Posterior Bayes Factors

Murray Aitkin

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 53, No. 1. (1991), pp. 111-142.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281991%2953%3A1%3C111%3APBF%3E2.0.CO%3B2-1

Journal of the Royal Statistical Society. Series B (Methodological) is currently published by Royal Statistical Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/rss.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Sun Mar 26 05:59:55 2006

J. R. Statist. Soc. B (1991)
53, *No.* 1, *pp.* 111–142

Posterior Bayes Factors

By MURRAY AITKIN†

Tel Aviv University, Israel

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, May 9th, 1990,* Professor D. V. Hinkley *in the Chair*]

SUMMARY

A general procedure for computing Bayes factors for the comparison of arbitrary models is described, based on the use of the posterior mean of the likelihood under each model rather than the usual prior mean. The use of the posterior mean has several advantages, including reduced sensitivity to variations in the prior and the avoidance of the Lindley paradox in testing point null hypotheses. The frequency properties of the new procedure are evaluated in standard examples, and a non-standard example is analysed to show the considerable differences possible between prior and posterior means of the likelihood. Several different justifications of the procedure are given, and a non-Bayesian direct likelihood interpretation is described.

Keywords: AKAIKE'S INFORMATION CRITERION; BAYES FACTORS; DIRECT LIKELIHOOD; LIKELIHOOD RATIO TEST; MODEL COMPARISONS; PENALIZED LIKELIHOOD

1. BAYES FACTORS

We are concerned in this paper with general methods for comparing different statistical models for the same data. We consider for simplicity the comparison of just two models M_1 and M_2 for data $\mathbf{y}$. Under model M_j, $\mathbf{y}$ has density or mass function $f_j(\mathbf{y}|\boldsymbol{\theta}_j)$ depending on a parameter $\boldsymbol{\theta}_j$ of dimension p_j. Given the data $\mathbf{y}$, the likelihood function under M_j is $L_j(\boldsymbol{\theta}_j)$. What evidence do the data provide about the two models?

1.1. *Neyman–Pearson Approach*

In the Neyman–Pearson framework, if the models are completely general then there is no optimal test for the hypothesis H_1: $\mathbf{y}$ has model M_1 against the alternative H_2: $\mathbf{y}$ has model M_2, unless $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are completely specified. Optimal tests exist under various restricted classes of models, the most commonly useful being when M_1 is a parametric submodel of M_2. In the general case the likelihood ratio test is commonly used: H_1 is rejected in favour of H_2 when $L_1(\hat{\boldsymbol{\theta}}_1)/L_2(\hat{\boldsymbol{\theta}}_2)$ is less than some constant c, chosen so that the test has level α under H_1.

The formulation of the likelihood ratio test requires an unambiguous specification of 'null' and 'alternative' models; this is clear in nested families of models but may be quite unclear in general models. The evaluation of c may involve intractable sampling distribution problems and simulation will generally be required. The size

†*Address for correspondence*: Department of Statistics, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel.

©1991 Royal Statistical Society 0035-9246/91/53111 $2.00

of the test will depend in general on $\boldsymbol{\theta}_1$, complicating further the evaluation of c.

The use of fixed conventional test sizes α like 0.05 or 0.01 leads to unreasonable tests in completely specified models. Let M_j be $y \sim N(\mu_j, \sigma^2)$ with $\sigma^2 = 1$ known and $\mu_1 = 0$, $\mu_2 = 1$. A sample of $n = 25$ gives $\bar{y} = 0.4$. If M_1 is taken as the null hypothesis and M_2 the alternative, the uniformly most powerful (UMP) test of size $\alpha = 0.05$ rejects M_1 in favour of M_2 when $z = n^{1/2}(\bar{y} - \mu_1)/\sigma > 1.64$, i.e. when $\bar{y} > 0.328$; the P-value of the observed $\bar{y}$ is 0.023. But the likelihood ratio $L(\mu_1)/L(\mu_2)$ is 12.18; model M_1 is much more strongly supported by the data than M_2, yet is rejected in favour of M_2. If M_1 is to be rejected in favour of M_2 only when the likelihood ratio is less than c, then the critical value of z is $n^{1/2}\delta/2 - \log c/n^{1/2}\delta$, which tends to infinity as $n \to \infty$, where $\delta = (\mu_2 - \mu_1)/\sigma$. Thus increasing sample size implies decreasing test size for a reasonable test. This simply restates the obvious fact that with increasing sample size both type I and type II error rates should tend to zero against a fixed alternative.

The same conclusion applies to other model comparisons problems in which the likelihood ratio test is UMP in a restricted class of models. For example, Dumonceaux *et al.* (1973) give (in their Table 3) percentage points of the likelihood ratio test of a log-normal against a two-parameter exponential distribution, for which the likelihood ratio test is UMP invariant. The tabulated points imply the rejection of the hypothesis with larger maximized likelihood for certain sample and test sizes.

It might be thought that this difficulty occurs only with simple alternative hypotheses, which are perhaps unrealistic, and that the role of the alternative is simply to indicate the direction of failure of the null hypothesis, rather than to be taken literally as an alternative scientific model (Cox and Hinkley (1974), pp. 88, 95). The real alternative, that is, is composite and unspecified. But for tests of simple hypotheses against composite alternatives, the use of conventional test sizes also causes difficulties. If in the example above μ_2 is unspecified in M_2, then the UMP unbiased test of size $\alpha = 0.05$ rejects M_1 in favour of M_2 when $|z| > 1.96$, i.e. when $|\bar{y}| > 0.392$; the P-value of the observed $\bar{y}$ is 0.046. But the maximized likelihood ratio $L(\mu_1)/L(\hat{\mu}_2)$ is $\exp(-\frac{1}{2}z^2) = 0.146$, which is not very small, and it clearly overstates the evidence against M_1, since the likelihood is maximized over M_2. For the maximized likelihood ratio to be less than c, the standardized mean $|z|$ has to exceed $(-2\log c)^{1/2}$, which is 2.45 for $c = 1/20$, corresponding to a test size of 0.014.

These and other difficulties with the interpretation of P-values have been intensively discussed in the Bayes framework by Berger and Sellke (1987), who give extensive references, and by Casella and Berger (1987).

1.2. *Bayes Approach*

In the Bayes framework we require the additional specification of prior densities $\pi_j(\boldsymbol{\theta}_j)$ and prior model probabilities π_j. Then the posterior odds on model 1 is

$$\frac{\pi(M_1|\mathbf{y})}{\pi(M_2|\mathbf{y})} = \frac{\bar{L}_1^{\mathrm{B}}}{\bar{L}_2^{\mathrm{B}}} \frac{\pi_1}{\pi_2}$$

where $B = \bar{L}_1^{\mathrm{B}}/\bar{L}_2^{\mathrm{B}}$ is the *Bayes factor* and

$$\bar{L}_j^{\mathrm{B}} = \int L_j(\boldsymbol{\theta}_j)\, \pi_j(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j$$

is the marginal probability of the data $\mathbf{y}$, or the 'prior mean' of the likelihood L_j.

The Bayes factor provides, in the Bayes framework, the sample 'weight of evidence' for model 1 over model 2. This weight of evidence depends on the priors $\pi_j(\boldsymbol{\theta}_j)$ and can be very sensitive to variations in the priors (see Section 5 for a simple example). If the prior $\pi_j(\boldsymbol{\theta}_j)$ accurately represents one's subjective belief about $\boldsymbol{\theta}_j$, then such sensitivity may not be a matter of concern, but many Bayesians and non-Bayesians feel more comfortable with Bayes conclusions which are insensitive to prior variations than with those which are very sensitive to such variations.

In the absence of a well-formulated subjective belief about $\boldsymbol{\theta}_j$ defining the prior $\pi_j(\boldsymbol{\theta}_j)$, one has to specify this prior in some reasonable way. The use of diffuse, vague, or Jeffreys priors as representations of prior ignorance is well established, but in the case of unbounded parameter spaces these can lead to the well-known Lindley paradox (Lindley, 1957) for a point null hypothesis.

1.3. *Lindley Paradox*

Let M_1 be $y \sim N(\mu_1, \sigma^2)$ with μ_1 specified, and M_2 be $y \sim N(\mu_2, \sigma^2)$ with μ_2 unspecified, σ^2 being known. Given vague prior information about μ_2, we specify a proper uniform prior for μ_2: $\pi(\mu_2) = 1/2C$ on $(-C, C)$, for C large. Then the Bayes factor for M_1 to M_2 from a sample of n observations is

$$B = 2C\frac{n^{1/2}}{\sigma}\phi(z) \Big/ \left\{\Phi\left(n^{1/2}\frac{\bar{y}+C}{\sigma}\right) - \Phi\left(n^{1/2}\frac{\bar{y}-C}{\sigma}\right)\right\},$$

where the denominator rapidly approaches unity as C increases. The value of B can be made arbitrarily large as $C \to \infty$ or as $n \to \infty$, whatever the fixed value of z. This is a consequence of the prior assigning increasing weight as $C \to \infty$ or $n \to \infty$ to values of μ_2 of negligible likelihood. The prior does not have to be uniform; any fixed proper prior will show the same effect as $n \to \infty$. The Bayes factor for M_1 to M_2 will tend to infinity.

These difficulties are well known and have often been discussed; see Shafer (1982) for an extensive recent discussion. Attempts to resolve these difficulties are of several different kinds. Since the paradox does not occur with models specifying *interval* hypotheses on the parameters (Casella and Berger, 1987), it has been argued that point null hypotheses are unreasonable—the problem is the hypothesis, not the analysis. Such an argument would limit Bayes analyses to interval hypotheses, thereby excluding most standard model comparisons problems, or at least requiring their reformulation.

It has been argued that the paradox occurs because of the inappropriate use of 'ignorance' priors. In the discussion of Shafer (1982), DeGroot said 'In summary, when diffuse prior distributions are used in Bayesian inference, they must be used with care. Although they can serve as convenient and useful approximations in some estimation problems, they are never appropriate for tests of significance. Under no circumstances should they be regarded as representing ignorance.'

This argument would limit Bayes analyses to informative priors, with the corresponding requirement to specify such priors as part of the analysis, and to assess the sensitivity of the conclusions to the choice of different informative priors. Such an approach is certainly possible, and has been implemented by Smith *et al.* (1985); it makes the Bayesian analysis of data quite a complex process.

Other approaches attempt to fix the prior specification by other arguments.

Spiegelhalter and Smith (1982) assigned a specific value to $2C$, the ordinate of the vague prior for μ_2, by the device of an 'imaginary training sample'. (They applied this approach to the more general case of unspecified location parameters under both M_1 and M_2.) Imagine that an additional data set (the 'training sample') is available which

(a) involves the smallest possible sample size permitting a comparison of M_1 and M_2 and
(b) provides maximum possible support for M_1.

Specifically, they assumed that the Bayes factor from this training sample is approximately unity. For the previous model comparison problem, one observation y^* is required which provides maximum support for M_1 and gives a Bayes factor of approximately unity:

$$1 \doteq \frac{2C}{\sigma}\phi\left(\frac{y^*-\mu_1}{\sigma}\right) \Big/ \left\{\Phi\left(\frac{y^*+C}{\sigma}\right) - \Phi\left(\frac{y^*-C}{\sigma}\right)\right\}.$$

The value $y^* = \mu_1$ gives maximum support to M_1, so

$$1 \doteq \frac{2C}{\sigma}\phi(0) \Big/ \left\{\Phi\left(\frac{\mu_1+C}{\sigma}\right) - \Phi\left(\frac{\mu_1-C}{\sigma}\right)\right\}$$

for which an approximate solution is, neglecting the denominator term,

$$2C/\sigma \doteq 1/\phi(0).$$

Then the Bayes factor for the comparison of M_1 and M_2 is

$$B = n^{1/2}\,\phi(z)/\phi(0) = n^{1/2}\exp(-\tfrac{1}{2}z^2),$$

where z is as in Section 1.1. This still increases with n whatever the value of z.

Smith and Spiegelhalter (1980) proposed the use of the prior $N(\mu_1, \sigma^2/n)$ for μ_2, in the context of local alternatives to the null hypothesis. This gives a Bayes factor for M_1 to M_2 of

$$B = 2^{1/2}\exp\left\{-\frac{n(\bar{y}-\mu_1)^2}{4\sigma^2}\right\} = 2^{1/2}\exp(-\tfrac{1}{4}z^2).$$

While this Bayes factor is not a function of n and therefore does not suffer from the Lindley paradox, it requires very large values of z for evidence against M_1: a value of B of 1/20, corresponding to a posterior probability of M_1 of 1/21 assuming uniform model priors, requires $|z| = 3.66$.

Smith and Spiegelhalter called this a 'local Bayes factor', since it gives increasing prior weight under M_2 to a local neighbourhood of M_1. While this may be appropriate in some problems, it is clear that as n increases the prior under M_2 becomes increasingly concentrated on the M_1 value, rather than on some value appropriate to the alternative model M_2. The approach leaves open the question of how to treat models symmetrically, without strong prior evidence in support of a local neighbourhood of one model, and how to compare non-nested models.

We leave aside here the logical status of the 'prior', depending as it does on the data through the sample size. Smith and Spiegelhalter (1980), p. 216, commented:

'The approach we adopt in examining an alternative prior specification may be viewed either as a genuine subjective Bayesian analysis, with respect to a particular form of prior belief, or simply as a formal analysis, intended as a theoretical *ad hoc* device for comparing $[M_1]$ with a "local" subset of the models contained within $[M_2]$'.

We now consider a variation on the Smith and Spiegelhalter approach which treats the models symmetrically. Since as n increases $\bar{y}$ approaches the true value of μ_2 under M_2 with variance σ^2/n, we propose the prior for μ_2 under M_2 to be $\mu_2 \sim N(\bar{y}, \sigma^2/n)$. This gives a Bayes factor of $2^{1/2}\exp(-\frac{1}{2}z^2)$, which also does not suffer from the Lindley paradox. It requires larger values of z as evidence against M_1 than does the ratio of maximized likelihoods, but smaller values than the Smith and Spiegelhalter factor: for a factor of 1/20, z must exceed 2.585, the 1% point of the normal distribution.

How can such a prior be justified objectively, beyond the kind of justification given by Smith and Spiegelhalter? We observe that the Lindley paradox occurs because values of μ_2 of negligible likelihood are assigned non-zero prior weight. In calculating the posterior distribution of μ_2, such an assignment does no harm, but in integrating the likelihood with respect to the prior to give the marginal probability of y, such an assignment reduces the average of the likelihood to zero as the prior weight increases on these values of μ_2.

