z_I *and* the observed response variables y_i^* for the selected population members, so we need a population model for the Y_I^* as well.

The fundamental probability relation is

$$\begin{aligned} \Pr[Y_I^*, Z_I] &= \Pr[Z_I \mid Y_I^*] \Pr[Y_I^*] \\ &= \Pr[Y_I^* \mid Z_I] \Pr[Z_I]. \end{aligned}$$

where $\Pr[Z_I \mid Y_I^*]$ is the sample selection model for Z, which specifies how the selection probability of population member I depends on the value of the response Y_I^* for that member – and $\Pr[Y_I^*]$ is the population model for Y. For simple random sampling with rate π,

$$\Pr[Z_I \mid Y_I^*] = \Pr[Z_I] = \pi^{Z_I}(1-\pi)^{1-Z_I}.$$

Correspondingly,

$$\Pr[Y_I^* \mid Z_I] = \Pr[Y_I^*]$$

– the model for the selected population values is the same as that for the unselected values. So

$$\Pr[Y_I^*, Z_I] = \Pr[Y_I^*] \Pr[Z_I],$$

and the likelihood is

$$\begin{aligned} L &= \Pr[y_1^*, \ldots, y_n^*] \cdot \Pr[Z_1, ..., Z_N] \\ &= \Pr[y_1^*, \ldots, y_n^*] \cdot \frac{1}{\binom{N}{n}} \\ &= \Pr[y_1^*, \ldots, y_n^*] \cdot \frac{n}{N} \frac{n-1}{N-1} \cdots \frac{1}{N-n+1}. \end{aligned}$$

The last term in the selection probabilities is completely known from the design – it is just a constant. In inferential statements about the parameters based on *ratios* of likelihoods, these constant terms *cancel*. Thus, regardless of the kind of model we might have for the Y_I^*, any inference through likelihood ratios does not depend on the sample design, if this is *non-informative*, in the sense described above – that $\Pr[Z_I \mid Y_I^*] = \Pr[Z_I]$ – membership of the I-th population member in the sample does not depend on the value of the response Y_I^*.

In the survey sampling paradigm, this difficulty is countered by the difficulty of the dependence of model-based inference, whether frequentist or Bayesian, on the correctness of the model – if this is incorrect, the conclusions from the analysis could be wrong. Since every model is *by definition* wrong as it is a simplification, the risk of wrong conclusions from the model-based approach is inherent in the approach. Survey samplers are frequently working to a time line for analysis, so they cannot spend much time on investigating and validating suitable models for the response Y. Also, if the sample design is *informative*, likelihood-based inference becomes much more difficult because the form of dependence in $\Pr[Z_I \mid Y_I]$ needs to be specified and included in the likelihood.

These arguments and positions have been stable for many years. As Kuhn notes (pp. 94 and 109):

> When paradigms enter, as they must, into a debate about paradigm choice, their role is necessarily circular. Each group uses its own paradigm to argue in that paradigm's defense .... To the extent ... that two scientific schools disgree about what is a problem and what is a solution, they will invariably talk through each other when debating the relative merits of their respective paradigms. In the partially circular arguments that regularly result, each paradigm will be shown to satisfy more or less the criteria that it dictates for itself and to fall short of a few of those dictated by its opponent.
>
> [Wh]atever its force, the status of the circular argument is only that of persuasion. It cannot be made logically or even probabilistically compelling for those who refuse to step into the circle. The premises and values shared by the two parties to a debate over paradigms are not sufficiently extensive for that. ... [In] paradigm choice there is no standard higher than the assent of the relevant community.

Aitkin (2008), extended in Chapter 4 of Aitkin (2010), revisited the multinomial model of Hartley and Rao (1968) for the population frequencies p_J of a variable Y defined on the distinct support values Y_J in the population. (They can always be taken as a finite discrete set because of the finite measurement precision of any recorded variable). He extended the Bayesian analysis given by Hartley and Rao, Ericson (1968) and Hoadley (1969) with a non-informative Dirichlet prior, to deal with clustering and stratification with differential sampling rates, and gave a complex Labor Force Survey example from Valliant, Dorfman and Royall (2000). The analysis used very simple draws to provide the marginal posterior distributions of any functions of the population proportions, including means, variances, percentiles, regression coefficients, variance components and small-area random effects. (A full book-length exposition of this approach is in progress.)

This paper and its analysis led to some striking conclusions:

- the analysis is *impregnably robust* to arbitrary population structures because the multinomial distribution is an *exact representation* of the population (of *any* population) and is not a simplification or approximation: the analysis is thus (population) *distribution-free*;
- the analysis is *fully efficient* since it uses the multinomial likelihood for inferential statements, and despite the apparently heavy parametrization, gives comparable precision to standard model-and design-based approaches in simulation studies from simple populations;
- the analysis handles differential strata sampling rates by simple weighting of the posterior draws across strata for inference at the population aggregate level: no formal weighting procedure is needed in the analysis;
- the analysis provides precision measures of the population estimates *automatically*, from the variation in their posterior draws: no standard error calculations are needed;

- the posterior simulation procedure in many cases reduces to an iterative weighted maximum likelihood procedure with posterior Dirichlet weights;
- the role of the "guiding model" remains important as providing a definition of the population parameters of interest, with the fully nonparametric analysis providing protection against *both* distribution *and* variance misspecification.

Thus the principal criticism of the model-based approach by the survey sampling school is negated by the multinomial model. (This model, and the non-informative Dirichlet prior, have been proposed recently by Gutiérrez-Pena and Walker 2005 and Walker and Gutiérrez-Pena 2007 as a *general* model for statistical inference.) At the same time the advantages of the model-based approach are gained: given the specific population parameters of interest, the multinomial/Dirichlet model and prior provide a full Bayesian analysis.

