Statistics and Computing **15**: 217–230, 2005
© 2005 Springer Science + Business Media, Inc. *Manufactured in The Netherlands.*

Bayesian point null hypothesis testing via the posterior likelihood ratio

MURRAY AITKIN, RICHARD J. BOYS and TOM CHADWICK

School of Mathematics and Statistics, University of Newcastle, UK

Received December 2003 and accepted February 2005

This paper gives an exposition of the use of the posterior likelihood ratio for testing point null hypotheses in a fully Bayesian framework. Connections between the frequentist P-value and the posterior distribution of the likelihood ratio are used to interpret and calibrate P-values in a Bayesian context, and examples are given to show the use of simple posterior simulation methods to provide Bayesian tests of common hypotheses.

Keywords: likelihood ratio, point null hypothesis, posterior distribution, Bayes factor

1. Introduction

Neyman-Pearson or frequentist inference and Bayes inference are most clearly differentiated by their approaches to point null hypothesis testing. With very large samples, the frequentist and Bayesian conclusions from a classical test of significance for a point null hypothesis can be contradictory, with a small frequentist P-value casting serious doubt on the null hypothesis, but a large Bayes factor or Bayesian Information Criterion (BIC) in favour of the null hypothesis.

A Bayesian approach by Dempster (1974, 1997) through the likelihood ratio between the null and alternative hypotheses, extended by Aitkin (1997), provides a different evaluation of the point null hypothesis, one in which frequentist and Bayesian conclusions are much closer. The discussion in Aitkin (1997) is restricted, in both the computational approach and the range of examples considered, and in this paper we extend both, by using simple posterior simulation methods for intractable integrations, and a range of examples of the standard frequentist hypothesis testing kind, to illustrate the broad generality of the approach. We also provide the usual Bayes factor comparisons where these are possible, to illustrate the differences in conclusions.

Section 2 of the paper gives a simple binomial example with no nuisance parameters to illustrate Dempster's original approach, and Section 3 gives the general result from Aitkin (1997). Section 4 illustrates the general approach with nuisance parameters using the two-parameter normal model, discussed analytically in Aitkin (1997) and by simulation methods in Chadwick (2002), and illustrates the role of posterior simulation in providing very simple solutions to the sometimes complex distributional problems of the likelihood ratio. Section 5 extends the "nested model" approach to encompassing models, and Section 6 shows that for the normal multiple regression model, straightforward posterior simulation methods give Bayesian analogues to backward elimination in frequentist theory. Section 7 illustrates the importance of parametrization with the binomial (N, p) model which has been considered by many authors. Section 8 discusses the Bayesian analysis of the 2×2 contingency table with a well-known example from a randomized clinical trial. Section 9 gives concluding discussion.

2. Simple null hypotheses

Consider the simple example due to Stone (1997) in the discussion of Aitkin (1997). A physicist runs a particle-counting experiment to identify the proportion θ of a certain type of particle. He has a well-defined scientific (null) hypothesis H_1 that $\theta = 0.2(= \theta_1)$ precisely. There is no specific alternative hypothesis, only the general H_2, that $\theta \neq \theta_1$. He counts $n = 527, 135$ particles and finds $r = 106, 298$ of the specified type. What is the strength of the evidence against H_1?

The binomial likelihood function

$$L(\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r} \approx L(\hat{\theta})\exp\left\{-\frac{(\theta-\hat{\theta})^2}{2SE(\hat{\theta})^2}\right\}$$

is maximized at $\theta = \hat{\theta} = 0.201652$ with standard error $SE(\hat{\theta}) = 0.0005526$. The standardized departure from the null hypothesis is

$$Z_1 = |\theta_1 - \hat{\theta}|/SE(\hat{\theta}) = 0.001652/0.0005526 = 2.9895,$$

with a two-sided P-value of 0.0028, strong evidence against the null hypothesis. The maximized likelihood ratio is $L(\theta_1)/L(\hat{\theta}) = 0.01146$.

The physicist uses the uniform prior $\pi(\theta) = 1$ on $0 < \theta < 1$ under the alternative hypothesis, and computes the Bayes factor

$$B = L(\theta_1) \Big/ \int_0^1 L(\theta)\pi(\theta)\,\mathrm{d}\theta.$$

The denominator is

$$\begin{aligned} L^B &= \binom{n}{r} \int_0^1 \theta^r (1-\theta)^{n-r} \mathrm{d}\theta \\ &= \binom{n}{r} B(r+1, n-r+1) \\ &\approx L(\hat{\theta}) \int_0^1 \exp\left\{ -\frac{(\theta - \hat{\theta})^2}{2SE(\hat{\theta})^2} \right\} \mathrm{d}\theta \\ &= \sqrt{2\pi} SE(\hat{\theta}) L(\hat{\theta}) \\ &= L(\hat{\theta})/f(\hat{\theta}), \end{aligned}$$

where $B(r+1, n-r+1)$ is the complete Beta function, and $f(\hat{\theta})$ is the normal posterior density $N(\hat{\theta}, SE(\hat{\theta})^2)$ of θ evaluated at the mean $\hat{\theta}$; since the sample size is so large the actual posterior Beta density is very nearly normal.

The Bayes factor is thus

$$\begin{aligned} B &= L(\theta_1)/L^B \\ &= f(\hat{\theta}) \cdot L(\theta_1)/L(\hat{\theta}), \end{aligned}$$

a simple multiple of the maximized likelihood ratio. In this example the multiplier is

$$f(\hat{\theta}) = \frac{1}{\sqrt{2\pi} SE(\hat{\theta})} = \frac{1}{0.0013851} = 721.937,$$

giving the Bayes factor

$$B = 721.937 \cdot 0.01146 = 8.27,$$

indicating evidence *in favour of* the null hypothesis. Thus the P-value and Bayes factor are in clear conflict. However the posterior distribution of θ is *not* in conflict with the P-value, since the posterior probability that $\theta > 0.2$ is

$$\Pr[\theta > 0.2 \mid \mathbf{y}] = \Phi(2.9895) = 0.9986 = 1 - P/2.$$

Any Bayesian using the uniform prior must have a very strong posterior belief that the true value of θ is larger than 0.2. Equivalently, the 99% equal-tailed Bayesian credible interval for θ is

$$\theta \in \hat{\theta} \pm 2.576 SE(\hat{\theta}) = (0.20023, 0.20308)$$

which is numerically identical to the 99% frequentist confidence interval, and excludes θ_1.

This example illustrates one of the difficulties of Bayesian analysis, that one may have to choose between "hypothesis testing" and "estimation" approaches when these are in conflict. Kass and Greenhouse (1989) and Kass and Raftery (1995) give clear statements of the difference between these approaches.