If the prior is intended to be 'objective', rather than to represent one's subjective belief, why should this objective prior assign weight to values of μ_2 which are untenable given the data, thus reducing the probability of the observed data to zero? If the probability of the observed data goes to zero under the integrated model, this surely means that the *prior assignment* is untenable. From this viewpoint, a Bayes factor approaching infinity does not mean strong support for the null model M_1, but strong rejection of the diffuse prior model M_2. It seems more appropriate to average with respect to one's *posterior* weight for μ_2 having observed the data. This is the variation we propose: that the *posterior* mean, rather than the prior mean, of the likelihood should be used in comparing the two models. We now state the result formally.

2. POSTERIOR BAYES FACTOR

With the same model notation as in Section 1, define

$$\bar{L}_j^{\mathrm{A}} = \int L_j(\boldsymbol{\theta}_j)\, \pi_j(\boldsymbol{\theta}_j \mid \mathbf{y})\, \mathrm{d}\boldsymbol{\theta}_j$$

where $\pi_j(\boldsymbol{\theta}_j \mid \mathbf{y})$ is the posterior density of $\boldsymbol{\theta}_j$:

$$\pi(\boldsymbol{\theta}_j \mid \mathbf{y}) = L_j(\boldsymbol{\theta}_j)\, \pi_j(\boldsymbol{\theta}_j) \Big/ \int L_j(\boldsymbol{\theta}_j)\, \pi_j(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j.$$

Then $\bar{L}_j^{\mathrm{A}}$ is the *posterior mean* of the likelihood function $L_j(\boldsymbol{\theta}_j)$ of the data $\mathbf{y}$ and parameters $\boldsymbol{\theta}_j$. Equivalently,

$$\bar{L}_j^{\mathrm{A}} = \int L_j^2(\boldsymbol{\theta}_j)\, \pi_j(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j \Big/ \int L_j(\boldsymbol{\theta}_j)\, \pi_j(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j.$$

The ratio of posterior means $\bar{L}_1^{\mathrm{A}}/\bar{L}_2^{\mathrm{A}}$ will be called the *posterior Bayes factor* denoted by A (for average):

$$A = \bar{L}_1^{\mathrm{A}}/\bar{L}_2^{\mathrm{A}}.$$

It is easily seen that $L_j(\hat{\boldsymbol{\theta}}_j) \geqslant \bar{L}_j^{\mathrm{A}} \geqslant \bar{L}_j^{\mathrm{B}}$.

We propose the use of A in the same way as B, as a measure of the weight of sample evidence in favour of M_1 compared with M_2, with the calibration that values of A less than 1/20, 1/100 or 1/1000 constitute strong, very strong and overwhelming sample evidence against M_1 in favour of M_2.

It is immediately clear from the equivalent definition of $\bar{L}_j^{\mathrm{A}}$ that it does not depend on the ordinate of a uniform prior density, for the constant $1/2C$ cancels from the ratio of integrals of L_j^2 and L_j. We may formally relate $\bar{L}_j^{\mathrm{A}}$, $\bar{L}_j^{\mathrm{B}}$ and the maximized likelihood $L_j(\hat{\boldsymbol{\theta}}_j)$ through the indexed integral family

$$\bar{L}_j^{(k)} = \int L_j(\boldsymbol{\theta}_j)\, w_j^{(k)}(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j$$

where

$$w_j^{(k)}(\boldsymbol{\theta}_j) = L_j^k(\boldsymbol{\theta}_j)\,\pi_j(\boldsymbol{\theta}_j) \Big/ \int L_j^k(\boldsymbol{\theta}_j)\,\pi_j(\boldsymbol{\theta}_j)\,\mathrm{d}\boldsymbol{\theta}_j$$

so that

$$\bar{L}_j^{(k)} = \int L_j^{k+1}(\boldsymbol{\theta}_j)\,\pi_j(\boldsymbol{\theta}_j)\,\mathrm{d}\boldsymbol{\theta}_j \Big/ \int L_j^k(\boldsymbol{\theta}_j)\,\pi_j(\boldsymbol{\theta}_j)\,\mathrm{d}\boldsymbol{\theta}_j.$$

Formally, $w_j^k(\boldsymbol{\theta}_j)$ is the posterior density of $\boldsymbol{\theta}_j$ from k 'copies' or replicates of the data $\mathbf{y}$ and prior density $\pi_j(\boldsymbol{\theta}_j)$. As $k \to \infty$, $w_j^{(k)}(\boldsymbol{\theta}_j)$ approaches a spike at the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}_j$, and hence $\bar{L}_j^{(k)} \to L_j(\hat{\boldsymbol{\theta}}_j)$ as $k \to \infty$. The case $k = 0$ gives the usual prior mean, and the posterior mean is the case $k = 1$.

3. REPEATED SAMPLING PROPERTIES

The repeated sampling properties of the posterior Bayes factor are easily determined in the usual nested model comparisons problems, along the lines of the discussion in Section 2.2 of Smith and Spiegelhalter. Let M_1 be a regular submodel of M_2 and assume that the likelihoods L_1 and L_2 are normal in the parameters, i.e. that the sample size is large compared with the number of parameters, and that the prior $\pi(\boldsymbol{\theta}_j)$ is locally uniform in the neighbourhood of $\hat{\boldsymbol{\theta}}_j$. Then

$$L_j(\boldsymbol{\theta}_j) = L_j(\hat{\boldsymbol{\theta}}_j)\, \exp\{-\tfrac{1}{2}(\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j)'\, \mathbf{I}_j (\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j)\}$$

where $\mathbf{I}_j$ is the observed information for M_j. It is easily shown that

$$\bar{L}_j^{\mathrm{A}} = 2^{-p_j/2}\, L_j(\hat{\boldsymbol{\theta}}_j)$$

and therefore that the posterior Bayes factor for M_1 to M_2 is

$$A = 2^{\nu/2}\, L_1(\hat{\boldsymbol{\theta}}_1)/L_2(\hat{\boldsymbol{\theta}}_2),$$

with $\nu = p_2 - p_1$, a *penalized* version of the ratio of maximized likelihoods. Write $\lambda = -2\log\{L_1(\hat{\boldsymbol{\theta}}_1)/L_2(\hat{\boldsymbol{\theta}}_2)\}$ for the usual likelihood ratio test statistic ('deviance difference') which is distributed as χ^2_ν under the model M_1. Then

$$-2\log A = \lambda - \nu \log 2,$$

a particular case of the general class of penalized likelihood ratio test statistics discussed by Smith and Spiegelhalter. This class

$$\Lambda(m) = \lambda - m\nu$$

includes Akaike's information criterion (Akaike, 1973) ($m=2$), a local Bayes factor obtained by Smith and Spiegelhalter using a particular uniform prior density on $\boldsymbol{\theta}_j$ rather than the normal density in Section 1 ($m=3/2$), the generalized linear model proposal of Nelder and Wedderburn (1972) to compare the deviance to its degrees of freedom ($m=1$) and the direct use of the ratio of maximized likelihoods ($m=0$) discussed by Edwards (1972) in the context of 'support tests'. The local Bayes factor obtained by the normal prior density argument leads to a different criterion, with

$$-2\log B=\tfrac{1}{2}\lambda-\nu\log 2$$

and Smith and Spiegelhalter do not consider its properties.

The B information criterion (BIC) or Schwarz (1978) criterion with $m=\log n$ is a special case of the (prior) Bayes factor when the information in a proper normal prior is proportional (in sample size) to that in the sample (Smith and Spiegelhalter (1980), pp. 214–215). In the posterior mean of the likelihood, the information matrix $\mathbf{I}_j$ cancels, and so the posterior Bayes factor does not depend explicitly on n or other aspects of the design matrix.

The posterior Bayes factor corresponds to $m=\log 2=0.693$. Smith and Spiegelhalter remark that 'generally speaking, values of $m<1$ tend to favour complex models unduly'. This remark appears to be based only on the test sizes resulting from the use of the Bayes factors as formal tests in the usual way. In the approach described here, the penalty arises naturally from the object of specifying the value of the likelihood that would have been achieved had the parameters been known. It is not an *a priori* penalty against complex models. We reproduce in our Fig. 1 part of the information on test sizes in Fig. 1 of Smith and Spiegelhalter supplemented by corresponding information for the posterior Bayes factor.

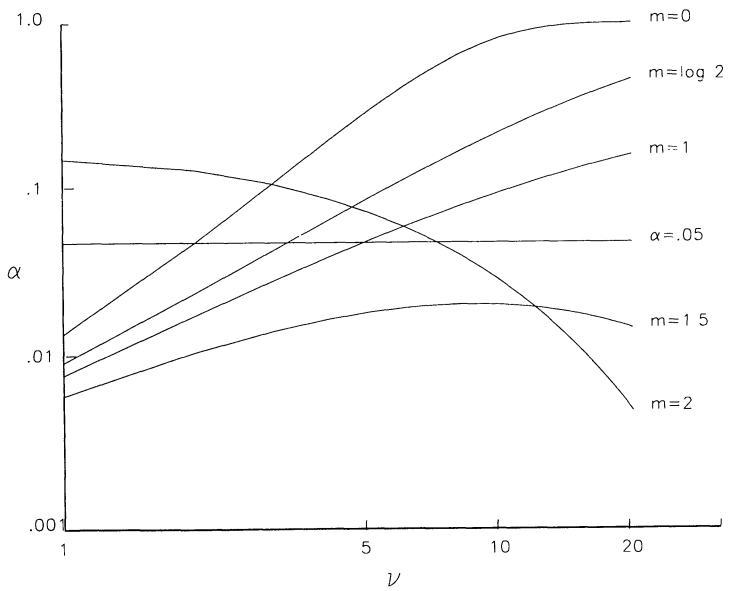

Fig. 1. Test sizes for penalized likelihood ratio test statistics ($c=0.05$)

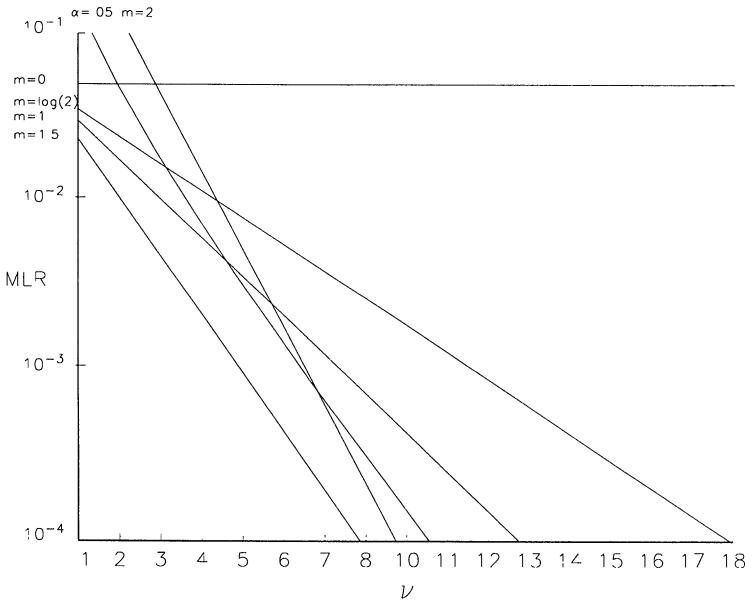

Fig. 2. Critical values of maximized likelihood ratios ($c=0.05$)

The standard use of Akaike's information criterion rejects M_1 in favour of M_2 when

$$\lambda - 2\nu > 0,$$

with equivalent test size

$$\alpha(2) = P(\chi^2_\nu > 2\nu).$$

We compare this test size with those denoted by $\alpha(m_c)$ resulting from comparisons of $\Lambda(m)$ for $m=1.5$, 1, $\log 2$ and 0 with $-2\log c$ for $c=0.05$, one of the values considered by Smith and Spiegelhalter. We choose this value because, for equal prior model probabilities, the posterior probability of M_1 is 1/21, sufficiently small to provide strong evidence against M_1. We also include the constant test size $\alpha=0.05$ for the usual use of the likelihood ratio test.

Our Fig. 1 shows the test size rising rapidly with the degrees of freedom ν to 1 for $m=0$, rising more slowly for $m=\log 2$ and $m=1$, being fairly stable for $m=1.5$ and decreasing rapidly for $m=2$. Fig. 2 shows the values on a logarithmic scale of the corresponding maximized likelihood ratios using the same 'critical values'.

The conventional likelihood ratio test of size α rejects M_1 when

$$L_1(\hat{\boldsymbol{\theta}}_1)/L_2(\hat{\boldsymbol{\theta}}_2) < \exp(-\tfrac{1}{2}\chi^2_{\nu,1-\alpha}).$$

The corresponding critical value of the maximized likelihood ratio is given in Table 1 for $\alpha=0.05$ and shown in Fig. 2. The critical value of the likelihood ratio decreases rapidly with ν and becomes extremely small for large ν. While maximization over $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ overstates the evidence against M_1, it is difficult to believe that this overstatement is so extreme that a maximized likelihood ratio of 10^{-6} for $\nu=20$ is still not convincing evidence against M_1. It appears that insisting on a fixed test size

TABLE 1

Dimension ν	*Critical value* $\exp(-\frac{1}{2}\chi^2_{\nu,0.95})$	*Dimension* ν	*Critical value* $\exp(-\frac{1}{2}\chi^2_{\nu,0.95})$
1	0.147	7	8.81×10^{-4}
2	0.050	8	4.29×10^{-4}
3	0.0201	10	1.06×10^{-4}
4	8.70×10^{-3}	20	1.51×10^{-7}
5	3.95×10^{-3}	30	3.13×10^{-10}
6	1.85×10^{-3}		

of 0.05, however many parameters there are in the model, leads to increasing conservatism with increasing degrees of freedom in the rejection of M_1 in favour of the better supported M_2.

Increasing the test size with the number of parameters is a common practice in multiple comparisons and multiple-testing problems, where it has long been recognized that conventional overall test sizes of 0.05 or less lead to increasingly conservative tests for individual effects or contrasts as the number of parameters increases. Overall test sizes as high as 0.50 have sometimes been used (Gabriel, 1964), and values like $1-(1-\alpha)^\nu$ have been used frequently in regression model simplification (Aitkin *et al.*, 1989).

Thus an increasing test size with increasing ν is not an *a priori* disadvantage of the posterior Bayes factor viewed as a formal test statistic. We now consider the advantages of the posterior Bayes factor in other model comparison problems.