This approach offers a *dramatic* reduction in the complexity of current survey analysis, which has to develop ad hoc methods for standard errors in complex designs by bootstrapping, jackknifing or the explicit design of the PSUs to allow for balanced half-sampling. These ad hoc procedures lie outside the paradigm, and have to be bootstrapped or jackknifed into it. The posterior distributions obtained by Dirichlet sampling provide the variation assessment *automatically*, and importantly, do not require the large-sample normality of the estimates for their calibration and interpretation.

4 Bayesian and frequentist paradigms

4.1 Priors

Prior distributions have always been a sticking-point for some frequentists, and the insistence by the subjective Bayesian sub-school on prior elicitation as an *essential* part of a principled Bayesian analysis has certainly not encouraged them to convert. The struggle of "objective Bayesians" (from Jeffreys onwards) to develop rules for the unique specification of non-informative priors has had some successes.

Frequentists still widely quote Fisher's criticism, from 1925 onwards, of the non-invariance of priors under monotone transformation, despite Geisser's (1988) note pointing out that if the parameter space for θ is finite and discrete, a monotone transformation $\phi = f(\theta)$ of a flat discrete prior on θ gives a flat discrete prior on ϕ – it is only the *spacing* on the ϕ scale which is transformed – the probabilities at the transformed values remain uniform. It is the transfer to a continuous parameter space which introduces the differential element $\frac{\partial \phi}{\partial \theta}$ which represents the change in measure; in the continuous space the different "packing density" of points has to be accounted for by an ordinate correction.

In the discrete finite space a uniform prior is a natural representation of equal support, and it is easy to see that for population means and proportions which are linear functions of the population values, the uniform prior is a natural

choice. Non-linear functions of the population values do not fit this pattern, but in the usual case where the information provided by the sample is large compared with that provided by the prior, the insensitivity of the posterior to the prior, and the aim of a *reference* analysis to determine "what the data say" make the uniform prior a reasonable choice, as it leaves the likelihood unchanged.

However an important difficulty in current Bayesian theory is a common confusion between aleatory and epistemic views of prior distributions. Probability models for observed data, or future data which might be observed, are representations of the outcomes of some random – "aleatory" – physical phenomenon. Epistemic probability represents our uncertainty, our degree of belief about uncertain events.

These have fundamentally important roles in Bayesian analysis, but these roles are different. Given enough observed data, a specific probability model can be assessed for goodness of fit to the data, and amended if it shows substantial discrepancies. This procedure is a standard part of statistical analysis, whether Bayesian or frequentist, and does not depend on a prior specification for the model parameters – it is the *shape* of the distribution which is being assessed, through a probability or QQ plot for example.

A prior distribution for the model *parameter* is not a stochastic model for a random act of a statistician, Nature, in drawing the one parameter value from some unknown probability distribution – it is simply a representation of our initial belief about the parameter values, prior to seeing the data. It is an essential part of Bayesian analysis, for without it we have only the likelihood, but its specification is not a critical model assumption for the Bayesian theory, as noted above: it does not have the same status as the probability model for the observed data. In particular, *the prior cannot be invalidated by the data*, unless it is very informative relative to the likelihood, and very different from it. This would imply that the data came from a different model family from the one implied by the highly informative prior. (If we *observed* a sample of one observation from a known parametric distribution, how could the single observation invalidate the distribution? What if the observation is drawn but not observed?)

This confusion is sometimes extended even further, to the model specification *for the data*, which is also treated by some Bayesians as an unknown infinite-dimensional parameter F requiring a prior, which is "got right" by embedding it in a nonparametric family like the Polya tree or Dirichlet process, or a mixture of Dirichlet processes. Draper asked, in the discussion of Walker et al (1999):

> Suppose that you take as your prior on F ... $p(F)$ = point mass on $N(\mu, \sigma^2)$ (with a prior on μ and σ) ... Now the data arrive and are strongly bimodal. What do you do? If you retain your original prior, then $p(\text{bimodal}|\text{data}) = 0$, which may be silly, but if you go back and change your prior (naively) you are cheating (using the data twice), and you risk poor calibration....

This difficulty is created by ignoring the finite measurement precision, which when recognised provides the finite-dimensional multinomial as the *always true*

model, with the Dirichlet prior on its parameters as described above. No larger "prior" family of arbitrary distributions is needed.

So the emphasis in many Bayesian discussions on *getting the prior right* is misplaced, apart from the logical contradiction of this argument with that of the subjective Bayesian sub-school, that it is critical for the prior to reasonably represent one's true belief, determined by elicitation if necessary. So one's prior beliefs cannot be recalibrated by getting them right from the data – that is the role of the posterior distribution. The example mentioned above, of prior and likelihood being inconsistent, can occur only with a highly informative prior. Even then to re-jig the prior to make it conform to the data likelihood would *really* be using the data twice. This important point is raised again in the following section.

However, we should not overstate the importance of these differences of view about priors; as Kuhn notes (p.44):

> [Scientists] can ... agree in their *identification* of a paradigm without agreeing on, or even attempting to produce, a full *interpretation* or *rationalization* of it. Lack of a standard interpretation or of an agreed reduction to rules will not prevent a paradigm from guiding research. (author's emphasis)

4.2 The Bayes factor anomaly

The major objection which frequentists (and others) can raise against the Bayesian paradigm as it is currently implemented is the Bayes factor *anomaly*, to use Kuhn's expression (pp. 5-6):

> Sometimes a normal problem, one that ought to be solved by known rules and procedures, resists the reiterated onslaught of the ablest members of the groups within whose competence it falls.... revealing an anomaly that cannot, despite repeated effort, be aligned with professional expectation.