In his 1974 conference paper, Dempster considered the *likelihood ratio* between the null and alternative hypothesis models:

$$LR(\theta) = L(\theta_1)/L(\theta).$$

Since θ is unknown under the alternative, $L(\theta)$ is also unknown, but is a function of θ and so, given the data, it has a posterior distribution $\pi[L(\theta) \mid \mathbf{y}]$ which can be derived from that of θ, $\pi(\theta \mid \mathbf{y})$. Since $L(\theta_1)$ is a known number, the likelihood ratio also has a posterior distribution, $\pi[LR(\theta) \mid \mathbf{y}]$. We may therefore find its posterior percentiles, and so can find

$$\Pr[LR(\theta) < 0.1 \mid \mathbf{y}]$$

for example. A likelihood ratio of 0.1 between fully specified simple hypotheses would be quite strong sample evidence against the "numerator" hypothesis; a posterior probability of 0.9 or more that the likelihood ratio was less than 0.1 would similarly be quite strong evidence against this hypothesis, and in general the posterior distribution of the likelihood ratio can be used to assess the strength of the evidence against (or *in favour of*) the null hypothesis.

In the Stone example, approximating the binomial likelihoods by the corresponding normal likelihoods gives the likelihood ratio as

$$LR(\theta \mid \mathbf{y}) \approx \frac{\phi([\theta_1 - \hat{\theta}]/SE[\hat{\theta}])}{\phi([\theta - \hat{\theta}]/SE[\hat{\theta}])},$$

or in terms of the "deviance" $D(\theta)$,

$$D(\theta) = -2 \log LR(\theta \mid \mathbf{y}) = Z_1^2 - Z^2,$$

where

$$Z = \frac{\theta - \hat{\theta}}{SE(\hat{\theta})}$$

Here Z has a posterior $N(0, 1)$ distribution, and Z_1 is Z with θ replaced by θ_1. Now $Z_1 = 2.9895$ and so

$$\begin{aligned} \Pr[LR(\theta) < 0.1 \mid \mathbf{y}] &= \Pr[D(\theta) > 4.605 \mid \mathbf{y}] \\ &= \Pr\left[Z^2 < Z_1^2 - 4.605 \mid \mathbf{y}\right] \\ &= \Pr\left[\chi_1^2 < 4.331\right] = 0.9626, \end{aligned}$$

while

$$\begin{aligned} \Pr[LR(\theta) < 1 \mid \mathbf{y}] &= \Pr[D(\theta) > 0 \mid \mathbf{y}] \\ &= \Pr\left[Z^2 < Z_1^2 \mid \mathbf{y}\right] \\ &= \Pr\left[\chi_1^2 < 2.9895^2\right] = 0.9972 \\ &= 1 - P \end{aligned}$$

where P is the frequentist P-value from the likelihood ratio test. This illustrates Dempster's fundamental result (which he gave for a p-parameter simple null hypothesis against a general alternative) that, with normal likelihoods and flat priors, *the P-value is equal to the posterior probability that the likelihood ratio is* *greater than 1, that is, that the data support the null hypothesis more strongly than the alternative.*

The above form of Bayesian analysis comes to the same conclusion as the frequentist analysis, that there is strong sample evidence against the null hypothesis. Why does the Bayes factor point in the opposite direction? One point which does not seem to have been noticed is that we intended to compare the null binomial model with "some other" binomial model, unspecified. But the binomial distribution integrated over the flat prior gives a uniform distribution with mass $1/(n+1)$ at the $n+1$ possible values of r. The Bayes factor is comparing the null binomial model with the uniform distribution for r. This was surely not our intention, since no binomial distribution is uniform. The integration has taken us outside the family of binomial distributions within which we wanted to compare the null model.

The general Bayesian opposition to the use of averaging over the sample space in frequentist testing is weakened in this approach, since the P-value has a fully Bayesian interpretation, though it might be argued that the P-value still overstates the strength of evidence against the null hypothesis since it refers only to a preference for the null hypothesis over the alternative. However we may compute any percentiles of the posterior distribution of the likelihood ratio; in the example above, there is strong posterior evidence that the likelihood ratio is less than 0.1, not just that it is less than 1. The information in the full posterior distribution of the likelihood ratio provides a richer analysis than just the frequentist P-value, and also calibrates the P-value from a Bayesian perspective.

This approach was extended to models with nuisance parameters in Aitkin (1997).

3. General point null hypothesis testing problems

We deal with a family of models M, determined by a probability model $f(y \mid \eta)$ depending on a vector-valued parameter $\eta^T = (\theta^T, \phi^T)$. It is helpful to consider the probability model in the context of a large but finite population of N members, in which θ and ϕ represent population properties like the mean and variance, which could be determined exactly by a census of the population, though we have only a sample of n values.

Some Bayesians (see for example Geisser 1993) deny the relevance of parameters, insisting that only random variables have a real existence, but most statisticians regard them as convenient model components, and survey sampling statisticians take finite population parameters as *the* essential feature for statistical inference.

The likelihood for the given data $\mathbf{y}$ is

$$L(\theta, \phi) = f(\mathbf{y} \mid \theta, \phi).$$

In our analysis there are *true values* of θ and ϕ; the prior distribution for these parameters represents our uncertainty about these true values.

We consider a null hypothesis H_1 which specifies the value θ_1 of θ, while ϕ is unspecified. An alternative hypothesis H_2 specifies *either* that θ is completely unspecified, *or* that θ has a different specified value θ_2. In either case ϕ is unspecified.

The joint prior distribution for θ and ϕ is $\pi(\theta, \phi)$. This may be proper or improper; we make particular use of flat priors to represent diffuse prior information, with the aim, following Berger and Bernardo (1989), of developing a *reference prior analysis* of these hypothesis testing problems.

The first class of testing problems with an unspecified alternative was considered by Aitkin (1997), and we review the results briefly. If the true value of ϕ, and the true value of θ under the alternative H_2 were known, the likelihood ratio between the hypotheses would provide the data evidence for H_1 against H_2; we write the likelihood ratio as

$$LR = LR(\theta, \phi) = L(\theta_1, \phi)/L(\theta, \phi),$$

where the dependence of LR on the data $\mathbf{y}$ and the known value θ_1 are suppressed, and the values of θ and ϕ are understood to be the true values.