4. NORMAL REGRESSION MODEL

Let model M_j be $\mathbf{y}\sim N_n(\mathbf{X}_j\boldsymbol{\beta}_j,\ \sigma^2\mathbf{I})$ with $\mathbf{X}_j$ of full rank p_j. Then omitting irrelevant constants

$$L_j(\boldsymbol{\beta}_j,\ \sigma)=\frac{1}{\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\left\{\mathrm{RSS}_j+(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)'\,\mathbf{X}_j'\mathbf{X}_j(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\right\}\right]$$

where RSS_j is the residual sum of squares for M_j. We take a diffuse prior for $\boldsymbol{\beta}_j$ and the improper prior σ^{r-1} for σ, where $r=0$ is the usual diffuse prior for $\log\sigma$. Then

$$\int L_j^k(\boldsymbol{\beta}_j,\ \sigma)\,\mathrm{d}\boldsymbol{\beta}_j\,\sigma^{r-1}\mathrm{d}\sigma=$$

$$2^{(nk-p_j-r-2)/2}k^{-(nk-r)/2}\,\Gamma\{(nk-p_j-r)/2\}|\mathbf{X}_j'\mathbf{X}_j|^{-1/2}\mathrm{RSS}_j^{-(nk-p_j-r)/2}(2\pi)^{p_j/2}c_j$$

for $k=1$, 2, and so

$$\bar{L}_j^{\mathrm{A}}=2^{-(n-r)/2}\,\frac{\Gamma\{(2n-p_j-r)/2\}}{\Gamma\{(n-p_j-r)/2\}}\,\mathrm{RSS}_j^{-n/2}$$

while

$$\bar{L}_j^{\mathrm{B}}=2^{(n-p_j-r-2)/2}\Gamma\{(n-p_j-r)/2\}|\mathbf{X}_j'\mathbf{X}_j|^{-1/2}\mathrm{RSS}_j^{-(n-p_j-r)/2}(2\pi)^{p_j/2}c_j,$$

where c_j is the ordinate of the diffuse prior. The (prior) Bayes factor depends in general on the scaling of both $\mathbf{y}$ and the explanatory variables in the model, as well as the ratio of the diffuse prior ordinates.

TABLE 2

r	-2	0	2
Penalty constant ($n=20$, $p_1=2$, $p_2=4$)	2.11	2.25	2.43

The posterior Bayes factor for M_1 to M_2 is

$$A=\bar{L}_1^{\mathrm{A}}/\bar{L}_2^{\mathrm{A}}=\frac{\Gamma\{(2n-p_1-r)/2\}}{\Gamma\{(n-p_1-r)/2\}}\frac{\Gamma\{(n-p_2-r)/2\}}{\Gamma\{(2n-p_2-r)/2\}}\left(\frac{\mathrm{RSS}_1}{\mathrm{RSS}_2}\right)^{-n/2}$$

which is invariant to scaling and is not subject to the Lindley paradox. It is again a penalized form of the ratio of maximized likelihoods. The penalty constant in A multiplying this ratio depends only weakly on the value of r, the 'hyperparameter' in the prior for σ. Table 2 gives the penalty constant for $n=20$, $p_1=2$, $p_2=4$ and $r=-2$, 0 and 2, representing diffuse priors for σ^{-2}, $\log\sigma$ and σ^2.

Thus even substantial variation in the prior distribution for σ has little effect on the posterior Bayes factor in reasonable-sized samples. If $p_1=p_2$, i.e. if we are comparing non-nested models with the same number of parameters, the penalty constant is unity whatever the value of r, and the posterior Bayes factor is simply the ratio of maximized likelihoods, which now has a direct interpretation as evidence for M_1 compared with M_2. This is an important benefit of the posterior Bayes factor, in allowing direct comparisons of non-nested models.

For nested models, the variation in the penalty constant with n, p_1 and p_2 serves the same function as the sampling distribution of the likelihood ratio test statistic—to adjust for the different numbers of parameters—without requiring this distribution. A detailed discussion of regression model choice using this approach will be presented elsewhere.

We conclude with a non-standard problem.

5. BINOMIAL SAMPLE SIZE

The binomial sample size problem has been discussed recently by Carroll and Lombard (1985), who gave historical background, and by Kahn (1987). Draper and Guttman (1971) gave a Bayes analysis (see also Raftery (1988)). A full discussion of the difficulties caused by the peculiar form of the likelihood function in this model is given by Aitkin and Stasinopoulos (1989).

We are given independent observations $s_1, \ldots, s_r$ from the binomial distribution $b(N, p)$ with both parameters unknown, and the problem is to draw conclusions about N, p being a nuisance parameter. The likelihood function is

$$L(N,p)=\prod_{i=1}^{r}\binom{N}{s_i}p^{s_i}(1-p)^{N-s_i}$$

$$=\left\{\prod_i\binom{N}{s_i}\right\}p^T(1-p)^{Nr-T}$$

where $T=\Sigma_1^r s_i$, and $0<p<1$, $N \geqslant \max_i s_i = N_L$. We illustrate with the sample of $r=5$ observations (16, 18, 22, 25, 27) taken from Olkin *et al.* (1981). The joint MLEs of p and N are (0.218, 99). For given N the MLE of p is

$$\hat{p}(N) = T/Nr = \bar{s}/N$$

and the profile likelihood in N is

$$L(N, \hat{p}(N)) = \left\{ \prod \binom{N}{s_i} \right\} \left(\frac{\bar{s}}{N}\right)^T \left(1 - \frac{\bar{s}}{N}\right)^{N^r - T}.$$

For the example, the profile likelihood is shown in Fig. 3. It rises rapidly from zero at $N=27$ to a poorly defined maximum value at $N=99$ and then declines very slowly, approaching an asymptote of $0.935L(\hat{N}, \hat{p})$ as $N \to \infty$.

A conditional likelihood approach is possible by conditioning on T, the sufficient statistic for p when N is known. The conditional likelihood is

$$C(N) = \left\{ \prod \binom{N}{s_i} \right\} \Big/ \binom{Nr}{T}$$

which is a conditional hypergeometric likelihood. The conditioning variable T has a distribution which also depends strongly on N, so there must be some loss of information about N in the conditioning.

The conditional likelihood also rises rapidly from zero but approaches its maximum as $N \to \infty$ and is flat for a large range of N (see Fig. 3).

Carroll and Lombard (1985) pointed out that $C(N)$ is effectively an integrated likelihood obtained by integrating out p from $L(N, p)$ with respect to the improper prior $1/p$. They considered the conjugate family of beta distributions for p and obtained point estimators of N as the maximizing values of the integrated likelihoods $\bar{L}(N)$ with respect to the beta distributions. For two proper priors (uniform and quadratic) they found that the point estimators of N have improved mean-square error relative to earlier estimators. Aitkin and Stasinopoulos (1989) show that for this example the two integrated likelihoods have well-defined maxima around 50–60 and approach zero as $N \to \infty$ (Fig. 3).

Raftery (1988) gave a Bayes analysis using independent priors, with p uniform and the prior for N proportional to N^{-1}. The posterior for N was then $N^{-1}\bar{L}(N)$.

Kahn (1987) noted that the upper tail behaviour of these integrated likelihoods is entirely predictable from the prior for p and does not depend on the data at all. If the prior is $\pi(p) = p^{a-1}(1-p)^{b-1}/B(a, b)$, the integrated likelihood $\int L(N, p)\pi(p)\,dp \to CN^{-a}$, where C is independent of N. For $a=0$ the tail of the integrated likelihood is constant, as already noted. Aitkin and Stasinopoulos (1989) show that the likelihood function $L(N, p)$ is extremely concentrated along the hyperbola $Np = \bar{s}$, and therefore the choice of prior distribution for p (if this is taken independently of N) has a critical effect on the integrated likelihood for N. Fig. 4 shows the two-parameter likelihood $L(N, p)$ for the example.

The difficulty with the Bayes approach with independent priors on p and N is evident from Fig. 4. If p is small, N must be large, and this is evident from prior consideration of the model, not just by inspection of the likelihood. Aitkin and Stasinopoulos (1989) show that reparameterizing the nuisance parameter to $\psi = Np$ gives a likelihood

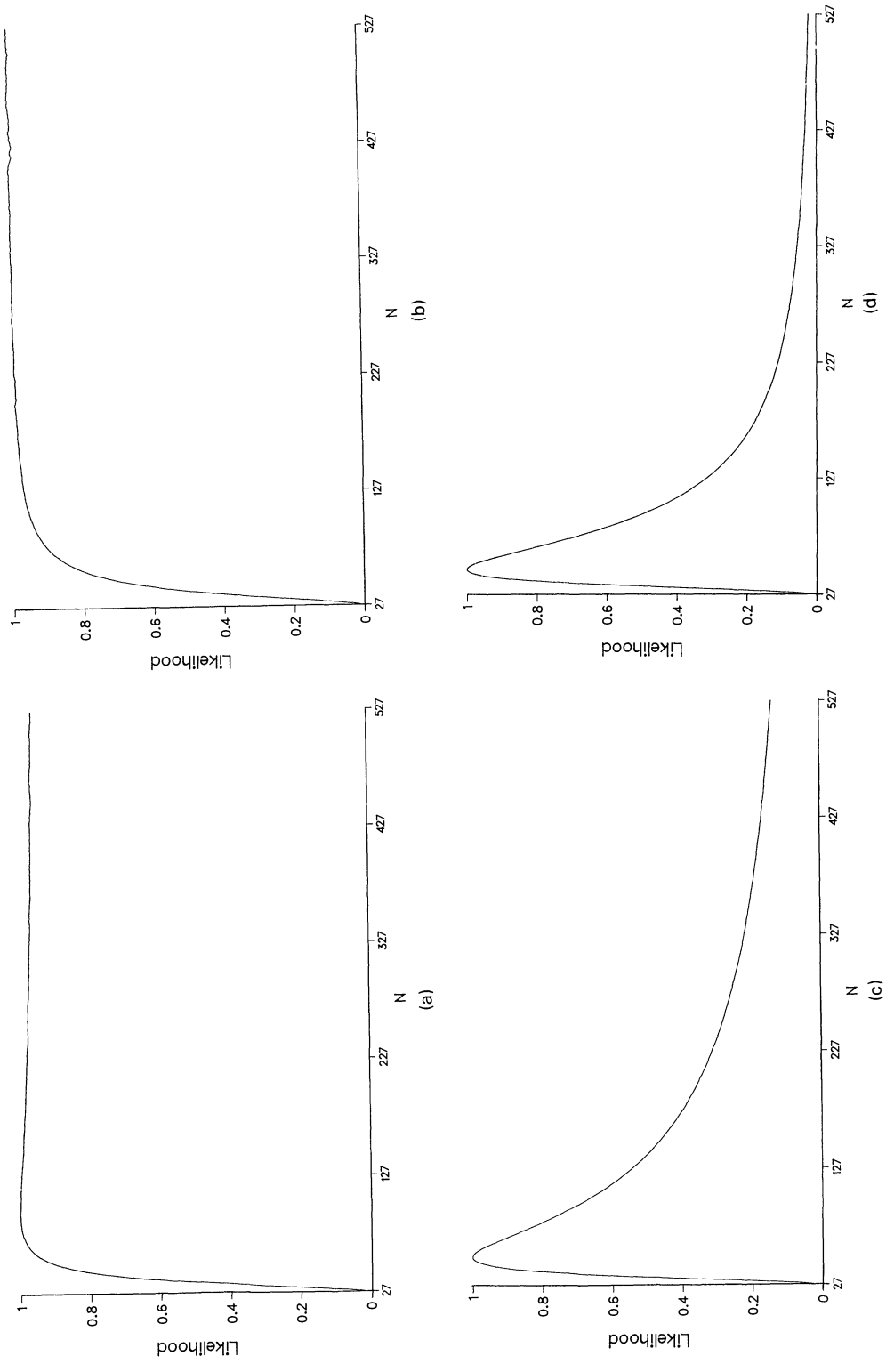

Fig. 3. (a) Profile; (b) conditional; (c) integrated $a = b = 1$; (d) integrated $a = b = 2$

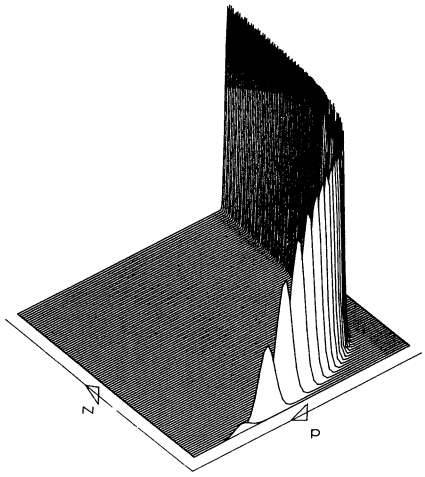

Fig. 4. Likelihood $L(N, p)$

in N and ψ which is almost orthogonal in the parameters, and the resulting likelihood in N is very close to the profile likelihood in the original parameterization.

To construct the posterior Bayes factor for N, we consider models M_1: $s \sim b(N_0, p)$ for fixed N_0 and M_2: $s \sim b(N, p)$ for unspecified N. We take p to have a uniform prior on (0, 1) for both models and N to have an independent diffuse prior on (N_L, ∞) for M_2. Then the posterior mean of the likelihood under M_1 is

$$\bar{L}_1^{\mathrm{A}}(N_0) = \int L^2(N_0, p)\,\mathrm{d}p \Big/ \int L(N_0, p)\,\mathrm{d}p$$

where

$$\int_0^1 L^k(N_0, p)\,\mathrm{d}p = \left\{ \prod_i \binom{N_0}{s_i} \right\}^k \int_0^1 p^{kT}(1-p)^{k(N_0 r - T)}\,\mathrm{d}p$$

$$= \left\{ \prod_i \binom{N_0}{s_i} \right\}^k B\{kT+1,\ k(N_0 r - T)+1\}$$

giving the posterior mean

$$\bar{L}_1^{\mathrm{A}}(N_0) = \prod \binom{N_0}{s_i} \frac{N_0 r + 1}{2N_0 r + 1} \frac{\binom{N_0 r}{T}}{\binom{2N_0 r}{2T}}.$$

Under M_2, the average likelihood is

$$\bar{L}_2^{\mathrm{A}} = \sum_{N=N_L}^{\infty} \int L^2(N, p)\,\mathrm{d}p \Big/ \sum_{N=N_L}^{\infty} \int L(N, p)\,\mathrm{d}p.$$

The value of $\bar{L}_2^{\mathrm{A}}$ has to be obtained numerically, but its value is irrelevant here as there is no formal hypothesis to test about the value of N: $\bar{L}_1^{\mathrm{A}}(N)$ describes the

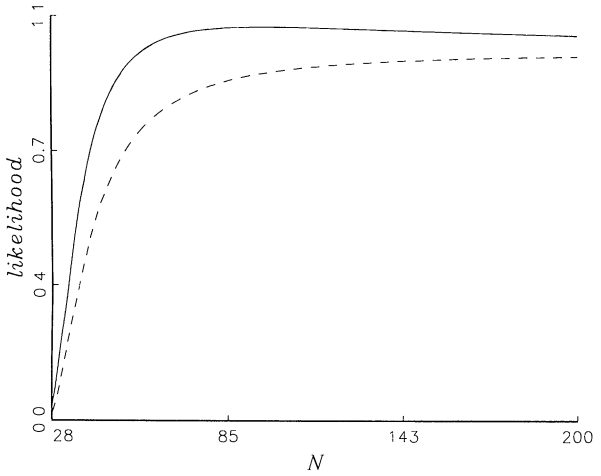

Fig. 5. Profile and average (——) and conditional (- - - -) likelihoods

posterior evidence for different values of N. As $N \to \infty$, the posterior mean $\bar{L}_1^{\mathrm{A}}(N)$ approaches a constant, like the profile and conditional likelihoods. For the example it has a poorly defined maximum at $N=98$. The asymptotic values of all three likelihoods are easily evaluated from Stirling's formula: we have

$$P(N) \to \exp(-T)\left(\frac{T}{r}\right)^T \Big/ S$$

$$C(N) \to (2\pi T)^{1/2} P(N)$$

$$\bar{L}_1^{\mathrm{A}}(N) \to 2^{-1/2} P(N)$$

where $S = \pi s_i!$. These three likelihoods are shown in Fig. 5, scaled by their asymptotic values.