The anomaly arises in the current Bayesian approach to testing point null hypotheses. In the frequentist theory, point null hypothesis testing about a parameter and confidence interval estimation for the parameter are consistent procedures – they are two sides of the same coin. In the current Bayesian theory, testing a point null hypothesis about a parameter through the Bayes factor – a form of generalized likelihood ratio – gives results which may be inconsistent with the credible interval for the parameter: the null value may lie well outside the credible interval, yet the Bayes factor may strongly accept the null value, or find no convincing evidence against it. So Bayes factors and credible intervals are two sides of different coins – their conclusions are incommensurate, as Kass and Raftery (1995, pp. 781-2) point out:

> In frequentist theory, estimation and testing are complementary, but in the Bayesian approach, the problems are completely different....

> It may happen that conclusions based on estimation [posterior distribution] seem to contradict those from a Bayes factor. In this case the data seem unlikely under H_0, but if the Bayes factor turns out to be *in favor of* H_0, then the data are *even more unlikely* under H_1 than they would have been under H_0. [authors' emphasis].

This creates a major difficulty for Bayesian statisticians, since they have in principle to decide whether their problem is one of inference about a model parameter through a credible interval, or whether it is a comparison of a null parameter value model with a general model. Having made this decision, the Bayesian presumably should not look at the other possible analysis – (s)he is not a frequentist.

(A further difficulty, which is not commonly mentioned, is that for regular models with large samples, the $100\alpha\%$ credible interval for the parameter agrees very well with the frequentist $100\alpha\%$ confidence interval, while the Bayes factor conclusion may differ substantially from the frequentist test conclusion.)

How do Bayesians deal with this anomaly? Kuhn describes the framework (p. 37):

> ... [O]ne of the things a scientific community acquires with a paradigm is a criterion for choosing problems that, while the paradigm can be taken for granted, can be assumed to have solutions. To a great extent these are the only problems that the community will admit as scientific or encourage its members to undertake. Other problems, including many that had previously been standard, are rejected as metaphysical, as the concern of another discipline, or sometimes just too problematic to be worth the time.

The Bayesian response is multi-dimensional:

- Adjust the prior: "... to avoid this difficulty, priors on parameters being tested [under the null hypothesis] generally must be proper *and not have too big a spread* ..." (Kass and Raftery p. 782, emphasis added – how big is too big?).
- It is the frequentist theory that is wrong: the P-value from the hypothesis test overstates the strength of evidence against the null model.
- The test of a *precise* null hypothesis is pointless, as we already know the hypothesis is false.
- Problems which are of scientific interest involve well-specified models in which we want to know the parameter ranges, we do not want to compare this model with another model or with a specialised form of this model.
- Model-checking can be done by posterior predictive checks, which compare features of the observed data with those of simulated data sets from the posterior predictive distribution.

These dismissals of hypothesis testing, or model comparison, do not fit easily with the major advantage claimed over the frequentist theory, of being able to compare non-nested models in the same way as nested models, which the frequentist theory cannot.

These responses are reflected in Bayesian textbook treatments of the two-sample and multi-sample normal mean problems, for which the t-test and analysis of variance are the standard tools of the frequentist theory. Almost all these textbooks are silent on the two-sample problem, except for the posterior distribution of the mean difference; there may be references to the overstatement of evidence by the P-value, clearly demonstrated by Berger and Delampady (1987) and Berger and Morters (1991) for example. There is no Bayesian equivalent of the t-test mentioned.

(There is a good reason for this – there was no published Bayesian version of the t-test until 2005. Now there are four: Aitkin, Boys and Chadwick 2005; Gönen, Johnson, Lu and Westfall 2005; Rouder, Speckman, Sun, Morey, and Iverson 2009; Wetzels, Raaijmakers, Jakab and Wagenmakers 2009.) Analysis of variance receives little attention, since this cannot easily be expressed in terms of the posterior distributions of mean differences.

Some textbooks give a full treatment of the Bayes factor approach, but have to ignore, or struggle with, its fundamental difficulty, discussed below. Others dismiss it on the scientific irrelevance grounds mentioned above. Gelman, Carlin, Stern and Rubin (2004) write:

> ... Bayes factors are rarely relevant in our approach to Bayesian statistics... (p. 192)
>
> ... this book has little role for the non-Bayesian concept of hypothesis tests, especially where these relate to point null hypotheses of the form $\theta = \theta_0$ .... most of the difficulties in interpreting hypothesis tests arise from the artificial dichotomy that is required between $\theta = \theta_0$ and $\theta \neq \theta_0$. (p. 250)

The key to the anomaly is the integration of the likelihood over the prior, and we discuss this at length, as it can severely affect the interpretation of Markov chain Monte Carlo (MCMC) analyses, a central contribution of the Bayesian school.

4.3 The integrated likelihood

The difficulty with point null hypothesis testing comes from the convention in current Bayesian theorey that when different models with unspecified parameters are being compared, this is to be done by integrating the likelihood, given these unknown parameters, with respect to the prior distribution of the parameters. The argument justifying this approach is taken from standard distribution theory: that a marginal distribution of one variable in a multivariate joint distribution is obtained by integrating out the other variables. This standard argument is applied to the product of likelihood and prior, with the likelihood

as the conditional distribution of the data given the parameter, and the prior distribution of the parameter, defining the joint distribution of data and parameter. The marginal distribution of the data is then obtained by integrating out the parameter, giving the integrated likelihood, often confusingly called the *marginal likelihood*; the same term is also commonly used for the likelihood from a two-level model integrated over the random effects.

The problem with this convention is immediate: in applying Bayes's theorem, the prior reflects our uncertainty about the parameter, often in a diffuse way with a flat prior. With an infinite parameter space the flat prior is improper, but this does not affect the posterior if the likelihood is informative: the posterior is the scaled likelihood.

However integrating the likelihood is a completely different matter: the prior cannot be improper, or the integrated likelihood will be undefined. So it must be proper, depending in general on (hyper-) parameters, which appear explicitly in the integrated likelihood, and determine its value. This effect does not disappear with increasing sample size; on the contrary, it is accentuated: as the likelihood becomes more precise with increasing sample size and the prior stays fixed, the integrated likelihood tends to zero.