In this approach the inferential function LR is the likelihood ratio defined by a *section* through the likelihood at the true value of the nuisance parameter ϕ, evaluated at the null hypothesis value θ_1 and at the true value of θ. Though the true values of ϕ and θ are unknown, their posterior distribution is known:

$$\pi(\theta, \phi \mid \mathbf{y}) = \frac{L(\theta, \phi) \cdot \pi(\theta, \phi)}{\int L(\theta, \phi) \cdot \pi(\theta, \phi) \mathrm{d}\theta \mathrm{d}\phi}$$

and therefore so is the posterior distribution of LR. In particular, we may evaluate the posterior probability

$$\Pr[LR < k \mid \mathbf{y}]$$

for any specified k, like 0.1 or 0.01. It will be convenient to evaluate such probabilities through the posterior distribution of the "true deviance" $D = -2 \log LR$.

For normal likelihoods with flat priors, Aitkin (1997) showed that the result due to Dempster, for a p-parameter simple null hypothesis, with normal likelihoods and flat priors:

$$P[LR < k \mid y] = F_p\left[F_p^{-1}(1 - P) + 2 \ln k\right]$$

applies also to nuisance-parameter models (where p is the dimension of θ, P is the frequentist P-value from the likelihood ratio test, and $F_p(x)$ is the cdf of the χ_p^2 distribution). In particular, for $k = 1$,

$$\Pr[LR < 1 \mid y] = 1 - P,$$

so again the P-value is the posterior probability that the likelihood ratio is greater than 1, that is that the null hypothesis is better supported than the alternative.

In finite samples with non-normal likelihoods these are asymptotic results and hence are insufficient. We now discuss simulation approaches to obtaining the posterior distribution of LR or D, in the context of the two-parameter normal model.

4. Example—the two-parameter normal model

The model for data y is $N(\mu, \sigma^2)$ with σ unknown. A null hypothesis H_1 specifies $\mu = \mu_1 = 0$; the alternative H_2 is general. A random sample of $n = 25$ observations gives $\bar{y} = 0.4$, and unbiased variance estimate $s^2 = 1$. What is the strength of the evidence against H_1 in favour of H_2? The t-statistic is $t = \sqrt{25} \cdot 0.4/1 = 2.0$, with a two-sided P-value of 0.057 from the t_{24} distribution.

The likelihood function is

$$L(\mu, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right\},$$

and given independent diffuse priors on μ and $\log\sigma$, the conditional posterior distribution of $\mu \mid \sigma$ is $N(\hat{\mu}, \sigma^2/n)$, and the marginal posterior distribution of s^2/σ^2 is $\chi^2_{n-1}/(n-1)$. The true deviance is

$$\begin{aligned} D = -2\log\left\{\frac{L(\mu_1, \sigma)}{L(\mu, \sigma)}\right\} &= \frac{n}{\sigma^2}[(\bar{y}-\mu_1)^2 - (\bar{y}-\mu)^2] \\ &= \frac{n(\bar{y}-\mu_1)^2}{s^2} \cdot \frac{s^2}{\sigma^2} - \frac{n(\bar{y}-\mu)^2}{\sigma^2} \\ &= t^2 \cdot W - Z^2 \end{aligned}$$

where Z has a posterior $N(0, 1)$ distribution independently of $W = s^2/\sigma^2$ which has the $\chi^2_{n-1}/(n-1)$ distribution. It follows immediately that

$$\begin{aligned} \Pr[LR < 1 \mid y] &= \Pr[D > 0 \mid y] \\ &= \Pr[Z^2/W < t^2 \mid y] = 1 - P, \end{aligned}$$

where P is the P-value 0.057 from the t_{n-1} distribution.

For other values of k the distribution of D has no simple analytic form, so we simulate it by generating N times a random value of W and an independent random value of Z, and computing the value of $D = t^2W - Z^2$ for the observed t. Figure 1 shows the posterior cdf of D from $N = 10,000$ simulations.

The simulated probability that $D > -2\log 1 = 0$ is 0.945, with simulation standard error 0.0023, in close agreement with the known value of $1 - P$ of 0.943, and the simulated probability that $D > -2\log 0.1$ is 0.157, with standard error 0.0036.

The probability that the $LR < 0.1$ is quite low—there is no convincing evidence against the null hypothesis. This is of course to be expected since the P-value does not reach even conventional levels.

The Bayes factor cannot be computed here due to the diffuse prior on μ.

5. An encompassing model

We now extend these results to the comparison of two specified values of θ, following Chadwick (2002). We illustrate with the two-parameter normal model.

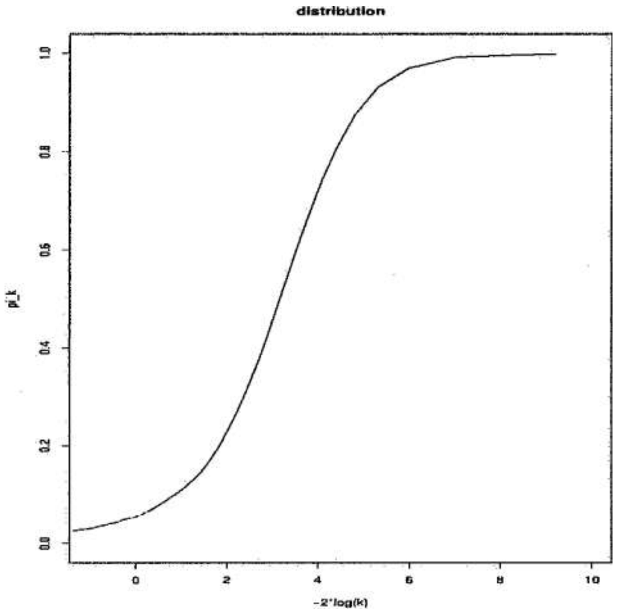

Fig. 1. *Posterior distribution of D*

The model and data are as in the previous example, but there are now *two* point hypotheses, $H_1 : \mu = \mu_1 = 0$ and $H_2 : \mu = \mu_2 = 1$. What is the strength of evidence against H_1 in favour of H_2?