The posterior mean likelihood is indistinguishable from the scaled profile likelihood. The reason is easy to see: for reasonably large N, the likelihood in p for given N is effectively normal, and integrating over p is equivalent to dividing the maximum of the likelihood $L(N, \hat{p}(N))$ by $\sqrt{2}$. Thus for large N the posterior mean likelihood will closely resemble the profile likelihood.

The difference between the prior and posterior means of the likelihood is very clearly evident here. The prior mean with respect to the general beta prior is

$$\bar{L}^{\mathrm{B}}(N) = \prod \binom{N}{s_i} B(T+a,\ Nr-T+b)/B(a,\ b)$$

while the posterior mean is

$$\bar{L}_1^{\mathrm{A}}(N) = \prod \binom{N}{s_i} B\{2T+a,\ 2(Nr-T)+b\}/B(T+a,\ Nr-T+b).$$

As $N\to\infty$, $\bar{L}^{B}(N)\to C_1 N^{-a}$, while $\bar{L}_1^{A}(N)\to C_2$, where C_1 and C_2 are independent of N. If the sample is not strongly underdispersed, as is the case here, the Poisson model will be a plausible alternative to the binomial model, and the likelihood as $N\to\infty$ should be appreciable. Under the prior mean, if $a>0$ then the prior probability of p goes to zero with p, so N cannot be large, and the Poisson model is made implausible by the prior specifications, whatever the data; the posterior mean as $N\to\infty$ is unaffected by this specification.

6. DISCUSSION

The posterior Bayes factor performs well in the examples discussed here, and in many others to be presented elsewhere. (Models with the number of nuisance parameters of the same order as the sample size—Neyman–Scott problems—suffer the same difficulties in this approach as in conventional maximum likelihood, and require an explicit model for the nuisance parameters, e.g. a variance component model.) We consider further its logical justification.

We note first that the use of the *posterior distribution of the likelihood* has been suggested before: Dempster (1974) proposed the use of the posterior means of the *log*-likelihoods for model comparisons, a difference of the order of log(1/20) being strong evidence. In an unpublished report, Raghunathan (1984) applied this approach to the variable subset selection problem in normal regression. With normal likelihoods another penalized maximized likelihood criterion results, with easily calculated penalty function: If $L(\boldsymbol{\theta})$ is normal with $\boldsymbol{\theta}$ of dimension ν, then

$$E\{\log L(\boldsymbol{\theta})|\mathbf{y}\}=\log L(\hat{\boldsymbol{\theta}})-\nu/2.$$

The corresponding penalty for $-2\log L(\hat{\boldsymbol{\theta}})$ is ν, giving the Nelder and Wedderburn proposal.

However, in other models the average log-likelihood is not simple. For example, in the normal regression model of Section 3, the posterior mean of the log-likelihood is

$$E(\log L_j|\mathbf{y})=\frac{n}{2}\left\{\psi\left(\frac{n-p_j}{2}\right)+\log\left(\frac{2}{\mathrm{RSS}_j}\right)\right\}\Big/\,\Gamma\left(\frac{n-p_j}{2}\right)\left(\frac{2}{\mathrm{RSS}_j}\right)^{(n-p_j)/2}-\frac{p_j}{2}$$

where ψ is the digamma function. The difference between the posterior mean log-likelihoods for two models is not a function of the ratio of residual sums of squares.

The use of the posterior mean of the likelihood has caused concern to referees because of the appearance of 'using the data twice', once to obtain the posterior distribution of $\boldsymbol{\theta}$, and again to average the likelihood—essentially the posterior—with respect to the posterior.

One response to this is simply to note that, once the data $\mathbf{y}$ are observed, any function of $\boldsymbol{\theta}$ and the data has a posterior distribution which can be derived from that of $\boldsymbol{\theta}$. The likelihood $L(\boldsymbol{\theta})$ is one such function, and we can in theory determine its full posterior distribution, as discussed by Dempster (1974). In practice this distribution is very complicated, and we settle for the first moment of the posterior distribution of $L(\boldsymbol{\theta})$, i.e. the posterior mean, which is easily evaluated in many models. From this point of view, the posterior mean of L has the same inferential status as the posterior mean of $\boldsymbol{\theta}$ or of any other function of $\boldsymbol{\theta}$, and the ratio $\bar{L}_1^{A}/\bar{L}_2^{A}$ of posterior means

of L_j for the different models M_j is a practical alternative to the likelihood ratio $L_1(\boldsymbol{\theta}_{1S})/L_2(\boldsymbol{\theta}_{2S})$ at the (unknown) fully specified values $\boldsymbol{\theta}_{1S}$ and $\boldsymbol{\theta}_{2S}$.

A second response is to note that the posterior averaging is equivalent to a form of penalty on the maximized likelihood, so that there is a close analytic connection between the conventional likelihood ratio test and the posterior Bayes factor. The form of penalty is similar to that of Akaike's information criterion in normal likelihoods, but it is model specific and does not require normality of the likelihood. The penalty serves the same purpose as the sampling distribution of the likelihood ratio test statistic—to allow for the maximizations over different parameter spaces.

A third response is to note that, given $\mathbf{y}$, the predictive distribution of new data $\mathbf{z}$ generated from the same model is

$$f(\mathbf{z}|\mathbf{y}) = \int f(\mathbf{z}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta}.$$

If we evaluate the predictive density for data $\mathbf{z}$ which has the same sufficient statistics as $\mathbf{y}$, then $f(\mathbf{z}|\boldsymbol{\theta})$ is identical with the likelihood function $L(\boldsymbol{\theta})$, and the posterior mean of the likelihood can be interpreted as the predictive probability of new data with the same sufficient statistics as the observed data. The ratio of two such predictive probabilities is then interpreted like a (prior) Bayes factor.

We note finally a non-Bayesian direct likelihood interpretation of the posterior mean. We may regard the normalized likelihood

$$w_j(\boldsymbol{\theta}_j) = L(\boldsymbol{\theta}_j) \Big/ \int L(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j$$

simply as a weight function without a Bayes posterior interpretation and regard

$$\bar{L}_j^{\mathrm{A}} = \int L_j^2(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j \Big/ \int L_j(\boldsymbol{\theta}_j)\, \mathrm{d}\boldsymbol{\theta}_j$$

as the best available one-point summary of the likelihood $L_j(\boldsymbol{\theta}_j)$ as the evidence for M_j. This is formally equivalent to the posterior mean with a diffuse prior for $\boldsymbol{\theta}_j$. In a non-Bayes framework the question of parameterization of L_j arises, since the integral form of $\bar{L}_j^{\mathrm{A}}$ is not invariant to arbitrary monotone transformation of $\boldsymbol{\theta}_j$. A reasonable approach to this problem is to note that, in real finite populations with finite measurement precision of $\mathbf{y}$, parameters like means or proportions can take values only on an equally spaced grid, and therefore there is a natural uniform scale $\boldsymbol{\theta}_s$ for such parameters. On this scale, the average of the likelihood is

$$\bar{L} = \sum_s L^2(\boldsymbol{\theta}_s) \Big/ \sum_s L(\boldsymbol{\theta}_s)$$

and the sums can be approximated by the corresponding integrals, since

$$\sum_s L^k(\boldsymbol{\theta}_s) \cdot \boldsymbol{\delta} \simeq \int L^k(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$$

where $\boldsymbol{\delta}$ is the grid interval for $\boldsymbol{\theta}$. A monotone transformation of $\boldsymbol{\theta}$ to $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ leaves the finite sums $\Sigma_s L^k(\boldsymbol{\phi}_s)$ unchanged, and so the correct discrete form of the average likelihood is invariant. To maintain this invariance, the integral approximations have to incorporate the differential of the transformation:

$$\int L^k(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \int L^k(\boldsymbol{\phi})\left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\boldsymbol{\phi}}\right)\mathrm{d}\boldsymbol{\phi}.$$

Thus with this convention the average likelihood is invariant to monotone parameter transformations: it is only necessary to determine the parameter scale on which uniform spacing is appropriate.

7. CONCLUSION

The comparison of models through the posterior Bayes factor is generally applicable for arbitrary models, requires no more than conventional diffuse prior specifications for the model parameters and does not suffer from Lindley's paradox. Variations in the prior specification have little effect on the posterior Bayes factor in reasonable sample sizes.

Applications of this approach to other model comparisons problems will be presented elsewhere.

ACKNOWLEDGEMENTS

I am grateful to Jim Berger, Bob Berk, Steve Fienberg, Camil Fuchs, Sam Oman, Richard Smith and David Steinberg for comments, and to referees of an earlier version of this paper for many helpful comments and references.

REFERENCES

Aitkin, M. A., Anderson, D. A., Francis, B. J. and Hinde, J. P. (1989) *Statistical Modelling in GLIM*, sect. 2.7. Oxford: Clarendon.

Aitkin, M. A. and Stasinopoulos, M. (1989) Likelihood analysis of a binomial sample size problem. In *Contributions to Probability and Statistics* (eds L. J. Gleser, M. D. Perlman, S. J. Press and A. R. Sampson). New York: Springer.

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

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. *J. Am. Statist. Ass.*, **82**, 112–122.

Carroll, R. J. and Lombard, F. (1985) Note on N estimators for the binomial distribution. *J. Am. Statist. Ass.*, **80**, 423–426.

Casella, G. and Berger, R. L. (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Am. Statist. Ass.*, **82**, 106–111.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

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. Blæsild and G. Sihon), pp. 335–352. Aarhus: University of Aarhus.

Draper, N. and Guttman, I. (1971) Bayesian estimation of the binomial parameter. *Technometrics*, **13**, 667–673.

Dumonceaux, R., Antle, C. E. and Haas, G. (1973) Likelihood ratio test for discrimination between two models with unknown location and scale parameters. *Technometrics*, **15**, 19–27.

Edwards, A. W. F. (1972) *Likelihood*. Cambridge: Cambridge University Press.

Gabriel, K. R. (1964) A procedure for testing the homogeneity of all sets of means in analysis of variance. *Biometrics*, **20**, 459–477.

Kahn, W. D. (1987) A cautionary note for Bayesian estimation of the binomial parameter *n*. *Am. Statistn*, **41**, 38–39.

Lindley, D. V. (1957) A statistical paradox. *Biometrika*, **44**, 187–192.
Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.
Olkin, I., Petkau, A. J. and Zidek, J. V. (1981) A comparison of n estimators for the binomial distribution. *J. Am. Statist. Ass.*, **76**, 637–642.
Raftery, A. E. (1988) Inference for the binomial N parameter: a hierarchical Bayes approach. *Biometrika*, **75**, 223–228.
Raghunathan, T. E. (1984) A new model selection criterion. *Research Report S-96*. Department of Statistics, Harvard University, Cambridge.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Shafer, G. (1982) Lindley's paradox. *J. Am. Statist. Ass.*, **77**, 325–351.
Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C. and Dransfield, M. (1985) The implementation of the Bayesian paradigm. *Communs Statist.* A, **14**, 1079–1102.
Smith, A. F. M. and Spiegelhalter, D. J. (1980) Bayes factors and choice criteria for linear models. *J. R. Statist. Soc.* B, **42**, 213–220.
Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factors for linear and log-linear models with vague prior information. *J. R. Statist. Soc.* B, **44**, 377–387.

DISCUSSION OF THE PAPER BY AITKIN

G. A. Barnard (Colchester): Murray Aitkin has presented us with an eminently discussable paper which raises many important issues and presents a most useful addition to likelihood methods. I must confine myself to comments on a few of the many introductory points he makes before coming to the main point, and then commenting on his final example.

His criticisms of conventional fixed significance levels are well taken and one can only wonder when 0.05 and 0.01 will disappear from the text-books. They had their origin in limitations of computing power which have long since disappeared. It surely cannot be long before quotation of attained mid-P-values becomes general. Anyone studying a single set of data can then set a critical P-value judged appropriate to the circumstances and, more importantly, it will become easy to combine independent sets of data by, for instance, simple addition of mid-P-values. It is high time that we recognized that situations where important issues turn on single data sets are to be avoided wherever possible.

One advantage that log-likelihood ratios have over P-values is that log-likelihood ratios from independent data sets combine unambiguously by simple addition to produce a combined value of the same kind, though that is not their only advantage. We do not have to leave the Neyman–Pearson framework to see the anomalies arising from fixed significance levels. A very slight extension of the argument used by Neyman and Pearson to derive their fundamental lemma shows that in a series of tests of simple hypotheses H_{0i} against a series of simple alternatives H_{1i}, $i=1, 2, \ldots$, we minimize the long run risk of error of the second kind, subject to a specified upper bound on the long run risk of error of the first kind, by keeping constant the critical likelihood ratio, not the critical α-level. This fact, noted by Pitman (1965) a quarter of a century ago, still seems not to have reached text-books following the Neyman–Pearson approach. As a result, concentration on the likelihood ratio is often thought of as a Bayesian idea—a tendency encouraged by the term 'Bayes factor'.

However, I do not wish here to attack the Bayesian position. I agree with Jack Good on the need for a 'Bayes–non-Bayes' compromise, though perhaps we approach the compromise from opposite sides. In this spirit I wish that we could agree on regularly quoting the observed likelihood ratio $L(H_1|\mathbf{y})/L(H_0|\mathbf{y})$ in addition to attained mid-P-values on H_0, and power on a specified alternative H_1. With Aitkin's example at the top of the second page such a practice would produce the correct conclusion—that model 1 fits much better than model 2, but neither model 1 nor model 2 fits at all well.