The effect of this is that if this model with some unspecified parameters is to be compared with a completely specified model with no unknown parameters, the Bayes factor – the ratio of the likehood for the fully specified model to the integrated likelihood – will tend to infinity, giving heavy weight to the fully specified model, *whatever it is*. A consequence of these difficulties is that *sensitivity analyses* are essential (Kass and Raftery p. 782), in which the prior is varied over some appropriate range to determine the effect on the integrated likelihood. This has to be carried out for each competing model, and if both integrated likelihoods are affected by these variations, it may be difficult to draw a soundly-based conclusion. The definition of the "appropriate range" again causes difficulty.

Aitkin (2010 pp. 47-52) gives a long discussion of these difficulties, first pointed out by Bartlett (1957) in correcting an error in Lindley (1957). Here we simply note that the prior distribution is being given the status of a *data model* as in a two-level model, with the corresponding requirement of *getting the prior right*, a phrase which recurs in the insistence of some Bayesians that the "model" is the data model *and* the prior "model", which have equal status. As noted above, this conflicts with the view that the prior should represent the analyst's true opinion about the parameter, and is not to be set to achieve some data-based end.

4.4 MCMC and tuning the priors to the data

The serious consequences of this approach have been illustrated in several discussions of the recession velocity "galaxy" data set of Roeder (1990). Aitkin (2001) compared several Bayesian analyses of this data set. These analyses diverged widely in their conclusions about the number of components in the fitted normal mixture distribution needed to represent the velocity data. He

asked rhetorically why these analyses were so diverse, and concluded:

> The complexity of the prior structures needed for Bayesian analysis, the obscurity of their interaction with the likelihood, and the widely different conclusions they lead to over different specifications, leave the user completely unclear about 'what the data say' from a Bayesian point of view about the number of components in the mixture. For this mixture parameter at least, the Bayesian approach seems unsatisfactory in its present state of development.

Bayesians who were asked this question directly have generally responded, as did some of the authors of these papers: "The results are different, because of the different priors used." This restatement of the problem does not throw light on its solution. Aitkin (2011) revisited the issue and concluded that the diversity is caused, not specifically by the different priors used, but by the settings of the hyper-parameters of these priors. In the galaxy data set these settings lead to positive posterior probability being assigned to unbelievable numbers of components in several of the analyses.

The issue is quite straightforward to understand. Most of the analyses used some form of MCMC with normal priors on the component means, inverse gamma priors on the variances, and a Dirichlet prior on the component proportions. Each number of components K was treated separately and an integrated likelihood obtained for it; the integrated likelihoods were converted to posterior component probabilities. (Several analyses used reversible jump MCMC in which K was included in the parameter space, and the posterior for K was obtained directly.)

Gibbs sampling is straightforward to implement with the component membership as a latent variable. However, random draws from proper but diffuse priors may leave the chain stuck in an area of flat likelihood far from the maximum, and to prevent this, informative priors are used which match – are *tuned* to – the data so that initial draws from these priors will avoid regions of flat likelihood and improve the rate of convergence.

As a method for speeding convergence, this is unexceptionable. We note for later reference that convergence could be *very much* accelerated by initial draws from a "prior" implied by the maximum likelihood estimates and their covariance matrix – convergence would be almost immediate! This seems absurd – at convergence the likelihood would then be effectively integrated with respect to the posterior! But the same problem afflicts the integration with the informative and tuned data-based priors to provide a single-number integrated likelihood for comparison across the number K of components. As is immediately obvious from the previous discussion, the integrated likelihoods will be explicit functions of the hyper-parameters in the various priors. Different settings of these hyperparameters, even with the same priors and MCMC structure, by different analysts will thus give different integrated likelihoods. With different priors, this difference in conclusions will be exacerbated.

Table 1 from Aitkin (2011), shows the posteriors from different analyses, scaled to reflect a common uniform prior on K (the authors' initials are decoded

in Aitkin 2001 and 2011). Tail probabilities marked by ? cannot be calculated from the published data, but are all decreasing with increasing K. (Probabilities given in the table are scaled to sum to 1, ignoring the unknown tail values.) The remarkable differences in posteriors are due to the settings of the hyper-parameters in the common priors, and to other prior differences.

K	3	4	5	6	7	8	9	10	11	12	13
EW		.01	.03	.07	.13	.18	.30	.28	?	?	?
PS				.00	.10	.21	.43	.26	?	?	?
S	.10	.25	.35	.29	?	?	?	?	?	?	?
RW	>.999	<.001	-	-	-	-	-	-	-	-	
RG	.06	.13	.18	.20	.16	.11	.07	.04	.02	.01	.01

Table 1: Posterior distributions for K (flat prior)

This difficulty is endemic in the comparison of models by integrated likelihoods (a second example is given below in §6.1), and Aitkin (2011) showed that an alternative approach using the full posterior distributions of the likelihoods under each model (described below) does not suffer from this difficulty, and leads to a reasonable conclusion about the number of components.

4.5 Frequentist difficulties

The frequentist difficulties are familiar; for our purposes it is sufficient to point to the major successes of MCMC in complex unbalanced crossed and nested multilevel GLMMS, and the widespread adoption of multiple imputation, with its steady development towards a fully Bayesian analysis with incomplete data.

While maximum likelihood analyses are available for some of these structures, standard errors for the parameter estimates in the ML approach become decreasingly reliable with increasing complexity, and better non-Bayesian precision expressions are almost impossible to obtain.