The t-statistic regarding H_1 as the "null" hypothesis is $t_1 = 2.0$ as before, while that regarding H_2 as the null is $t_2 = 3.0$. Clearly H_1 is better supported. The maximized likelihood ratio is

$$\begin{aligned} LR_{\max} &= \frac{L(\mu_1, \widehat{\sigma_1(\mu_1)})}{L(\mu_2, \widehat{\sigma_2(\mu_2)})} \\ &= \left[\left(1 + \frac{t_2^2}{n-1}\right)\Big/\left(1 + \frac{t_1^2}{n-1}\right)\right]^{-n/2} = 7.80, \end{aligned}$$

where

$$\widehat{\sigma_j^2(\mu_j)} = [(n-1)s^2 + n(\bar{y}-\mu_j)^2]/n, \quad j = 1, 2.$$

The true deviance is now

$$\begin{aligned} D &= -2\log\left[\frac{L(\mu_1, \sigma)}{L(\mu_2, \sigma)}\right] \\ &= \frac{n}{\sigma^2}[(\bar{y}-\mu_1)^2 - (\bar{y}-\mu_2)^2] \\ &= \frac{s^2}{\sigma^2}\left[t_1^2 - t_2^2\right]. \end{aligned}$$

The only nuisance parameter is σ, and as before s^2/σ^2 has the $\chi^2_{n-1}/(n-1)$ marginal posterior distribution. So for the upper tail,

$$\begin{aligned} \Pr[LR > k \mid \mathbf{y}] &= \Pr[D < -2\log k \mid \mathbf{y}] \\ &= \Pr\left[\chi^2_{n-1} > -2(n-1)\log k\big/\left(t_1^2 - t_2^2\right)\right]. \end{aligned}$$

We drop the conditioning on $\mathbf{y}$ for notational convenience. For $k = 1$ we have immediately

$$\Pr[LR > 1] = 1,$$

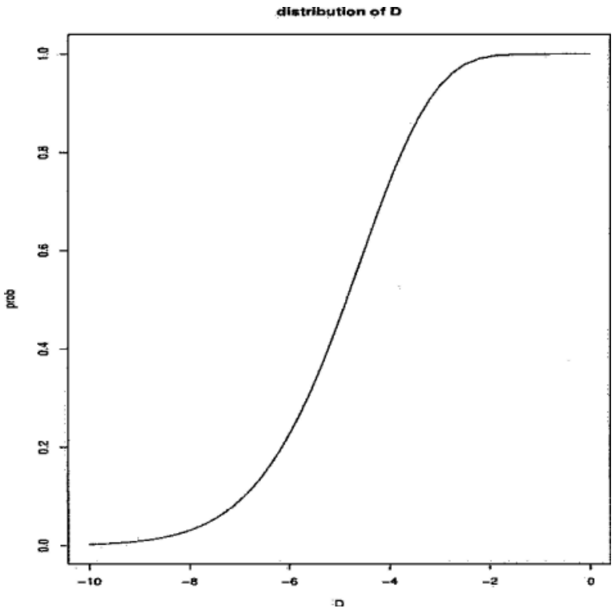

Fig. 2. *Posterior distribution of D*

and for $k = 10$,

$$\Pr[LR > 10] = \Pr\left[\chi^2_{24} > 24 \cdot 4.605/5 = 22.10\right] = 0.573.$$

So H_1 is certainly better supported, but the evidence in favour of H_1 is not very strong—the chance that the true *LR* is greater than 10 is not much over 50%.

The posterior distributions of D and of *LR* are graphed in Figs. 2 and 3. It is of interest that

$$\Pr[LR > 7.80] = \Pr\left[\chi^2_{24} > 19.72\right] = 0.713,$$

so the maximized likelihood ratio is at the 29-th percentile of the posterior distribution of the true *LR*; in this case the maximized likelihood ratio appears to *understate* the strength of evidence.

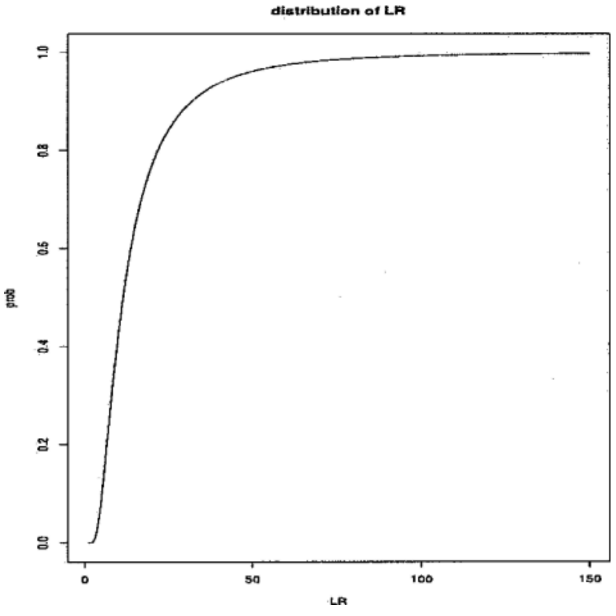

Fig. 3. *Posterior distribution of LR*

Since the parameter space has the same dimension under both models, the Bayes factor can be computed with the same diffuse prior on $\log \sigma$, assuming that the same (arbitrary) prior constant is used. The integrated likelihood over σ is

$$L^B(\mu) = \frac{1}{2(2\pi)^{n/2}}\Gamma(n/2)\left[\frac{2}{(n-1)s^2 + n(\bar{y} - \mu)^2}\right]^{n/2}$$

and hence

$$\begin{aligned} BF &= \left[\frac{1 + t_2^2\big/(n-1)}{1 + t_1^2\big/(n-1)}\right]^{-n/2} \\ &= LR_{\max} = 7.80. \end{aligned}$$

Thus the Bayes factor gives the same understatement of strength of evidence as the maximized likelihood ratio in this example.

We turn now to more complex examples.

6. Multiple regression

Consider the normal regression model with n observations on a response Y and a $p + 1$-vector x of p explanatory variables and 1, with the model

$$Y \mid \mathrm{x} \sim N(\mu, \sigma^2), \quad \mu = \beta^T \mathrm{x}.$$

Our aim is to assess the important variables through a series of model comparisons expressed in terms of partitions $\beta = (\beta_j, \gamma_j)$ and hypotheses $H_j : \gamma_j = 0$ in model M_j. Paralleling backward elimination methods in frequentist theory, we examine the strength of evidence for the various models in the backward elimination sequence. The approach does not depend on the choice of variables—any sub-model can be compared with the full model in the same way.

We use a well-known data set—the gas consumption data of Henderson and Velleman (1981), which has observations on the fuel consumption, expressed in miles per (US) gallon, of 32 cars with 10 design variables on the engine and transmission.

We follow the backward elimination analysis of Aitkin, *et al.* (1989, p. 140), using log(mpg) as the response variable and the explanatory variables, listed in order of backward elimation: c, drat, s, t, log(disp), cb, g, and log(hp). The corresponding backward elimination t-statistics for these variables are −0.082, −0.320, −0.340, −0.461, −0.723, −1.070, 1.052 and −4.372. Elimination ceases with a final model using log(wt) and log(hp).

In the backward elimination sequence the sums of squares of eliminated variables are pooled with the error sum of squares from the full model, so the degrees of freedom of the t-statistics change at each step.