The reference on the third page to a 'well-formulated subjective belief about θ_j' worries me. I would be less worried if 'subjective belief' were replaced by 'agreed further assumption'. The primary role of the statistician as such is to present his client(s) with the inferences that flow from the data together with any assumptions concerning the model or the parameters which have been accepted as appropriate—usually on the basis of past experience. If these do not answer all the client's questions he should be told that further assumptions, for which he must take responsibility, are needed. In the model choice

problem these further assumptions will typically change the question being asked—from the choice between models M_j with parameters unknown to the choice between models M_j^* with parameters having specified distributions. It is very difficult to imagine circumstances in which such a process could properly be described as 'subjective'. It is true that the use of diffuse, vague or Jeffreys priors is 'well established'—or, at least, very common; but one wishes the use of likelihood plots, which usually serve their purposes better, was as well established among statisticians as it is among geneticists and some physicists.

The author's reference to our late friend Morris DeGroot's outstanding contribution to the Shafer (1982) discussion is very welcome. The only point DeGroot made that seemed to me to call for further comment was his reference to the fact that an assumption of a uniform prior for an unknown θ implies that an equally unknown $\eta = \exp\theta/(1+\exp\theta)$ was almost certainly equal to zero or to unity. The lack of invariance of a uniform prior under non-linear transformation was decisive for Fisher in his rejection of a Bayesian approach; but we should not forget that Fisher's mentor 'Student' failed to distinguish between Fisher's likelihood and a posterior relative to a uniform prior. Neither Gosset nor Fisher was lacking in logical penetration, and we should ask why their views differed in this respect. I think that it was because Gosset's parameters were always measured on a natural scale, as so many bushels per acre, for example. But Fisher's genetical recombination fractions, interpretable though they might have been in terms of chromosome distances, did not have any clearly defined natural scale. So Gosset would have had very strong grounds for objecting to any transformation such as that from θ to η whereas Fisher would not. This distinction bears on the choice of parameters for Professor Aitkin's posterior Bayes factors.

In the model choice problem, when the parameter distributions are unknown, Aitkin has had the ingenious idea of allowing the data themselves, so far as possible, to supply the additional assumption. Using a uniform prior for all the parameters in each model, the data provide posterior distributions $\pi_j(\boldsymbol{\theta})$ for the parameters; and then the problem can be restated as that of comparing models M_j^*, where M_j^* consists of M_j together with the distribution $\pi_j(\boldsymbol{\theta}_j)$ for its parameters. The ratio of marginal probabilities of the data on the two models M_j^* then provides the required likelihood ratio or 'posterior Bayes factor'. Provided that the client accepts the choice between the M_i^* as a proper reformulation of his problem, the likelihood ratio needs no reinterpretation in terms of test sizes—a consistent use of, say, a critical value of 10/1 in a long run of cases means that, on the assumption that one or other of the M_i^* is correct, then the correct choice will be made at least 10 times as often as the incorrect choice. The restriction that the choice lies between the two M^* models is important, and it is worthwhile to point out that situations will arise where both M^* models fail to be credible in the light of the data on the basis of a test procedure not necessarily related to the choice between them. And what is said above about the possible inappropriateness of taking uniform priors as a starting point shows that care is needed to ensure that the scales of measurement of the $\boldsymbol{\theta}$ can be regarded as natural. Provided that these cautions are borne in mind, the symmetry between the two models implicit in the Aitkin procedure, and the fact that it depends only on the data and its parameterization, must make it a very valuable addition to our collection of standard procedures. Simple addition of log-likelihoods from independent data sets will no longer apply, but the specification of the π_j for each data set will make combination quite easy.

As already indicated, there are sound reasons, independent of Bayesian doctrine, for treating the likelihood ratio as being at least on a par with test sizes as indicators of strength of evidence, but the fact that in many cases the Aitkin procedure is effectively equivalent to penalized likelihood helps to relate it to the now large amount of work that has been done since the pioneering efforts of Jeffreys concerning model choice. Jeffreys's simplicity postulate represented a profoundly original insight into the way that scientific progress is made. The fact that Jeffreys tentatively related his measure of simplicity to the number of coefficients in a differential equation suggested that he was primarily concerned with physical science. And in some scientific fields theories may be more acceptable than others just because they are not over-simple in relation to their subject matter. The general notion of acceptability of a model and the notion of 'naturalness' of a particular parameterization are topics that both need to be addressed if we wish to extend our statistical methods towards a general theory of scientific inference.

I have left myself very little time to comment on the binomial index problem. The current problems relating to the acquired immune deficiency syndrome epidemic illustrate that statisticians must sometimes be prepared to make estimates based on the most inadequate data. But a check on the references given for practical applications of the binomial index problem leave me with the impression that, if a statistician

is provided only with the observations $s_1, s_2, \ldots, s_r$ and is asked on that basis to estimate N, he should very often refuse to do so. Unless the index of dispersion is significantly low, the Poisson possibility, equivalent to infinite N, cannot be ruled out. In one allegedly practical case, the s_i represented the number of appliances brought in for repair over a given period. It was suggested that these could be used to estimate the total number of appliances in use, knowledge of the rate of breakdown of appliances being absent. For a binomial model accurately to underlie such data would require an unbelievable absence of what in the motor trade are referred to as 'Friday' or 'Monday' products; even given such uniformity of production the time and effort needed to present the small amount of information about N contained in the data would almost certainly have been better spent in obtaining further data bearing directly on the p-parameter. None-the-less Aitkin's analysis of this theoretical problem reminds us not only of the much greater graphical facilities that are now at our disposal for likelihood plots—and the value of careful choice of the parameters for these plots—but also of his other valuable contributions to likelihood theory—his work on profile and on canonical likelihood.

It gives me the greatest pleasure to move a warm vote of thanks.

Professor D. V. Lindley (Minehead): The method of posterior Bayes factors is seriously flawed and cannot be recommended.

To demonstrate this, consider four models, M_{11}, M_{12}, M_{21} and M_{22}, for some data. These models are supposed simple, i.e. each completely specifies the probability distribution for the data without a nuisance parameter ($\boldsymbol{\theta}$ in the paper). The likelihoods are L_{ij} and the models have probabilities π_{ij}. In addition, take two other models, $M_{11} \cup M_{21}$ and $M_{12} \cup M_{22}$. These are composite, saying that the model is either specified by the distribution corresponding to M_{1j} or by that of M_{2j} ($j = 1, 2$). Effectively there are nuisance parameters taking two values only in each. We calculate some posterior Bayes factors and to do this we need the expression for the posterior mean likelihoods given near the beginning of Section 2, namely

$$\int L^2(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \Big/ \int L(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}, \tag{1}$$

where suffixes have been omitted.

In the first row of Table 3, three model comparisons are proposed. In the first two the simple models with common first suffix are being compared with respect to their second suffix. In the last comparison, $M_{11} \cup M_{21}$ is taken with $M_{12} \cup M_{22}$, each being a union of first suffixes. The posterior Bayes factors are easily found from expression (1). With the simple models, the integrals have only one term and ratio (1) is simply L. The factors are listed in the second row of Table 3. Now the likelihoods are arbitrary positive values, as are the probabilities with the additional requirement that they add to unity. The third row of Table 3 gives the numerical values of the factors for the case described below it. The result is that M_{11} is thought more plausible than M_{12} (a factor of 1.5) and M_{21} more plausible than M_{22} (1.5 again), yet the union of M_{11} and M_{21} is thought less plausible than the union of M_{12} and M_{22} (the factor is 0.77). The effect of the unions is to reverse the comparisons of second suffixes.

This is ridiculous. Consider models for the next person to enter the room. Let the first suffix correspond to height with tall (short) meaning above (below) average height, and the second to sex. Then advocates of posterior Bayes factors could find themselves saying 'It is more likely to be

TABLE 3†

Comparison	M_{11} v M_{12}	M_{21} v M_{22}	$(M_{11} \cup M_{21})$ v $(M_{12} \cup M_{22})$
Posterior Bayes factor	$\dfrac{L_{11}}{L_{12}}$	$\dfrac{L_{21}}{L_{22}}$	$\dfrac{L_{11}^2\pi_{11} + L_{21}^2\pi_{21}}{L_{11}\pi_{11} + L_{21}\pi_{21}} \dfrac{L_{12}\pi_{12} + L_{22}\pi_{22}}{L_{12}^2\pi_{12} + L_{22}^2\pi_{22}}$
Numerical value	1.5	1.5	0.77
Odds	$\dfrac{L_{11}\pi_{11}}{L_{12}\pi_{12}}$	$\dfrac{L_{21}\pi_{21}}{L_{22}\pi_{22}}$	$\dfrac{L_{11}\pi_{11} + L_{21}\pi_{21}}{L_{12}\pi_{12} + L_{22}\pi_{22}}$

†The numbers in the third row are derived by taking
$L_{11} = 9$ $L_{12} = 6$ $L_{21} = 3$ $L_{22} = 2$,
$\pi_{11} = 0.05$ $\pi_{12} = 0.45$ $\pi_{21} = 0.45$ $\pi_{22} = 0.05$.

a tall man than a tall woman,
a short man than a short woman,
a woman than a man'.

One hardly advances the respect with which statisticians are held in society by making such declarations.

A natural question that arises is whether there is a method that avoids this difficulty, for perhaps it is inevitable that such contradictions occasionally arise. It is not; there is a sensible way to proceed. The fourth row of Table 3 gives the comparisons using the odds. The numerator (denominator) of the factor for comparison of the composite models is the sum of the numerators (denominators) for the comparisons of the simple models. So if the numerators exceed the denominators in the simple cases, they will continue to do so in the composite case. The contradiction cannot occur.

This is all easy. However, there is a deeper result. A theorem says that essentially only the odds will avoid the difficulty. This was first proved by Ramsey, though appreciation of it dates from other proofs by Savage, de Finetti and Jeffreys. The best exposition is to be found in chapter 6 of DeGroot (1970), which I recommend as required reading for all statisticians. The requirement that the above contradiction does not arise is his assumption SP_2.

The theorem says that the only way to compare models is through the probabilities of these models and hence, with two models, the ratio. In other words, to avoid situations like that developed above for posterior Bayes, you must act like a Bayesian. Any method that is not Bayesian will contain contradictions. Experience shows me that it is often difficult to exhibit explicitly a counter-example to a non-Bayesian proposal. I have been able to do so here because Aitkin's ideas are clear cut and not hedged about with vague qualifications. For this reason I have pleasure in seconding the vote of thanks for a well-written paper.

The vote of thanks was passed by acclamation.

Sir David Cox (Nuffield College, Oxford): I agree with Professor Barnard that we must distinguish between

(a) the assessment of the relative fit of two models, M_1 and M_2, assuming provisionally that one of the models is 'true', and
(b) analysis of the adequacy of M_1 looking for departures in the direction of M_2, and vice versa.

In (b), the conclusion may be that the fit of both, one or neither model is adequate.

The intriguing introduction of A in Section 2 of the paper can be assessed by calibration, i.e. by examining performance under hypothetical repetitions (and I would have liked to have seen more explicit discussion of this, particularly for some well-known treacherous examples with many nuisance parameters), or by comparison with logical principles, such as those of Bayesian theory. It is unlikely on general grounds that the effectively data-dependent prior could be consistent with the usual Bayesian requirements and Professor Lindley has shown this via a striking example. It helps, however, to consider the form of A in the light of the distinction between coherence and temporal coherence, the latter term due, I think, to P. Suppes; I. J. Good has written about dynamic probability.

Suppose that in the planning of an investigation we consider hypotheses H_i; denote the data, yet to be obtained, by $\mathbf{y}$. Then coherence asserts

$$P(H_i|\mathbf{y}) \propto P(H_i)\ P^*(\mathbf{y}|H_i).$$

The factor with an asterisk is hypothetical, whereas $P(H_i)$ concerns our current knowledge about H_i from other sources. Time passes, data $\mathbf{y}_0$ are collected and now coherence asserts

$$P(H_i|\mathbf{y}_0) \propto P^*(H_i)\ P(\mathbf{y}_0|H_i)$$

where $P^*(H_i)$ is hypothetical in the sense that it concerns knowledge which we would have had in the absence of the $\mathbf{y}_0$ which we do have. Temporal coherence requires $P(H_i) = P^*(H_i)$. Often, but certainly not always, this is a very reasonable working assumption. It may, however, be that the work of data collection, or the data themselves, lead to the formulation of an H previously rejected or ignored, especially when the data show some unanticipated effect. The latter is the Bayesian analogue of data-snooping with corresponding dangers (and advantages). In many fields plausible *ex post* explanations of 'strange' effects can be constructed. It is important not to dismiss such hypotheses generated by the data, but to recognize them as in some way distinct from hypotheses considered *a priori*.

One interpretation of Professor Aitkin's A is that it represents a mechanical adjustment of $P(H_i)$,

in the light of $\mathbf{y}_0$. I am very uneasy at this as a general procedure, but I do not think that it is logically incorrect in the above framework; calibration is an ultimate test!

R. L. Smith (University of Surrey, Guildford): My comments are concerned with the binomial population size problem discussed in Section 5 and, at greater length, in the cited paper of Aitkin and Stasinopoulos (1989).

First, let us be clear about what the problem is here. It is not so much the discreteness of N, since this can almost be treated as a continuous variable, but the fact that N is in effect an end point of the distribution, far away from any of the data points. As such, the issues raised are similar to those which arise in statistical inference about extreme values.

Aitkin and Stasinopoulos showed that, typically, the profile likelihood for N is flat but a Bayesian analysis, based on independent prior distributions for N and p, leads to a sharply peaked posterior density for N. However, if the problem is reparameterized in terms of N and $\psi = Np$, and independent prior distributions taken for N and ψ, then the posterior density of N is very similar to the profile likelihood.

To me this is an important point. If nothing else, it serves as a salutary warning of the difficulties of defining an uninformative prior distribution in multiparameter problems.

In his paper, Professor Aitkin takes this analysis a step further, showing that if posterior Bayes factors are used to compare different values of N, even in the original parameterization, the results are again very similar to those deduced from the profile likelihood.

I would like to raise two questions for Professor Aitkin. First, the posterior Bayes factors approach is itself not parameterization invariant, so it might still be desirable to compare the effects of different parameterizations of the problem. Would this have any material effect on the results?