On the foundational side, the arguments over the reference set for conditional analyses, and those over conditional versus unconditional analyses remain unresolved, and can lead to a wide variety of P-values, as in the ECMO example (Bartlett et al 1985), where the adaptive randomization used to assign babies to treatments led to a remarkable table of outcomes (Table 2):

Table 2: Babies surviving or dying under CMT and ECMO

	Response		
Treat	Survived	Died	Total
CMT	0	1	1
ECMO	11	0	11
TOTAL	11	1	12

Many P-values were given for this table (Ware 1989, Begg 1990) by different conditional and unconditional frequentist arguments: from 0.001 to 0.62.

It is fair to say that the frequentist paradigm is coming to the end of its useful life; one sign of a dying paradigm is the proliferation of new "flexible" methods untrammelled by the paradigm: regression trees, with their recipes for growing and pruning, and the grandiose claims once made for neural networks, now made for support vector machines.

This is not a criticism of the frequentist paradigm – it has made a profound contribution to the development of statistical theory and data analysis – but this contribution has nearly stalled, and to continue to develop complex analyses, in a much simpler theoretical framework, requires a fully Bayesian approach. The outstanding problem is how to resolve the Bayesian anomaly.

5 Resolving the anomaly

The foundation of the resolution was laid by Dempster (1974) for simple null and composite alternative hypotheses in a conference paper, subsequently published formally (Dempster 1997) together with Aitkin's extension of it to composite null hypotheses (Aitkin 1997). A book-length exposition of the approach is in Aitkin (2010).

The idea is simple, and completely within the Bayesian paradigm. A great strength of this paradigm over the frequentist paradigm is that, given the model $p(y \mid \theta)$, data $\mathbf{y}$, likelihood $L(\theta)$ and prior $\pi(\theta)$, the posterior distribution of *any* function of the parameters and the data, $g(\theta, \mathbf{y})$ say, follows immediately by standard distribution theory, and where this is difficult it can be obtained practically by mapping random draws $\theta^{[m]}$ from the posterior of θ into random draws $g^{[m]} = g(\theta^{[m]}, \mathbf{y})$ from the posterior distribution of g. This is particularly valuable for non-linear functions of the model parameters, for which the frequentist theory has to apply approximate linearisation through Taylor expansions to obtain approximate standard errors, from the already approximate standard errors from the information matrix.

Dempster's 1974 innovation was to take g as the likelihood, and obtain the posterior distribution of the likelihood. In practice it is much simpler to use the posterior distribution of the deviance (-2 log likelihood), from which that of the likelihood may be directly obtained. For two non-nested models which we want to compare, this process is repeated using independent draws from each posterior, which are then randomly paired to generate the posterior distribution of the likelihood ratio between the models.

With many models, we plot the cumulative distributions of the deviances for each model to determine whether the distributions are *stochastically ordered*, so that a random draw from the Model 1 likelihood is stochastically larger than a random draw from the Model 2 likelihood. If so, then Model 1 is preferred, and the strength of preference is assessed from the proportion of draws in which the likelihood ratio for Model 1 to Model 2 is greater than 1 (or another value like 3 or 10). This calibration is the same as that used for Bayes factors or likelihood ratios between fully specified models.

This approach is illustrated for the galaxy example in Figure 1, based on

10,000 draws from the posterior for each number of components, using diffuse normal, inverse gamma and Dirichlet priors for the means, variances and proportions respectively. The draws were kindly provided by Christian Robert, and are used in the paper by Celeux et al (2006) on the properties of the DIC. Aitkin (2010 pp. 211-221) discusses this example at length, and the figure below is used in Aitkin (2011) also.

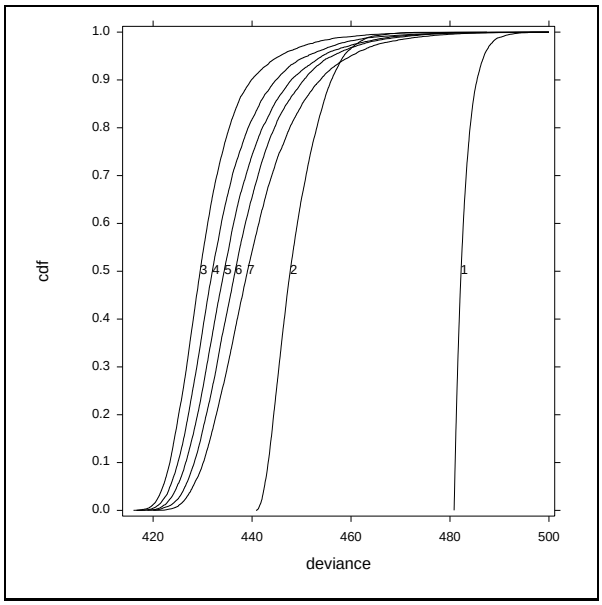

Figure 1: Deviances for 1-7 components

The interpretation of this figure can be simply summarised:

- The deviance distribution for $K = 2$ greatly improves on that for the single normal.
- The improvement continues for $K = 3$.
- As the number of components increases beyond three the deviance distributions move steadily to the right, to larger values (lower likelihoods).
- They also become more *diffuse*, with increasing slope.

So the deviance distribution for $K = 3$ is stochastically smaller than any of the others – this model is the most preferred, but that for $K = 4$ is not much less preferred. A simple parallel to the posterior model probabilities can be obtained by using the *median* deviances as "representative" values from the deviance distributions, and transferring these to likelihoods and then model probabilities, giving Table 3:

K	1	2	3	4	5	6	7
dev	482.2	448.9	431.0	433.9	436.2	438.3	440.8
prob	0	0	0.745	0.175	0.055	0.019	0.006

Table 3: Posterior median deviances and model probabilities

Three and four components are well-supported, but five or more have probability only 0.08. This contrasts very sharply with the Bayesian analyses reported in Table 1.