In the analysis below we maintain the posterior distribution of β and σ from the full model, as this gives a more realistic picture of the information about many parameters from the small sample.

The likelihood for the full model is

$$\begin{aligned} L(\beta, \sigma) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)\right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}[RSS \right. \\ &\qquad \left. + (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})]\right\} \end{aligned}$$

where X is the $n \times (p + 1)$ design matrix,

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}, \quad RSS = (\mathbf{y} - X\hat{\beta})^T(\mathbf{y} - X\hat{\beta}).$$

We take flat priors on β and $\log\sigma$ to give the usual joint posterior distribution, with

$$\beta \mid \mathbf{y}, \sigma \sim N(\hat{\beta}, \sigma^2(X^T X)^{-1}), \quad RSS/\sigma^2 \mid \mathbf{y} \sim \chi^2_{n-p-1}.$$

Consider the null hypothesis $H_j : \beta^T = (\beta_j^T, 0^T)$ for some partition. The true likelihood ratio and deviance are

$$\begin{aligned} LR &= \frac{L(\beta_j, 0, \sigma)}{L(\beta, \sigma)} \\ D &= \frac{1}{\sigma^2}\big[RSS_j - RSS + (\beta_j - \tilde{\beta}_j)^T X_j^T X_j(\beta_j - \tilde{\beta}_j) \\ &\quad - (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})\big] \end{aligned}$$

where RSS_j is the residual sum of squares from model M_j, X_j is the partition of X corresponding to β_j, and $\tilde{\beta}_j$ is the MLE of β_j in model M_j.

The posterior distribution of *D* or *LR* is easily simulated, by generating a random σ from its marginal posterior distribution, and a random β from its conditional posterior given the generated σ. The MLEs and residual sums of squares from the models are known, and the quadratic forms in β are evaluated from the MLEs and X matrices.

We show in Figs. 4–11 the posterior distributions of *D*, based on 1000 simulations, for the successive omitted partitions corresponding to the backward elimination *t*-statistics: $\{c\}$, $\{c, \text{drat}\}$, $\{c, \text{drat}, s\}$, $\{c, \text{drat}, s, t\}$, $\{c, \text{drat}, s, t, \log(\text{disp})\}$, $\{c, \text{drat}, s, t, \log(\text{disp}), cb\}$, $\{c, \text{drat}, s, t, \log(\text{disp}), cb, g\}$, $\{c, \text{drat}, s, t, \log(\text{disp}), cb, g, \log(\text{hp})\}$.

The distributions are all remarkably diffuse, with very large variances, reflecting the very small degrees of freedom of the residual sum of squares. The tail probabilities that $D < 4.605$, that is that $LR > 0.1$ (weak evidence against the null hypothesis of zero regression coefficients), are given in Table 1, together with the *P*-values from the relevant *t*-distributions.

Despite this diffuseness the message is very clear: The early distributions have large probabilities for $D < 4.605$, around 0.5 for the first four variables eliminated; this drops to around 0.16 at step 5 but increases again to around 0.2 in step 7. At step 8 the distribution changes drastically, with a tail probability below 4.605 of only 0.036. These results are completely consistent with the backward elimination t-statistics, though the latter are not *pooled* tests of all the variables being eliminated.

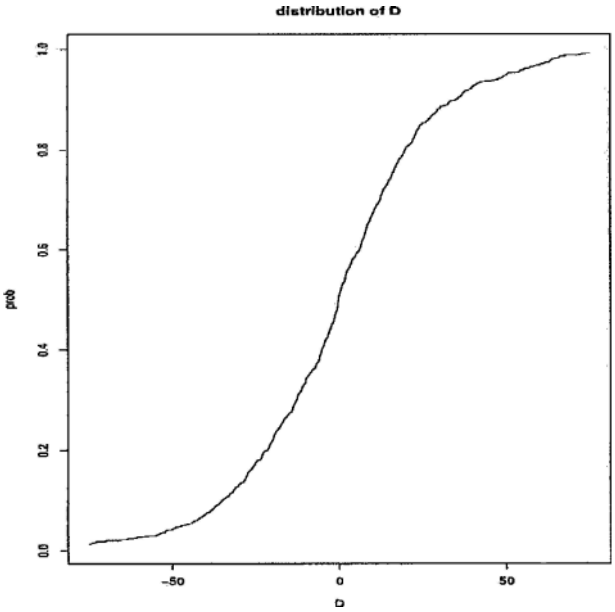

Fig. 4. *Step 1*

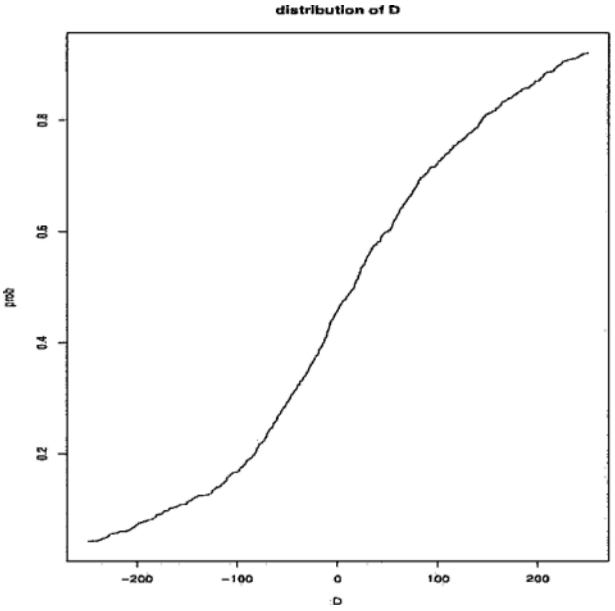

Fig. 5. *Step 2*

7. The importance of parametrization

In his discussion of Aitkin (1997), Dempster expressed concern about the extension of the approach proposed there, in the sense of the dependence of the true *LR* on the parametrization of the nuisance parameter.