Secondly, the broad conclusion with this example seems to be that, once the first Bayesian analysis is discarded, the others are equivalent to the profile likelihood. I found this an unsatisfactory conclusion. A possible explanation is that, for this problem, the structure of the nuisance parameter is very simple—if we assume that N is known, then we all know what to do about p. Would a problem with a more complicated, perhaps multidimensional, nuisance parameter require a more detailed comparison of the different procedures available?

Professor A. F. M. Smith (Imperial College of Science, Technology and Medicine, London): Suppose that data $\mathbf{x}$ are to be used to compare or choose among a set of alternative models $\mathfrak{M} = \{M_i, i \in I\}$. Any criterion for model comparison should surely depend on two things: first, the decision context in which the comparison is taking place; secondly, the perspective from which the model set $\mathfrak{M}$ is viewed.

Figs 6 and 7 illustrate two possible such decision contexts. In the first case, given data $\mathbf{x}$ we choose M_i; subsequently some 'state of the world' $\mathbf{w}$ obtains and the utility $u(M_i, \mathbf{w})$ ensues. In the second

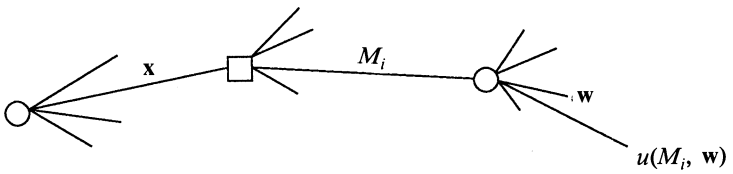

Fig. 6. Decision problem involving model choice only

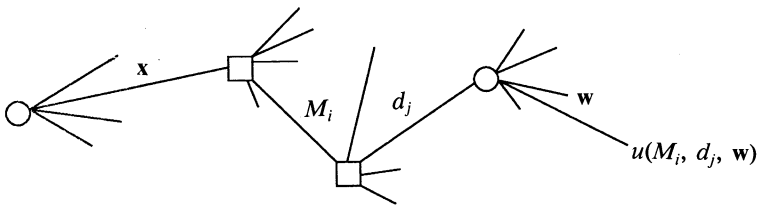

Fig. 7. Decision problem involving model choice and subsequent decision

case, given data $\mathbf{x}$ we choose M_i and, based on M_i, take a decision d_j; some state of the world $\mathbf{w}$ then obtains and the utility $u(M_i, d_j, \mathbf{w})$ ensues. In both cases, the optimal model choice is M^*, where

$$\bar{u}(M^* | \mathbf{x}) = \sup_i \{\bar{u}(M_i | \mathbf{x})\},$$

with

$$\bar{u}(M_i | \mathbf{x}) = \int u(M_i, \mathbf{w})\, p(\mathbf{w} | \mathbf{x})\, d\mathbf{w}$$

in the first case, and in the second case

$$\bar{u}(M_i | \mathbf{x}) = \int u(M_i, d_i^*, \mathbf{w})\, p(\mathbf{w} | \mathbf{x})\, d\mathbf{w},$$

where d_i^* is the d_j that maximizes $\int u(M_i, d_j, \mathbf{w})\, p_i(\mathbf{w} | \mathbf{x})\, d\mathbf{w}$.

In the last expression, $p_i(\mathbf{w} | \mathbf{x})$ denotes beliefs about $\mathbf{w}$ conditional on $\mathbf{x}$ and model M_i. Throughout, $p(\mathbf{w} | \mathbf{x})$ represents beliefs about $\mathbf{w}$ conditional on $\mathbf{x}$ and one's actual perspective on 'the true model', making clear that the perspective from which $\mathfrak{M}$ is viewed is a crucial part of the analysis.

There are three possible perspectives, which, based on joint work with J. M. Bernardo, can be identified as follows.

(a) The $\mathfrak{M}$-*closed* view corresponds to acting as if M_i, $i \in I$, exhaust all conceivable modelling possibilities; equivalently, one of the M_i is assumed to be the true model. In this case, we have

$$p(\mathbf{w} | \mathbf{x}) = \sum_i p(\mathbf{w} | M_i, \mathbf{x})\, p(M_i | \mathbf{x}),$$

and the required expected utility calculations proceed relatively straightforwardly.

(b) The $\mathfrak{M}$-*completed* view acknowledges M_i, $i \in I$, to be convenient possible proxies for an identified distinct model M_t, which is regarded as the true model but is perhaps too cumbersome to use in routine practice. In this case,

$$p(\mathbf{w} | \mathbf{x}) = p(\mathbf{w} | M_t, \mathbf{x})$$

and the required calculations typically involve extensive numerical integrations.

(c) The $\mathfrak{M}$-*open* view again acknowledges M_i, $i \in I$, to be proxies but does not identify M_t, so that $p(\mathbf{w} | \mathbf{x})$ is not directly available.

The (prior) Bayes factor is a key ingredient in the case of Fig. 6, with 0–1 utilities and an $\mathfrak{M}$-closed perspective. However, as this discussion hopefully makes clear, this is just a very particular, reasonably well-illuminated, corner of a much bigger, murkier world of model choice problems.

A. C. Davison (University of Oxford): One motivation for the work described by Professor Aitkin is the difficulties associated with discriminating between separate, i.e. non-nested, families of hypotheses (Cox, 1961). I have two comments on Section 1.1 of the paper.

First, the problem that the size of tests depends on the values of parameters unknown under the null hypothesis can sometimes be eliminated, along with the parameters themselves, by conditioning on statistics sufficient under the null hypothesis. This approach is taken by Pace and Salvan (1990). In general approximate conditional tests would be obtained by conditioning on maximum likelihood estimates obtained under the null hypothesis.

When the competing models are specified up to unknown parameters, the simulation effort required to estimate significance levels can be reduced by generating data not from the null but from the alternative hypothesis, and weighting the results appropriately. As an example, consider the data due to Bliss (1935) and subsequently analysed by Pregibon (1980), concerning the deaths of flour beetles at different doses of a poison. The deviances for a binomial model with the logit and complementary log–log link functions are 11.22 and 3.44, both on six degrees of freedom. The second model fits better, but is the difference of deviances significant? The usual χ^2 result does not apply because the models are not nested. We aim to estimate the probability that the difference in deviances exceeds 7.78 under each model. Importance sampling from the complementary log–log model to fix things up *as if* the sampling were from the logistic distribution gives estimated significance level 0.0033, with standard error 1.8×10^{-4} based on 1000 simulations. The variance of this estimate is approximately 100 times smaller than would have been the case for simulation from the null distribution; theoretical calculations by Anna Gigli at Imperial College bear this gain out. This is a worthwhile gain in efficiency, but there is more. The simulation was performed from the complementary log–log model, and hence the significance level taking

this model as the null hypothesis can be obtained directly in the usual way. Thus importance sampling not only gives an improved estimate of one tail probability; it also gives the other one *gratis*. More details will be written up shortly by Dr Gigli and myself.

Michael Goldstein (University of Durham): Consider the following counter-example to the interpretation of posterior Bayes factors.

From a random number table, I select a random integer between 1 and 1000. I do not reveal this value, which serves as the parameter θ. I now generate a data value Y, which I will report to you. There are two different models under which the data can be generated. Under model 1, the value Y that I report is equal to the previously selected value θ. Under model 2, I ignore θ and instead select and report a value Y which is a random integer, between 1 and 1000.

Having seen Y, you must guess whether the data value was generated under model 1 or model 2, i.e. whether you are seeing the first or the second of two random integers. The data value carries no information to distinguish between the two models, and this is what a statistician should report.

The usual Bayes factor is unity. However, because the likelihoods are all unity or zero under model 1, but all 1/1000 under model 2, the posterior Bayes factor favours model 1 by a factor of 1000 to 1. (In the author's words, this should constitute 'overwhelming' sample evidence against model 2 in favour of model 1.) We can easily elaborate this example so that for half of the sample values the posterior Bayes factor gives overwhelming support to model 1, for half to model 2, and yet, as above, there is no information in the data to distinguish the two models.

What do counter-examples such as this imply? Not that we should necessarily follow the Bayes route, but that at the least, when we change the rules, we should expect various pathologies, even in the simplest problems, which must be carefully addressed (somehow!) before we can hope that the new methods will be reliable for difficult problems.

Dr T. Fearn (University College London): My analysis of the logic of this paper is as follows.

(a) Model comparison depends critically on the prior distributions for the unknown parameters in the models.
(b) Actually specifying these priors would be difficult and subjective (and by unspoken implication bad).
(c) Here is an arbitrary prior (which is 'objective' and therefore good). Objective here means that it can depend on the experiment, the observations or anything else as long as it *does not* depend on your prior beliefs about the parameters in question.
(d) Use it and the model comparison does not depend on the prior.

The beginning and end points of this sequence seem contradictory—Aitkin's paradox perhaps? This paradox at least is easily resolved. The illogical step is the sleight-of-hand whereby the calculated Bayes factor—which *is* a Bayes factor for one very special (and difficult to defend) choice of prior—is implicitly regarded as the Bayes factor for the diffuse prior that happened to be involved in its calculation.

The author, as he points out, has a perfect right to calculate the posterior mean of any quantity he wishes—what he cannot do is to interpret the result as a Bayes factor, except in so far as it corresponds to one very special (and highly informative) prior.

The only justification for diffuse/ignorance/reference (call them what you will) priors is that the results that they give in some situations approximate those that would result from a wide range of actual prior beliefs. The objective priors used here do no such thing—since you need to know the outcome of the experiment to specify them they cannot approximate *any* coherent prior belief.

Incidentally, problems caused by diffuse priors will still occur more often than the author admits. A good example of the sort of problem that can arise is the equation at the foot of page 123. Since $\backslash L$ and $\backslash L^2$ both have $1/N$ tails this looks like ∞/∞. One cannot be quite so cavalier with improper priors and get away with it all the time.

Professor D. A. Sprott (University of Waterloo): The example described in Section 1.3 hardly seems to be a 'paradox'. It seems only to illustrate the way in which the Bayesian approach using *ad hoc* diffuse priors leads to an endless thrashing around to find one that makes sense. The simplest and most direct approach would be the use of the likelihood ratio mentioned by Professor Aitkin. It has a simple interpretation in terms of the ratio of objective frequencies and avoids the special pleading required by the use of diffuse priors.

If diffuse priors are used, Professor Aitkin's procedure which uses posterior Bayes factors seems preferable, and to produce more convincing results, than the use of Bayes factors.

In the comparison of two models M_1 and M_2 care has to be used in obtaining the likelihood functions when one model is not a submodel of the other. In writing down the likelihood $L_j(\boldsymbol{\theta}_j)$ it must be remembered that j is also a parameter. This determines which constants are irrelevant. In this regard the description at the foot of the fifth page of $L_j(\boldsymbol{\theta}_j)$ as the 'likelihood function of the data $\mathbf{y}$ and parameters $\boldsymbol{\theta}_j$' is misleading. For example, in Section 4 the factor $\exp(-\mathrm{RSS}_j/2\sigma^2)$ in the expression for $L_j(\boldsymbol{\beta}_j, \sigma)$ is still relevant if σ is known, whereas it is not if L_j is the likelihood function of $\boldsymbol{\beta}_j$ only.

Finally, from Figs 3 and 5 it appears that the conditioning in Section 5 to form the conditional likelihood loses somewhat less information than do the profile and average likelihoods.

Professor A. P. Dawid (University College London): Aitkin dismisses the (proper) Bayes approach to model comparison with the comment that it can be 'very sensitive to variations in the priors'. However, this so-called sensitivity deserves closer attention. For large samples, it turns out that (in regular problems, with the parameters consistently estimable) the prior density $\pi_j(\cdot)$ enters $\overline{L}_j^{\mathrm{B}}$ only through a factor $\pi_j(\hat{\boldsymbol{\theta}}_j)$. Changes to the prior densities will thus introduce an asymptotically constant factor into the posterior odds. This will be swamped by the other, data-derived, terms which determine the asymptotic behaviour (tending to zero or infinity depending on which model holds) of the posterior odds. Moreover, the appropriateness of this prior-based term is clear when we realize that, with extensive data, we are effectively observing $\boldsymbol{\theta}_j = \hat{\boldsymbol{\theta}}_j$, and so a comparative assessment of different priors $\pi_j(\cdot)$ and $\pi_j^*(\cdot)$ for the same model would be based on the likelihood ratio $\pi_j(\hat{\boldsymbol{\theta}}_j)/\pi_j^*(\hat{\boldsymbol{\theta}}_j)$ generated by such an observation.

Asymptotic sensitivity analysis in the proper Bayesian approach is thus straightforward, while small sample sensitivity is no harder to assess than in Aitkin's approach. True, improper priors cause trouble, and the above analysis shows why—they yield $\pi_j(\hat{\boldsymbol{\theta}}_j) = 0$. Thus while I agree with Aitkin that 'this argument would limit Bayes analysis to informative priors', I cannot accept his inference that this conclusion (the very basis of much exciting work in modern Bayesian statistics) is something to be avoided, by the development of such *ad hoc* constructions as posterior Bayes factors.

The problem of Section 5 is interesting since (on account of the Poisson approximation to the binomial distribution) for large N and small p the data are effectively informative only about the single parameter $\psi = Np$. Consequently, although posterior inferences about ψ will feature only the unimportant sensitivity to the prior discussed above, those for other parameters will remain highly sensitive. But it should be pointed out that Aitkin's approach does *not* remove this sensitivity. In particular, if he were to re-do his analysis with a prior incorporating independence between N and ψ, which seems perfectly sensible, he would obtain entirely different results. Where then is the practical pay-off from Aitkin's theoretical compromise?

The following contributions were received in writing after the meeting.

Professor H. Akaike (Institute of Statistical Mathematics, Tokyo): It is often dangerous to advance an argument related to the likelihood of a model without proper understanding of the basis for the use of the likelihood as a criterion. The repeated use of one and the same sample in the posterior mean of the likelihood function for the definition and evaluation of a posterior density certainly introduces a particular type of bias that invalidates the use of the mean as the likelihood of the model.

Consider two models $y \sim N(\mu, \sigma^2)$ and $y \sim N(\mu, \tau^2)$ with a common diffuse prior for μ, where both σ^2 and τ^2 are known and $\sigma^2 > \tau^2$. The posterior mean of the likelihood function for the first and second model is given by $(2\pi^{1/2}\sigma)^{-1}$ and $(2\pi^{1/2}\tau)^{-1}$ respectively, and the posterior Bayes factor is given by $A = \tau/\sigma$ which is always less than unity. This shows that, other things being equal, the second model is always preferred to the first.

The entropic or information theoretic interpretation of log-likelihood as developed in Akaike (1985) will be useful to avoid this difficulty.