An important indirect contribution of this approach is that it frees MCMC from the constraint of tuning priors and its effect on the integrated likelihood. Since this approach does not use integration over the full parameter space, the chain can be initialised, and greatly accelerated, with draws from a normal distribution with mean the MLEs and covariance matrix the inverse information matrix, suitably expanded to allow for the greater variation in the posterior distribution. This gives maximum likelihood estimation, when it can be applied, an important role in MCMC.

For nested models which have the same parameter space, it is the likelihood *ratio* between the models (in practice the deviance difference between them) whose posterior distribution is *directly* simulated, as in Dempster's original binomial null hypothesis example. This gives a more precise result, as the posterior distribution of only the alternative model parameters is needed for that of the likelihood ratio.

It also gives in general, in nested models with normal likelihoods and flat priors, Dempster's remarkable result that the P-value from the frequentist likelihood ratio test is numerically equal to the posterior probability that the likelihood ratio, of null to alternative model, is greater than 1. So a P-value of 0.05 means only that the posterior probability that the null hypothesis has lower likelihood than the alternative is 0.95. While 0.95 is an impressive probability, a likelihood ratio of only less than 1 is not compelling evidence against the null hypothesis – we would want a likelihood ratio of something like 10 in favour of the alternative, for a probability of 0.95 *of this event* to be persuasive evidence against the null. So, independently of the demonstrations by Berger and others, the posterior likelihood approach shows that the P-value is indeed an overstatement of strength of evidence against the null hypothesis, and it also can be used to give a re-calibration of the P-value (Aitkin 1997) to provide a more reliable measure of strength of evidence.

Aitkin (2010) applies the posterior likelihood approach across a wide range of models, giving in particular new Bayesian versions of the t-test (first given by Aitkin, Boys and Chadwick 2005) and other standard frequentist procedures. A particular strength of this approach is that conclusions from the posterior likelihood model comparison or hypothesis test are consistent with the posterior distribution of the model parameter being tested; that is, this approach does not suffer from the Bayes factor anomaly.

In addition, the approach provides Bayesian alternatives to the common

"distribution-free " tests like the Wilcoxon-Mann-Whitney; these have the same advantage as in the survey sampling application, of a full model specification and a multinomial likelihood analysis.

A further benefit of the approach is that it provides a *general* procedure for assessing the probability model assumption, by comparing the specific parametric model with the nonparametric multinomial/Dirichlet model described earlier. Simulation studies reported in Aitkin (2010, pp. 188-194) show that for single sample problems the correct generating model for the data can be identified consistently from a candidate list as the sample size increases, though discrimination (for example between the lognormal and gamma) may require large samples.

Thus this approach removes the anomaly, and provides a consistent Bayesian paradigm for parameter inference and model comparisons which requires only standard non-informative priors, while at the same time extending the scope of Bayesian model comparison procedures into model validation and diagnostics.

6 Discussion

6.1 The need for a new paradigm

Bayes's theorem is the central plank of Bayesian theory. Its remarkable simplicity, and the immediate provision of the posterior distribution from the likelihood and the prior, make it an outstanding candidate for a paradigm. The rapid development of MCMC methods since 1990 shows how powerful the computational Bayesian approach is, relative to the severe difficulties of the frequentist school in dealing with the very complex models which have become tractable by MCMC methods. The Bayesian theory *was* the (only!) paradigm for a century and a half, and was replaced by the frequentist theory, but as Kuhn notes, that is not a barrier:

> (p. 108) [Paradigm] change [in the explanation of gravity] has been reversed, and could be again.
>
> (p. 76) So long as the tools a paradigm provides continue to prove capable of solving the problems it defines, science moves fastest and penetrates most deeply through confident employment of those tools. The reason is clear. As in manufacturing so in science – retooling is an extravagance to be reserved for the occasion that demands it. The significance of crises is the indication they provide that an occasion for retooling has arrived.

Why is a change needed in the current paradigm? My concern is twofold:

- The continuation of multiple inconsistent paradigms is damaging to the profession, and is holding back changes which could dramatically improve analyses, especially in the survey sampling school.

- The current Bayesian paradigm is unable to deal effectively with model comparisons, and is in serious danger of coming to misleading conclusions from the use of Bayes factors.

In support of the first point, I have already described the benefits to the survey sampling school of adopting the multinomial/Dirichlet approach. On damage to the profession, I have personally been involved as an expert witness in a court case (Aitkin 1992) in which statisticians advised both the prosecution and defense on the statistical interpretation of evidence for a change in the standard risk of brain damage from MMR vaccination. The statisticians presented frequentist and Bayesian analyses from which the conclusions were inconsistent. The three-judge panel later criticised the statisticians for their inability to agree on the meaning of the evidence – the statisticians contributed nothing to the evaluation of the case. We cannot be "expert" witnesses in the presence of conflicting paradigms.

One might argue that the statisticians should have met before the hearing and thrashed out their differences. While this would be beneficial, and could be workable for agnostic statisticians, Kuhn's comments about the arguments between paradigms are relevant – the statisticians' paradigms may prevent them from coming to an agreement.

On the second point, the galaxy data example leads to no consistent conclusions across the Bayes factor approaches discussed above, so there is little danger of misleading conclusions from them, rather, there is *no* conclusion. For a specific example of misleading conclusions, we refer to the recent papers by Lee (2004) and Liu and Aitkin (2008) in the Journal of Mathematical Psychology.

These papers discuss a data set on memory recall and its decay over time. The data set is made up of 15 annual re-tests of subjects for television programme recall, aggregated across subjects and modelled by five different normal distribution generalized linear models, with response linear regressions on time and different link functions, different transformations of the time scale, and a fixed standard deviation of 0.25. These are labelled as linear, exponential, hyperbolic, logarithmic and power.