The ideal parametrization would have fully orthogonal parameters θ and ϕ, with likelihood of the form

$$L(\theta, \phi) = L_1(\theta)L_2(\phi).$$

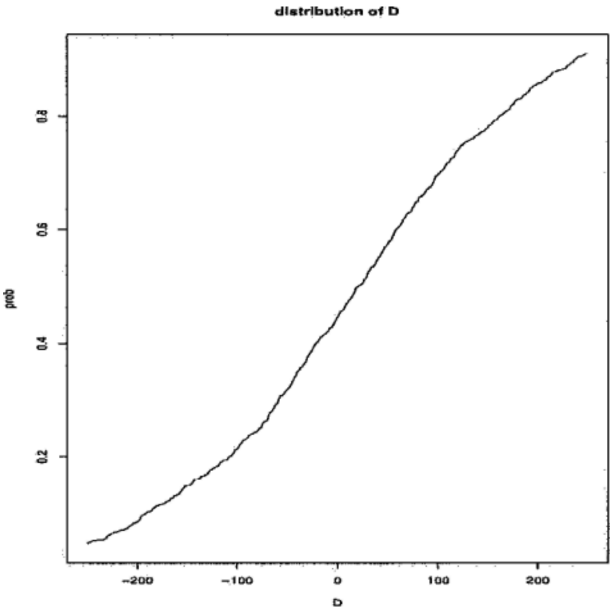

Fig. 6. *Step 3*

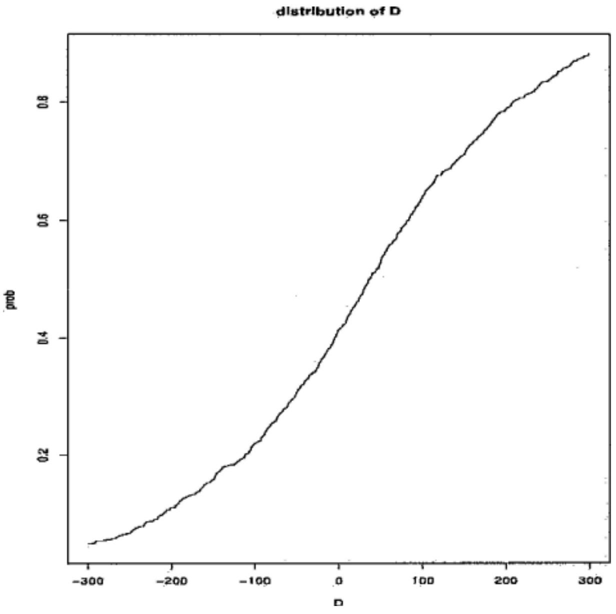

Fig. 7. *Step 4*

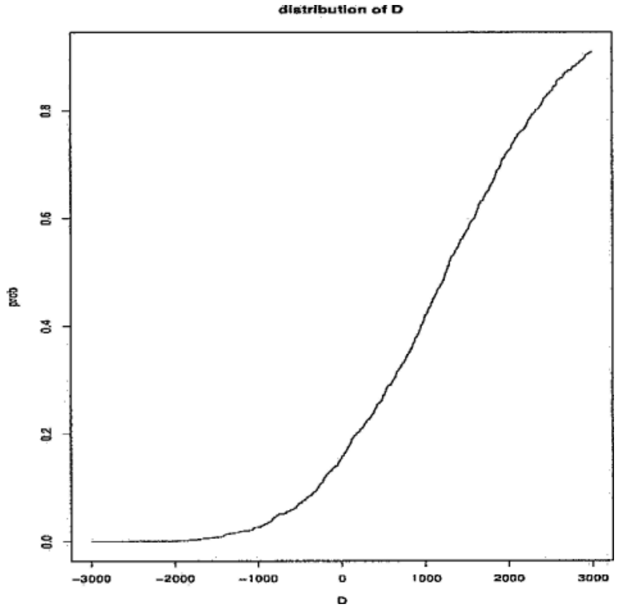

Fig. 8. *Step 5*

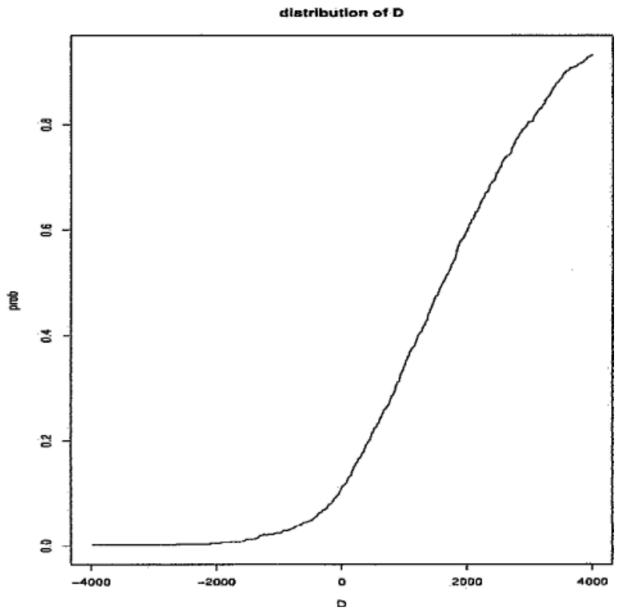

Fig. 9. *Step 6*

Then the true likelihood ratio for a null hypothesis $H_1 : \theta = \theta_1$ is

$$LR = \frac{L(\theta_1, \phi)}{L(\theta, \phi)} = \frac{L_1(\theta_1)}{L_1(\theta)}$$

for any ϕ, so the nuisance parameter is irrelevant—any prior distribution for it gives the same likelihood ratio for θ.

This parametrization does not generally exist; the next best has orthogonality in the observed or expected information matrix (Cox and Reid 1987); as the sample size increases and if the likelihood approaches normality in the parameters this will give approximately orthogonal parameters. For such a parametrization, independent priors will be a natural choice and their effect will dissipate rapidly with increasing sample size.

The importance of orthogonality is clear from the following example.

The binomial sample size

Given a sample $y_1, \ldots, y_n$ from a binomial distribution $b(N, p)$ with both parameters unknown, what can be said about N? The