Jack Cuzick (Imperial Cancer Research Fund, London): The most problematic aspect of Professor Aitkin's proposal is to use the same data twice for making a single inference. Such an approach has been successful in empirical Bayes estimation, where the data are used once to estimate the prior distribution of some variable and then again to estimate its value for each observation, because the two uses are in a sense orthogonal. If the prior distribution is estimated by all data points except that under consideration, the answer does not change much. However, this is not so here, and absurd results

will result in limiting cases. A more defensible procedure would be to split the data into two samples, one for estimating the prior and the second for computing the expected likelihood.

Consider for example the normal regression model $Y=N(X'\mu, \sigma^2 I)$, $\mu \in R^p$, $Y \in R^n$, σ^2 known, and split the observations into two samples of size n_1 and n_2, defining $Y_1=N(X_1'\mu, \sigma^2 I)$ and $Y_2=N(X_2'\mu, \sigma^2 I)$. If we assume $I_j=X_jX_j'$, $j=1, 2$, are both of full rank and that a vague prior is assumed for the first sample, the expected likelihood for the second sample, with the posterior from the first sample used as the prior, denoted $E_1(L_2)$, is given by

$$(2\pi\sigma^2)^{-n_2/2}|I_2|^{-1/2}|I_*|^{1/2}\exp\left[-\frac{1}{2\sigma^2}\{R_2+(\hat{\mu}_1-\hat{\mu}_2)'I_*(\hat{\mu}_1-\hat{\mu}_2)\}\right]$$

where $\hat{\mu}_j$ are the maximum likelihood estimates for each sample, $R_2=(Y_2-X_2'\hat{\mu}_2)'(Y_2-X_2'\hat{\mu}_2)$ is the residual sum of squares for the second sample and $I_*=(I_1^{-1}+I_2^{-1})^{-1}$.

If n_1 and n_2 are large and the two samples are chosen at random, then $n_1^{-1}I_1$ and $n_2^{-1}I_2$ are approximately equal and by reparameterizing we can take them to be the identity. With this approximation we obtain the simpler form

$$E_1(L_2)\cong(2\pi\sigma^2)^{-n_2/2}\left(\frac{n_1}{n_1+n_2}\right)^{p/2}\exp\left[-\frac{1}{2\sigma^2}\{R_2+n^*\|\hat{\mu}_1-\hat{\mu}_2\|^2\}\right],$$

where $n^*=(n_1^{-1}+n_2^{-1})^{-1}$.

The first factor is a constant and can be ignored, the second is a penalty for large p, while R_2 measures the maximum likelihood fit of the second sample, and $\|\hat{\mu}_1-\hat{\mu}_2\|$ measures the comparability of the two samples. Several features require comment. If $n_2=p$, then $R_2=0$, $X_2'\hat{\mu}_2=Y_2$ and $n_1/(n_1+n_2)\cong 1$ so that this is essentially cross-validation. The opposite extreme of letting $n_1=p$ is discussed by Atkinson (1978). Even when $n_1=n_2$, the samples are not treated symmetrically and averaging $E_1(L_2)$ with $E_2(L_1)$ and other randomly selected subsamples, either on the additive or logarithmic scale, might be considered.

This approach seems to have a more secure basis than that proposed and in principle overcomes the problems which the author set out to deal with. Practical issues require further consideration.

Dr A. O'Hagan (University of Warwick, Coventry): Professor Aitkin suggests that we obtain the posterior distribution using the whole data vector $\mathbf{y}$, and then reuse all these data to effect model comparisons. Such a procedure is quite evidently non-coherent, as Professor Lindley and Professor Goldstein have demonstrated. Yet Professor Aitkin has persuasively argued some advantages of his method. I believe that those advantages may be obtained without abandoning coherence, as follows. Divide the data into two parts, $\mathbf{y}=(\mathbf{y}_1, \mathbf{y}_2)$, use $\mathbf{y}_1$ to update the prior distribution to a posterior density $\pi_j(\theta_j|\mathbf{y}_1)$, then use $\mathbf{y}_2$ to provide model comparisons. Like Professor Aitkin's posterior Bayes factor, we obtain a method that avoids Lindley's paradox, but for any chosen partition of the sample it is a genuine Bayesian method, and so coherent. The difference between this and the usual full Bayes factor is simply that we do not use $\mathbf{y}_1$ for model comparisons.

If a proportion p of the sample is used to update the prior distribution, yielding a 'partial' Bayes factor of C, then following the argument about the asymptote of Section 3 of the paper we find

$$-2\log C=(1-p)\{\lambda-\nu m(p)\}, \tag{2}$$

where $m(p)=-\log p/(1-p)$. Unlike Professor Aitkin's $m=\log 2$, this penalty factor $m(p)$ exceeds unity for all p. Akaike's $m=2$ corresponds to using a proportion $p=0.203$, or a fifth, of the sample for updating the prior distribution. The Smith and Spiegelhalter factor $m=1.5$ corresponds to $p=0.417$. However, the term $1-p$ in equation (2) represents the overall loss of information in not using the whole sample for model comparison and argues for smaller p rather than larger.

Dr L. I. Pettit (Goldsmiths' College, London): The examples presented by Professor Lindley and Professor Goldstein illustrate why we should not be using posterior Bayes factors. However, I would like to examine the relationship between the posterior Bayes factor and the Spiegelhalter and Smith (1982) Bayes factor for normal regression models. In Section 4 Professor Aitkin uses a standard non-informative prior. We adopt the Spiegelhalter and Smith limiting normal inverse χ^2 conjugate prior,

equation (4) of their paper, which leads to a Bayes factor which is invariant to scale changes in the dependent variable. It follows that

$$\bar{L}_j^{\mathrm{A}} = \frac{\Gamma(n)}{\Gamma(n/2)} 2^{(n+p_j)/2} \mathrm{RSS}_j^{-n/2}$$

and hence

$$A = \frac{\bar{L}_1^{\mathrm{A}}}{\bar{L}_2^{\mathrm{A}}} = 2^{(p_2-p_1)/2} \left(\frac{\mathrm{RSS}_1}{\mathrm{RSS}_2} \right)^{-n/2}.$$

If we compare this with equation (5) in Spiegelhalter and Smith (1982) we see that the posterior Bayes factor in this case is formally equivalent to taking

$$\frac{c_1}{c_2} = 2^{(p_1-p_2)/2} \left(\frac{|A_2^{\mathrm{T}} A_2|}{|A_1^{\mathrm{T}} A_1|} \right)^{-1/2}.$$

This may be interpreted as equivalent to a training sample, with the same design matrix as the data, which gives maximal support to the smaller model being given a Bayes factor of $2^{(p_2-p_1)/2}$. I prefer the Spiegelhalter and Smith procedure involving a *minimal* training sample which is *not data dependent* and which leads to a Bayes factor of unity with maximal support.

If Professor Aitkin is concerned about the effect of the prior he could adopt the suggestions of Lempers (1971) and Atkinson (1978) to set aside part of the data randomly as a real training sample to give posteriors. He would then not be using the same data twice. If there is not enough data to make this viable then I suggest that the methods of Spiegelhalter and Smith (1982) and Smith and Spiegelhalter (1981) are preferable.

Professor T. E. Raghunathan (University of Washington, Seattle): What do data tell us about the possibly many models that are justifiable from prior experience and the scientific context of the problem? Once the model is specified then the likelihood contains all the information about the parameter under that model. However, the uncertainty about the parameter naturally leads to uncertainty about this informal 'informational measure', i.e. the likelihood function. It is natural, at least from a Bayesian perspective, in model comparisons to study various characteristics of the posterior distribution of the likelihood (Raghunathan, 1984). The posterior mean of the distribution of the likelihood is one such characteristic. It is dangerous just to compare this mean across models and to stop at that. Given that two models have the same posterior Bayes factor, it does not mean that the data favour the two models equally. The whole posterior distribution of the likelihood should be considered before settling on one model.

An easy way to do this is to use the importance sampling to simulate the posterior distribution of the likelihood under different models. Let $L_j(\boldsymbol{\theta}^j | \mathbf{x})$ denote the likelihood function under the jth model. Let $p_j(\boldsymbol{\theta}^j | \mathbf{x})$ denote the posterior density of $\boldsymbol{\theta}^j$ under the jth model where $j = 1, 2, \ldots, M$ and $\boldsymbol{\theta}^j$ denote the parameters of the jth model. For simplicity let us assume that the parameters in these models represent similar characteristics (like location, scale, skewness, kurtosis etc.) of the distributions. This can be achieved in most cases by reparameterization.

Suppose that $g_q(\boldsymbol{\theta}^q)$ is a reasonable approximation for $p_q(\boldsymbol{\theta}^p | \mathbf{x})$ and it is easy to draw values from g_q for some q. For each drawn value θ_* from g_q attach an importance ratio $i_*^j = p_j(\theta_* | \mathbf{x})/g_q(\theta_*)$ up to a multiplicative constant. Then $(i_*^j, \theta_*, j = 1, 2, \ldots, M; * = 1, 2, \ldots, N)$ can be used to construct the posterior density of the likelihood under different models if N is sufficiently large. We need to draw values θ_* only once and to recompute the importance ratios when a new model is introduced.

It is possible that the posterior distribution of the likelihood under a subset of models will be similar in many aspects. Then as a Bayesian I see no alternative but to draw inference under these models and to combine them using the posterior probabilities of the models being correct (calculated usually under the discrete uniform prior) as weights. For instance, the prediction scheme that makes sense is to consider a weighted average of predictions with the posterior probabilities as weights. Should we not take this approach regardless and abandon all the model selection procedures?

Dr Trevor J. Sweeting (University of Surrey, Guildford): Results given by the posterior Bayes factor will usually seriously conflict with Bayesian analyses based on proper prior distributions for the nuisance parameters and, as already demonstrated by previous discussants, with common sense. I would therefore like to explore a little further the arguments leading to the definition of the posterior Bayes factor. The motivation seems to come from statements such as '. . . why should this objective prior assign weight to values of μ_2 which are untenable given the data . . .?' (in discussion of the example in Section 1.3). Continuing with this example, this leads to the intuitively plausible idea that we could first use the data $\mathbf{y}$ to revise the prior distribution of μ_2, and then use $\mathbf{y}$ to calculate the Bayes factor based on this revised distribution. The distribution used in the definition of the posterior Bayes factor A, however, is *not* the posterior distribution of μ_2, but the posterior distribution of μ_2 given that the data are actually *generated* from M_2. If the data are known to be generated from M_1, then they provide *no* information relating to the likely value of μ_2 under M_2. Thus values of μ_2 specified by the prior outside the main range of its posterior distribution (conditional on M_2) are only 'untenable given the data' if we know at the outset that $P(M_1|\mathbf{y})$ is small; but this is precisely the quantity that we are attempting to evaluate! Likewise, a large Bayes factor only implies 'strong rejection of the diffuse prior model M_2' (see Section 1.3) when we *know* that $P(M_1|\mathbf{y})$ is very small.

Let us pursue the author's ideas, using the correct posterior distribution of μ_2, and introduce $\bar{L}_2^{\mathrm{C}}$ (C for correct average) given by

$$\bar{L}_2^{\mathrm{C}} = \int L_2(\mu_2)\ \pi(\mu_2|\mathbf{y})\ \mathrm{d}\mu_2.$$

Then, since $\pi(\mu_2|\mathbf{y}, M_1) = \pi(\mu_2)$, we find that

$$\bar{L}_2^{\mathrm{C}} = \bar{L}_2^{\mathrm{B}}\ P(M_1|\mathbf{y}) + \bar{L}_2^{\mathrm{A}}\ P(M_2|\mathbf{y}).$$

Now let $C = p(\mathbf{y}|M_1)/\bar{L}_2^{\mathrm{C}}$. By equating posterior odds to prior odds $\times C$, we can obtain C as the positive solution of a quadratic equation which involves A, the Bayes factor B and the prior odds. I am not suggesting that anyone should *use* this, any more than the posterior Bayes factor A. The point is that, if we pursue the author's motivation a little more carefully, we see that it is just not possible to eliminate entirely the Bayes factor B from consideration. Furthermore, the prior odds ratio also becomes entangled with the 'Bayes factor'.

Professor H. Tong (University of Kent, Canterbury): Sir David Cox's comment has reminded me of the role of predictive distributions in modelling. Given (past) data $(y_1, y_2, \ldots, y_n)$, the predictive distribution of a (future) observation, y_{n+1}, may be defined as the mean of the data distribution $p(y_{n+1}|\boldsymbol{\theta})$ with respect to the posterior distribution $\pi(\boldsymbol{\theta}|y_1, \ldots, y_n)$. Akaike (1980) has discussed the use of predictive distributions in the context of what he called 'the commonsense approach to Bayesian statistics'.

Joe Whittaker (Lancaster University): In this interesting paper Murray Aitkin defines and evaluates the posterior Bayes factors. I should like his opinion on the following construction which indicates that these posterior factors may have an undesirable property for use in inference.

We wish to compare two models denoted by $j = 1, 2$ based on data y, in a situation complicated by an unknown nuisance parameter θ, where the distribution of Y, $p(y|j, \theta)$, depends on both j and θ. If θ has a known prior distribution (for example θ may index a random choice of experiment) and is distributed independently of j, we may integrate over θ to obtain the conditional distribution $p(y|j)$ of Y given just j. The argument is that

(a) if, in this distribution, Y is independent of j (so that an observation of Y cannot be informative about j) and
(b) if the posterior Bayes factors for j differ with differing values of y,

then the use of these posterior factors for inference is unwarranted.

To see that such an example exists consider the following example.

Suppose that the distribution of Y given j and θ is determined by the conditional probabilities in Table 4, where these given numbers satisfy

$$\alpha_1 + \beta_1 + \gamma_1 = 1 = \alpha_2 + \beta_2 + \gamma_2,$$

and $\alpha_1 + \alpha_2 = \frac{2}{3}$, $\beta_1 + \beta_2 = \frac{2}{3}$ and $\gamma_1 + \gamma_2 = \frac{2}{3}$. Further assume that the nuisance parameter θ takes the values 1 and 2 with probability $\frac{1}{2}$, irrespective of the value of j.