Lee integrated the likelihoods for the five models over a common flat prior for the slope and intercept on the (0,2) square, which covered the areas of high likelihood under all the models; this was aimed to ensure "fairness" in the comparison. The preference ordering of models by their integrated likelihoods – hyperbolic, exponential, linear, power and logarithmic – was quite different from that by their maximized likelihoods – logarithmic, power, hyperbolic, exponential and linear – as Lee noted, explaining that the models with the highest maximized likelihoods had very concentrated contours and so had lower integrated likelihoods than those with diffuse contours.

Liu and Aitkin examined a range of other Bayesian and frequentist model comparison methods for the recall data. *All* except the Bayes factor approach gave the same ordering as that from the maximized likelihoods. Figure 2 shows the posterior deviance distributions with Lee's priors for the five models from 10,000 draws.

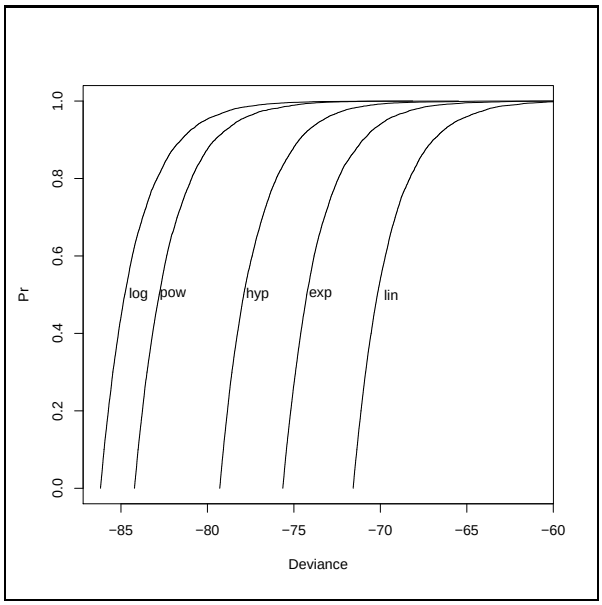

Figure 2: Posterior deviances, memory recall data

The distributions are stochastically ordered, and follow very closely their asymptotic form for regular models with flat priors (Aitkin 2010 p. 53):

$$-2\log L(\theta) \sim -2\log L(\hat{\theta}) + \chi^2_p,$$

where p is the model parameter dimension, 2 for all the models. Thus the cdfs of the models in this example are essentially exponential cdfs with the frequentist deviances as origins.

A critically important aspect of this result is that it depends on the data only through the maximized likelihood and the parameter dimension: there is no specific penalty for *model complexity*, other than the parameter dimension. For regular models with flat priors and the same dimensionality, the asymptotic ordering of the models is by their maximized likelihoods. This approach gives conclusions consistent with those of all the other Bayesian and frequentist criteria for model comparisons, except Bayes factors.

Lee gives a counter-argument supporting the models with diffuse likelihood contours (p. 314):

> It is also clear, however, that the good fits of the logarithmic and power functions occur at a very narrow set of parameter values, while the hyperbolic function fits reasonably well at a large number of parameterizations. These differences in complexity are even clearer in [his] Fig. 2, which shows contour plots of the same information. [These plots are also given in Liu and Aitkin and in Aitkin (2010).] It can be seen that the hyperbolic function fits across a broader range

> of parameter values than the exponential function, which in turn has a broader range than the linear, power and logarithmic functions. ...
>
> In terms of the balance between goodness-of-fit and complexity, these results indicate that the power and logarithmic models are capable of achieving better fit to the data, but are more complicated than the hyperbolic and exponential functions, because their fit is less robust across parametric variation.

This curious argument suggests that a model with a flat likelihood – unidentified – would be preferred to one with very narrow likelihood contours – very well identified – if it would integrate to a larger value, across the full parameter space over which it has constant likelihood, than the precise likelihood, which would be zero over much of the same parameter space.

This is precisely the anomaly caused by the integrated likelihood. Lee concludes (p.315)

> ... at the estimated level of data precision, the hyperbolic model constitutes the best balance between fit and inherent complexity, and is most strongly supported by [the] data.

We argue that these conclusions are unsound, and Lee's conviction is unsafe; what is of great concern is that this may be a widespread but latent phenomenon.

6.2 Reactions to the paradigm

Kuhn has discussed the reaction to new paradigms in detail:

> (pp. 77 and 81) ... what scientists never do when confronted by even severe and prolonged anomalies. ...[T]hey do not renounce the paradigm that has led them into crisis. They do not, that is, treat the anomalies as counter-instances... They will devise numerous articulations and *ad hoc* modifications of their theory in order to eliminate any apparent conflict. ...[E]ven a discrepancy unaccountably larger than that experienced in other applications of the theory need not draw any profound response. There are always some discrepancies. Even the most stubborn ones usually respond at last to normal practice. Very often scientists are willing to wait, particularly if there are many problems available in other parts of the field.
>
> (p. 7) The invention of ... new theories regularly, and appropriately, evokes the same response [of resistance] from some of the specialists on whose areas of special competence they impinge. For these men the new theory implies a change in the rules governing the prior practice of normal science. Inevitably, therefore, it reflects upon much scientific work they have already successfully completed. That is why a new theory, however special its range of applications, is seldom or never just an increment to what is already known. Its

> assimilation requires the reconstruction of prior theory and the re-evaluation of prior fact, an intrinsically revolutionary process that is seldom completed by a single man and never overnight.
>
> (p. 151) The transfer of allegiance from paradigm to paradigm is a conversion experience that cannot be forced. Lifelong resistance, particularly from those whose productive careers have committed them to an older tradition of normal science, is not a violation of scientific standards but an index of the nature of scientific research itself. The source of resistance is the assurance that the older paradigm will ultimately solve all its problems, that nature can be shoved into the box the paradigm provides. ... [This] assurance is what makes normal or puzzle-solving science possible.
>
> (p. 151 quoting Max Planck) A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.