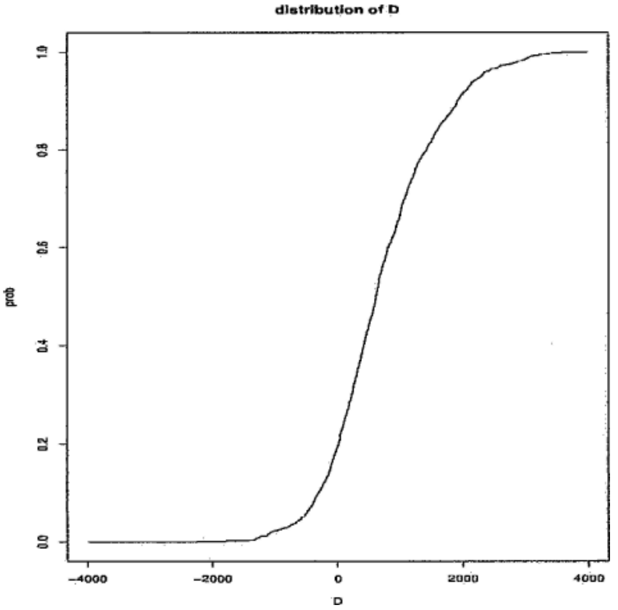

Fig. 10. *Step 7*

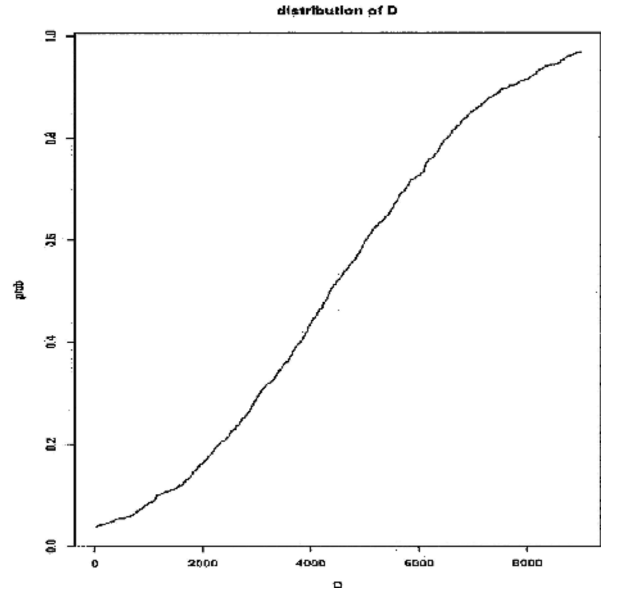

Fig. 11. *Step 8*

likelihood is

$$L(N,p) = \prod_{i=1}^{n} \binom{N}{y_i} p^{y_i}(1-p)^{N-y_i}$$
$$= \left[\prod_{i=1}^{n} \binom{N}{y_i}\right] p^{t}(1-p)^{Nn-t}$$

where $t = \sum_1^n y_i$.

This problem was originally considered by Olkin, Petkau and Zidek (1981) in the framework of the "instability" of the MLE $\hat{N}$ from samples in the "near-Poisson" region where the sample mean and variance were close.

Table 1. Pr[D < 4.605] *and t P-value for subset elimination*

Step	Variable omitted	Pr[D < 4.605]	P-value
1	c	0.585	0.935
2	drat	0.473	0.752
3	s	0.458	0.737
4	t	0.420	0.649
5	log(disp)	0.159	0.476
6	cb	0.111	0.294
7	g	0.196	0.302
8	log(hp)	0.036	0.00014

A recent discussion from a Bayesian perspective, with some references, was given by Berger, Liseo and Wolpert (1999) who argued for the general use of integrated likelihoods for the elimination of nuisance parameters, and gave this model and the following data (considered by Olkin *et al.* and later authors) as a persuasive example.

The data from a sample of $n = 5$ are 16, 18, 22, 25, 27. The sample mean is $\bar{y} = 21.6$ and the (biased) variance estimate s^2 is 17.04, giving moment estimates of $\tilde{p} = 1 - s^2/\bar{y} = 0.211$ and $\tilde{N} = \bar{y}/\tilde{p} = 102.3$. These estimates are highly unstable, as are the MLEs, in the sense that small changes in the largest observation produce very large changes in $\tilde{N}$: for example, if the largest observation is changed to 28, then $\bar{y} = 21.8, s^2 = 19.36, \tilde{p} = 0.112, \tilde{N} = 194.8$.

The profile likelihood in N is nearly flat, with a very poorly defined maximum, and the conditional likelihood conditioned on t has no internal maximum at all, approaching its maximum as $N \to \infty$. Berger *et al.* concluded that "These [likelihoods] are nearly constant over a huge range of N and are clearly useless for inference." They proposed the uniform or Jeffreys priors for this problem; these give well-defined modes in the integrated likelihood for N.

Kahn (1987) had earlier considered general conjugate beta priors

$$\pi(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)},$$

and had shown that the integrated likelihood in N,

$$\left[\prod_{i=1}^{n} \binom{N}{y_i}\right] B(t+a, Nn-t+b),$$

in this example is extremely sensitive to the value of the first beta parameter a, which controls the location of the mode and the heaviness of the tail of the posterior distribution of N; for $a = 0$ this tail is flat, giving an essentially uninformative posterior for N, equivalent to the conditional likelihood.

A detailed comparison of profile likelihood and integrated likelihood inference for this example was given in Aitkin and Stasinopoulos (1989), who also showed the likelihood in N and p, which has extremely concentrated banana-shaped contours along the curve $N_p = \bar{y}$ (Fig. 12).

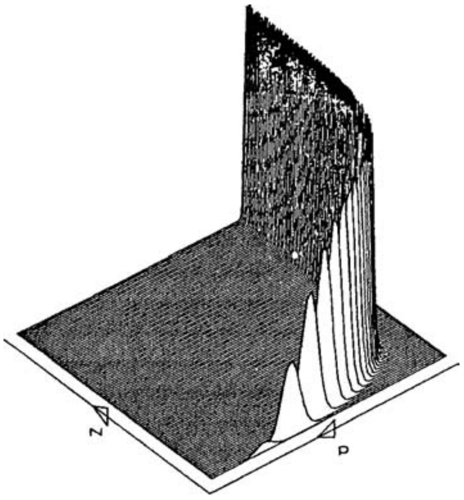

Fig. 12. *Likelihood in (N, p)*

This very strong association between the parameters emphasises the difficulty of drawing marginal inferences about *N*, at least in this parametrization.

Aitkin and Stasinopoulos derived the (expected) information-orthogonal nuisance parameter transformation $\psi = Np$, by solving a partial differential equation, following Cox and Reid (1987). The joint likelihood in *N* and ψ, shown in Fig. 13, is almost orthogonal, and whether this joint likelihood is maximized or integrated over ψ, the resulting likelihood is essentially the profile likelihood (which is invariant over nuisance parameter transformation).

The true likelihood ratio shows clearly the difficulty in the N, p parametrization. Consider two candidate values of the binomial index, N_1 and N_2. The true likelihood ratio is, as an explicit function of *p*,

$$\begin{aligned} LR(p) &= \frac{L(N_1, p)}{L(N_2, p)} \\ &= (1-p)^{(N_1-N_2)n} \prod_i \left[\binom{N_1}{y_i} \Big/ \binom{N_2}{y_i} \right]. \end{aligned}$$

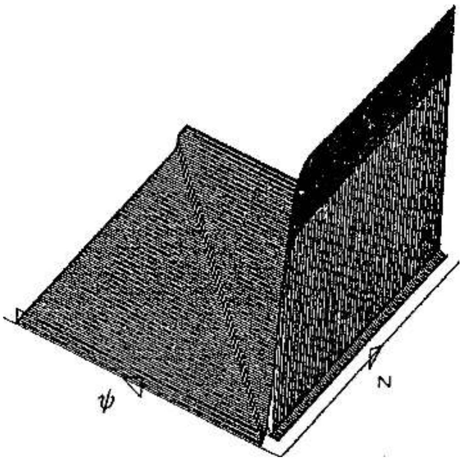

Fig. 13. *Likelihood in (N, ψ)*

The ratio of products of binomial coefficients can be expressed simply using Stirling's formula (since all of *N* and the y_i are large) as

$$\prod_i \left[\binom{N_1}{y_i} \Big/ \binom{N_2}{y_i} \right] \approx \left(\frac{N_1}{N_2} \right)^t.$$

For even moderate differences between N_1 and N_2 and even small values of *n*, the term $(1-p)^{(N_1-N_2)n}$ depends very strongly on the prior distribution for *p*.