TABLE 4

	$j=1$		$j=2$	
	$\theta=1$	$\theta=2$	$\theta=1$	$\theta=2$
$y=1$	α_1	α_2	β_1	β_2
$y=2$	β_1	β_2	α_1	α_2
$y=3$	γ_1	γ_2	γ_1	γ_2

The distribution of Y given j must then be

$$p(y|j)=\tfrac{1}{3}$$

for $y=1, 2, 3$ whatever j, and so is uninformative about j. However, the posterior mean of the likelihood function $\bar{L}_j$, defined at the beginning of Section 2 of the paper, in this example becomes

$$\bar{L}_j=\int p(y|\theta, j)^2\, p(\theta)\, \mathrm{d}\theta/p(y),$$

so that the posterior Bayes factors for model $j=2$ against model $j=1$ are $(\beta_1^2+\beta_2^2)/(\alpha_1^2+\alpha_2^2)$, $(\alpha_1^2+\alpha_2^2)/(\beta_1^2+\beta_2^2)$ and 1, corresponding to observing $y=1$, $y=2$ and $y=3$ respectively. For a numerical example, set $\alpha_1=1/6$, $\beta_1=2/6$ and $\gamma_1=3/6$, so that the posterior Bayes factors are 8/10, 10/8 and 1 respectively.

This example is a version of the well-known counter-example to the assertion that, for random variables X, Y and Z, the independence of X from Y and of X from Z implies the independence of X from (Y, Z). For instance, see Whittaker (1990).

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

I thank the discussants for their comments. I shall deal first with comments on my examples, then with the counter-examples, and finally make some general comments.

First, the use of the posterior Bayes factor (PBF) is not restricted to the comparison of just two models. Professor Barnard and Sir David Cox refer in the 'simple/simple' models of Section 1.1 to the possibility of a third composite model, which I will take as M_3: $\mathbf{y} \sim N(\mu, \sigma^2)$ with μ unknown. The average likelihoods for the three models are proportional to $L_1=\exp(-2)=0.135$, $L_2=\exp(-4.5)=0.011$ and $\bar{L}_3=1/\sqrt{2}=0.707$. Model 3 is best supported and model 1 is tenable, with $L_1/\bar{L}_3=0.191$, but model 2 is firmly ruled out, with $L_2/\bar{L}_3=0.016$. Such comparisons can be extended directly to any number of models of any complexity.

The examples in Sections 4 and 5, and the discussion of the Lindley paradox, were meant to show the failure of the Bayes approach with conventional priors when these are used to calculate the Bayes factor, and the corresponding success of the PBF. Very few comments refer to these examples, which I take to establish that the analyses are not challenged. In the regression example, Dr Pettit proposes the improper limiting normal inverse χ^2 conjugate prior for $\boldsymbol{\beta}$ and σ, in which the prior for $\boldsymbol{\beta}$ depends on σ, so that the joint prior is *strongly* informative about σ, though this is hardly the kind of informative prior resulting from prior information, either subjective or objective. Nevertheless the PBF is almost unaffected, and in fact has the asymptotic value of Section 3 appropriate for likelihoods that are normal in the parameters. Dr Pettit finds this unsatisfactory because it apparently corresponds to an imaginary training sample which is not the imaginary sample imagined in the thought experiment by Spiegelhalter and Smith. The PBF does not require imaginary training samples, or any other thought experiments, and is not changed in value by a thought experiment, performed or not performed.

Professor Dawid asks whether the posterior mean of the likelihood in N in the binomial example would be changed by a prior independent in N and $\psi=Np$. The answer is no, as described in the paper and in more detail in Aitkin and Stasinopoulos (1989). The same likelihood in N is obtained, with the two-parameter likelihood almost orthogonal in N and ψ. Dr Fearn points out that the integrated likelihood with respect to the beta prior does not have a finite sum over N unless $a>1$. Thus a uniform prior on N and p will give an improper posterior mass function for N unless N has a finite maximum value. This was noted by Draper and Guttman (1971) who set a finite limit *a priori* on N. However, this does not affect the shape of the integrated likelihood, or the essential difference between the prior and posterior means over p of the likelihood.

Professor R. L. Smith asks whether this example is over-simple, and whether a more complex nuisance parameter structure might throw more light on the different approaches. The analysis of more complex capture–recapture models with multiple nuisance parameters will be reported elsewhere; the present example is simple, but it shows clearly the extreme differences possible between prior and posterior means of the likelihood. It appears to Professor Sprott that the conditional likelihood loses 'somewhat less' information than the profile or average likelihoods, though he does not say why. The conditional likelihood loses the information in the likelihood from the conditioning variable T, which has the $b(Nr, p)$ distribution. The average likelihood recovers the 'average' information about N in T by integrating out p from this marginal likelihood with respect to the posterior distribution of p given N. There is not much information in this marginal likelihood, but there is some, and the average likelihood loses less than the conditional likelihood which ignores T.

Now to the counter-examples: Professor Lindley gives an example in which (to use his illustration) the PBF very weakly supports (by a factor of 1.5) the hypothesis 'tall man' over 'tall woman', very weakly supports by the same factor 'short man' over 'short woman', but supports even more weakly the composite hypothesis 'woman' over 'man'. None of these PBFs is conclusive sample evidence, but Professor Lindley says that this is ridiculous. He does not give the value of the ordinary Bayes factor for the composite hypothesis, which is 0.64, compared with 0.77 for the PBF. Since the PBFs for the simple hypotheses are the same as the prior Bayes factors (since both are just the likelihood ratio) Professor Lindley's argument would make the ordinary Bayes factor conclusions even more ridiculous. Perhaps Lindley realizes this, for he seems to reject Bayes factor as well, claiming that only posterior odds ratios can serve as inferential summaries. But posterior odds ratios are not inconsistent with the sample information in likelihood ratios or Bayes factors—they simply incorporate the given non-sample prior model information. In the example the prior odds on tall man relative to tall woman are 1/9, giving posterior odds of 1/6, and those on short man to short woman are 9/1, giving posterior odds of 13.5. The prior odds on man to woman are 1/1, and the posterior odds are 0.64. The only difference that the PBF makes is to change the last value to 0.77. If these inversions of weak implications from sample evidence in this example are Professor Lindley's only criterion for the performance of inferential methods, then the ordinary Bayes factor fares worse than the PBF. It is this criterion, not the PBF result, that is ridiculous.

Professor Goldstein and Dr Whittaker give examples of the same form: $f_j(\mathbf{y} \mid \boldsymbol{\theta}_j)$ are different, but $\pi_j(\boldsymbol{\theta}_j) = \pi(\boldsymbol{\theta}_j)$ are the same, and so are the marginal densities $f_j(\mathbf{y}) = f(\mathbf{y})$. Given one observation $\mathbf{y}$, the ordinary Bayes factor based on the marginal distributions is 1, but the PBF is not. Professor Goldstein makes the PBF 1000 in his example. Professor Akaike gives a similar example in which the marginal distribution of a single observation actually degenerates, if the diffuse prior is taken literally, and the prior Bayes factor can take any value. The common feature of all these examples is that *only one observation is allowed* from $f_j(\mathbf{y} \mid \boldsymbol{\theta}_j)$: with more than one observation the marginal distributions of $\mathbf{y}$ are different. In Professor Goldstein's example a second observation immediately identifies the correct model. These examples are slight extensions of the well-known examples of so-called 'paradoxes' of likelihood inference with one observation from a two-parameter distribution. Professor Barnard warns against settling important issues on the basis of single data sets. This warning seems even more cogent when the data set consists of just one observation.

Now to general issues: Dr Davison refers to the recent paper by Pace and Salvan (1990) giving general conditional tests for separate families problems. The same conditional test is given for comparing the log-normal with the exponential distribution and the log-normal with the gamma distribution. This is clearly unreasonable and I shall report elsewhere the PBF approach to this problem. The comparison of different link functions for the same binomial data is an excellent example of the value of the PBF approach. If the sample sizes are reasonably large so that the likelihoods in the parameters can be taken as normal, the deviances for the two models can be compared directly as in Section 3. With the same number of parameters in the models, the deviance difference is 7.78, so the PBF is $\exp\{-\frac{1}{2}(7.78)\} = 0.020$. The complementary log–log link is strongly supported over the logit. If the sample sizes are not large more accurate approximations to the average likelihood using higher derivatives of the sample log-likelihood are required. No bootstrap sampling is needed. Details for the logit model are given in Aitkin (1990).

Sir David Cox gives an interpretation of the PBF in terms of the principle of temporal coherence, which it violates. Whatever the status of this principle, Sir David proposes calibration as the ultimate test, i.e. the performance of the PBF under hypothetical replications. This is given in Section 3, which incidentally also shows the unsatisfactory behaviour of the Akaike information criterion ($m = 2$), with

its very heavy penalty against complex models. Models with the number of nuisance parameters of the same order as the sample size cause difficulties in all likelihood approaches, but these difficulties arise because of the failure to model the nuisance parameters. *Ad hoc* conditional or marginal likelihood approaches to such models work in some cases, namely those for which conditioning or marginalizing is equivalent to integrating out the nuisance parameters with respect to a distribution. It seems simpler to approach the problem directly and to specify a variance component model for such situations at the beginning, rather than to introduce it implicitly later. The PBF approach then applies directly.

Professor A. F. M. Smith says, if I understand him correctly, that model choice problems are more complicated than a simple Bayes factor calculation. Perhaps it would clarify matters if the question of *evidence* for the competing models were separated from the question of *decision* to choose one of the models, and the resulting utility. The posterior mean of the likelihood provides the sample *evidence* for the model, and the PBF provides the weight of sample evidence for one model over another. This is the important contribution of the data analysis, and is separate from the question of possible actions to be taken given the evidence, and their consequences.

Several discussants (Dr Cuzick, Dr Pettit and Dr O'Hagan) suggest that we use part of the sample to obtain the posterior density of θ_j, and the rest to compare the models given these independent posterior densities. Dr O'Hagan's proposal is the most formal. He also shows most clearly the difficulties in such arbitrary sample divisions: different splits of different sizes will produce different 'partial' Bayes factors. On what non-arbitrary basis will the actual division be made? Such proposals are made to avoid 'using the data twice'. Suppose that an experimenter wants the posterior density of θ_1, given a model M_1. She generates data $\mathbf{y}$ and obtains $\pi_1(\theta_1 | \mathbf{y})$, based on prior $\pi_1(\theta_1)$. A second experimenter analyses the same data using a different model M_2; he obtains $\pi_2(\theta_2 | \mathbf{y})$, using prior $\pi_2(\theta_2)$. Later the experimenters wish to compare their models—which is better supported by the data? The PBF is the relevant comparison: surely we do not require that the experimenters return to their prior densities for θ_j, given their information about the *particular value of* θ_j *that actually applied in this experiment*, nor that they generate independent data from a new experiment, to settle the issue of which model is better supported by the *previous* experiment.

Dr Fearn misrepresents my references to subjective priors. I have no objection to the subjective Bayesian approach. If one has subjective (individual) priors $\pi_j(\theta_j)$, whether these are based on experimental data or are agreed assumptions for the client, then the calculation of the (prior) Bayes factor is a formal probability result. The question is not its correctness, but its appropriateness for inference. Bayesians point out the unreasonableness of Neyman–Pearson post-data averaging over the sample space, with respect to samples which might have been observed, but were not. But the same type of averaging is used in the ordinary Bayes factor. Before the data are obtained, $\pi_j(\theta_j)$ is the distribution of θ_j, but, once the sample value of θ_j is generated and the data $\mathbf{y}$ observed, averaging over values of θ_j which might have generated $\mathbf{y}$, but did not, is no longer appropriate. As Professor Dawid notes, with extensive data we are effectively observing $\theta_j = \hat{\theta}_j$. The PBF uses this information, averaging over the values of θ_j near $\hat{\theta}_j$ which *could* have generated the data, not those which could not. Professor Dawid underlines the arbitrariness of the ordinary prior mean of the likelihood, which is changed in large samples by the factor $\pi_j(\hat{\theta}_j)/\pi_j^*(\hat{\theta}_j)$ by a change of prior from $\pi_j(\theta)$ to $\pi_j^*(\theta)$. The PBF is invariant to this change. Prior Bayes factors cannot use diffuse priors; PBFs can. The claims of 'absurd' or 'dangerous' results, from 'using the data twice', are philosophical statements that are not supported by evidence.

In conclusion, PBFs stand as a general framework or paradigm for model comparisons, i.e. for parametric statistical inference. If a Bayes–non-Bayes compromise is necessary, they provide one.

REFERENCES IN THE DISCUSSION

Aitkin, M. (1990) Model choice in contingency table analysis using the posterior Bayes factor. In *Proc. 5th Int. Workshop Statistical Models, Toulouse, July 1st–8th*. Submitted to *Comput. Statist. Data Anal.*

Aitkin, M. A. and Stasinopoulos, M. (1989) Likelihood analysis of a binomial sample size problem. In *Contributions to Probability and Statistics* (eds L. J. Gleser, M. D. Perlman, S. J. Press and A. R. Sampson). New York: Springer.

Akaike, H. (1980) Likelihood and the Bayes procedure. In *Bayesian Statistics* (eds J. M. Bernado, M. H. DeGroot, D. V. Lindley and A. F. M. Smith). Valencia: University of Valencia Press.

——— (1985) Prediction and entropy. In *A Celebration of Statistics* (eds A. C. Atkinson and S. E. Fienberg), pp. 1–24. New York: Springer.

Atkinson, A. C. (1978) Posterior probabilities for choosing a regression model. *Biometrika*, **65**, 39–48.

Bliss, C. I. (1935) The calculation of the dosage-mortality curve. *Ann. Appl. Biol.*, **22**, 134–167.

Cox, D. R. (1961) Tests of separate families of hypotheses. In *Proc. 4th Berkeley Symp. Mathematics, Probability and Statistics*, vol. 1, pp. 105–123. Berkeley: University of California Press.

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

Draper, N. and Guttman, I. (1971) Bayesian estimation of the binomial parameter. *Technometrics*, **13**, 667–673.

Lempers, F. B. (1971) *Posterior Probabilities of Alternative Linear Models*. Rotterdam: Rotterdam University Press.

Pace, L. and Salvan, A. (1990) Best conditional tests for separate families of hypotheses. *J. R. Statist. Soc.* B, **52**, 125–134.

Pitman, E. J. G. (1965) *Some Remarks on Statistical Inference: Bernoulli, Bayes, Laplace*, pp. 209–216. New York: Springer.

Pregibon, D. (1980) Goodness of link tests for generalized linear models. *Appl. Statist.*, **29**, 15–24.

Raghunathan, T. E. (1984) A new model selection criterion. *Research Report S-96*. Department of Statistics, Harvard University, Cambridge.

Shafer, G. (1982) Lindley's paradox. *J. Am. Statist. Ass.*, **77**, 325–351.

Smith, A. F. M. and Spiegelhalter, D. J. (1981) Bayesian approaches to multivariate structure. In *Interpreting Multivariate Data* (ed. V. Barnett), pp. 335–348. Chichester: Wiley.

Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factors for linear and log-linear models with vague prior information. *J. R. Statist. Soc.* B, **44**, 377–387.

Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Wiley: Chichester.