It has been argued that the approach to Bayesian model comparisons described here is a new paradigm, and is not Bayesian. My view is that it not a new paradigm, but a new articulation of the Bayesian paradigm following the solution of a puzzle, not a true anomaly (in Kuhn's terms). However these are largely matters of definition.

There have been two reviews of Aitkin (2010) of quite different natures. In the discussion to be added, I will refer to these and any other reviews then available, in addition to responding to the discussants.

7 Acknowledgements

I have benefitted greatly from discussions with and comments from Charles Liu and Andrew Robinson which have improved an earlier draft of this paper.

8 References

Aitkin, M. (1992) Evidence and the posterior Bayes factor. *The Mathematical Scientist* **17**, 15-25.

Aitkin, M. (1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood (with Discussion). *Statistics and Computing* 7, 253-272.

Aitkin, M. (2001) Likelihood and Bayesian analysis of mixtures. *Statistical Modelling* 1, 287-304.

Aitkin, M. (2008) Applications of the Bayesian bootstrap in finite population inference. *Journal of Official Statistics* 24, 21-51.

Aitkin, M. (2010) *Statistical Inference: an Integrated Bayesian/Likelihood Approach*. Boca Raton, Chapman and Hall/CRC Press.

Aitkin, M. (2011) How many components in a finite mixture? in *Mixture Estimation and Applications*, eds. K. L. Mengersen, C. P. Robert and D. M. Titterington. New York, Wiley, pp. 277-292.

Aitkin, M. and Aitkin, I. (2011) *Statistical Modeling of the National Assessment of Educational Progress*. New York, Springer.

Aitkin, M., Boys, R.J. and Chadwick, T. (2005) Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing* 15, 217-230.

Barnett, V. (1999) *Comparative Statistical Inference* (3rd edn). Wiley: New York.

Bartlett, M.S. (1957) A comment on D.V. Lindley's statistical paradox. *Biometrika* 44, 533-534.

Bartlett, R.H., Roloff, D.W., Cornell, R.G., Andrews, A.F., Dillon, P.W. and Zwischenberger, J.B. (1985) Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* 76, 479-487.

Begg, Colin B. 1990. On inferences from Wei's biased coin design for clinical trials (with discussion). *Biometrika* 77, 467-484.

Berger, J.O. and Delampady, M. (1987) Testing precise hypotheses. *Statistical Science* 3, 317-352.

Berger, J.O. and Mortera, J. (1991) Interpreting the stars in precise hypothesis testing. *International Statistical Review* 59, 337-353.

Celeux, G., Forbes, F., Robert, C.P. and Titterington, D.M. (2006) Deviance information criteria for missing data models. *Bayesian Analysis* 1, 651-674.

Dempster, A.P. (1974) The direct use of likelihood in significance testing. In *Proceedings of the conference on foundational questions in statistical inference*, ed. Ole Barndorff-Nielsen, P. Blaesild and G. Sihon, 335-352.

Dempster, A.P. (1997) The direct use of likelihood in significance testing. *Statistics and Computing* 7, 247-252.

Ericson, W.A. (1969) Subjective Bayesian models in sampling finite populations (with Discussion). *Journal of the Royal Statistical Society B* 31, 195-233.

Fienberg, S.E. and Tanur, J.M. (1996) Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review* 64, 237-253.

Fisher, R.A. (1912) On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* 41, 155-160.

Fisher, R.A., (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A* 222, 309-368.

Fisher, R.A. (1925) Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* 22, 700-725.

Geisser, S. (1984) On prior distributions for binary trials. *The American Statistician* 38, 244-247.

Gelman, A., Carlin, J.B., Stern, H. and Rubin, D.B. (2004) *Bayesian Data Analysis* (2nd edn). Boca Raton: Chapman and Hall/CRC Press.

Gönen, M., W.O. Johnson, Y. Lu and P.H. Westfall. 2005. The Bayesian two-sample t test. *The American Statistician* 59, 252–257.

Groves, R.M. and Lyberg, L. (2010) Total survey error: past, present, and future. *Public Opinion Quarterly* 74, 849-879.

Gutiérrez-Pena, E. and Walker, S.G. (2005) Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review* 73, 309-330.

Hartley, H.O. and Rao, J.N.S. (1968) A new estimation theory for sample surveys. *Biometrika* 55, 547-557.

Hoadley, B. (1969) The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association* 64, 216-229.

Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association* 90, 773-795.

Kuhn, T.S. (1996) *The Structure of Scientific Revolutions* (3rd edn.) The University of Chicago Press.

Lee, M.D. (2004) A Bayesian analysis of retention functions. *Journal of Mathematical Psychology* 48, 310-321.

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

Liu, C. C. and Aitkin, M. (2008) Bayes factors: prior sensitivity and model generalizability. *Journal of Mathematical Psychology* 52, 362-375.

Neyman, J. and Pearson, E.S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 20A, 175-240.

Neyman, J. and Pearson, E.S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A* 231, 289-337.

Neyman, J. (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 97, 558-625.

Roeder, K. 1990. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association* 85, 617-624.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D. and Iverson, G. (2009) Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review* 16, 225-237.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model-Assisted Survey Sampling*. New York: Springer.

Smith, T.M.F. (1976) The foundations of survey sampling: a review (with Discussion). *Journal of the Royal Statistical Society A*, 139, 183-204.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000) *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

Vermunt, J.K. (2004) An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, 58, 220-233.

Walker, S.G. and Gutiérrez-Pena, E. (2007) Bayesian parametric inference in a nonparametric framework. *Test* 16, 188-197.

Ware, J.H. (1989) Investigating therapies of potentially great benefit: ECMO (with discussion). *Statistical Science* 4, 298-340.

Welsh, A.H. (1996) *Aspects of Statistical Inference*. New York, Wiley.

Wetzels R., Raaijmakers J.G., Jakab E., Wagenmakers E.J. (2009) How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian t-test. *Psychonomic Bulletin and Review* 16, 752-760.