In the N, ψ parametrization, the likelihood is

$$L(N, \psi) = \prod_{i=1}^{n} \binom{N}{y_i} \left(\frac{\psi}{N} \right)^t \left(1 - \frac{\psi}{N} \right)^{Nn-t}$$

and the true LR becomes

$$\begin{aligned} LR(\psi) &= \frac{L(N_1, \psi)}{L(N_2, \psi)} \\ &\to \left(\frac{N_2}{N_1} \right)^t \prod_i \left[\binom{N_1}{y_i} \Big/ \binom{N_2}{y_i} \right] \\ &\to 1 \end{aligned}$$

since for large *N* the last term in the likelihood tends to $\exp(-n\psi)$. Thus in the ψ parametrization the likelihood ratio $LR(\psi)$, based on the section through the $L(N, \psi)$ likelihood at ψ, does not depend on ψ, nor on the data, and approaches 1. That is, the tail of the likelihood is flat in *N* for any given ψ.

This is in accord with the "near-Poisson" nature of the sample—the maximized likelihood ratio for Poisson to "best binomial" is 0.935 (Aitkin and Stasinopoulos)—and with the profile likelihood which exhibits this asymptotic behaviour. Since ψ is bounded above by *N*, the parameter spaces for *N* and ψ are not independent. However the likelihood in ψ goes to zero rapidly with ψ when far from *N*, and so the upper bound on the range for ψ has no practical consequences.

We note finally that independent flat priors on *p* and *N* transform to a prior in *N* and ψ of the form $\pi(N, \psi) = 1/N$, and it is this term in $1/N$ which "pulls down" the flat *N* tail of the likelihood in the N, ψ parametrization, giving the well-defined mode in *N* in the *N, p* parametrization with the uniform prior in *p*.

Thus the "useless" profile or conditional likelihoods are in fact conveying correctly the information *in the data* about *N*—the well-defined modes in the integrated likelihoods for the uniform and Jeffreys priors are direct consequences of these priors, and give a misleading impression of the information *in the data* about *N*. As we noted earlier, the use of independent flat priors may have a strong effect on the marginal posteriors if the parameters are strongly associated in the likelihood.

8. The 2 × 2 table for randomized clinical trials

In the 2 × 2 randomized clinical trial, subjects are randomized to one of two treatment conditions, giving n_1 patients in treatment

Table 2. *ECMO trial outcome*

	ECMO	CMT	Total
Recover	11	0	11
Died	0	1	1
Total	11	1	12

1 and n_2 in treatment 2. The response to treatment is the binary event of "success" or "failure", suitably defined. The response probability in treatment j is p_j, and of the n_j patients treated, r_j are successes, with $r = r_1 + r_2$, $n = n_1 + n_2$. What can be said, in a Bayesian framework, about:

- the attributable risk $\Delta = p_1 - p_2$;
- the relative risk $\rho = p_1/p_2$;
- the odds ratio $\psi = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$ and
- the number needed to treat $nnt = \frac{1}{p_2} - \frac{1}{p_1}$?

The likelihood is

$$L(p_1, p_2) = p_1^{r_1}(1-p_1)^{n_1-r_1} \cdot p_2^{r_2}(1-p_2)^{n_2-r_2},$$

and using independent conjugate Beta priors

$$\pi(p_j) = p_j^{a_j-1}(1-p_j)^{b_j-1}/B(a_j, b_j),$$

the posterior distribution of p_1, p_2 is the product of independent Beta posteriors

$$\pi(p_j \mid y) = p_j^{r_j+a_j-1}(1-p_j)^{n_j-r_j+b_j-1}/B(r_j + a_j, n_j - r_j + b_j).$$

Exact results for the posterior distribution of the attributable risk or any of the other measures of difference are complex and involve sums of hypergeometric probabilities (Altham 1969, Hashemi *et al.* 1997). However the marginal posterior distribution of any parametric function of p_1 and p_2 can be simulated directly, by generating N realizations from the posterior distributions of p_1 and p_2, and calculating the appropriate function (Tanner 1996). This is an extremely simple calculation. We illustrate with the following table, from the ECMO study of Bartlett *et al.* (1985). This study compared the ECMO (extra corporeal membrane oxygenation—oxygenation of the blood outside the body) treatment for respiratory failure in newborn babies with CMT (conventional medical treatment—oxygen under pressure in a respirator). The "play the winner" randomization method used is discussed below; it led to the treatment of 11 babies with ECMO, of whom all recovered, and 1 baby with CMT, who died. We use *uniform* priors here initially; we show the effect of non-uniform priors below, and comment on the general use of uniform priors.

The posterior distribution of p_1 (for ECMO) is

$$\pi(p_1 \mid y) = 12p_1^{11},$$

and that of p_2 for CMT is

$$\pi(p_2 \mid y) = 2(1 - p_2).$$

What is the posterior probability that $p_1 > p_2$? We have immediately that

$$\begin{aligned}\Pr[p_1 > p_2 \mid y] &= \int_0^1 2(1-p_2)\mathrm{d}p_2 \int_{p_2}^1 12p_1^{11}\mathrm{d}p_1\\ &= 2\int_0^1 (1-p_2)\cdot\left(1-p_2^{12}\right)\mathrm{d}p_2\\ &= 0.989.\end{aligned}$$

Thus there is *strong* evidence that ECMO is better. Altham (1969) gave this probability calculation for the general 2 × 2 table in terms of hypergeometric probabilities; it is expressed there in terms of the odds ratio being greater than 1. Altham showed that the Fisher P-value exceeds the posterior probability for all priors with common indices $a_j = b_j = c$ for $0 \le c \le 1$, but this result need not hold for $c > 1$.

The superiority of ECMO holds for all the measures of discrepancy above. However to determine the extent of its superiority, we need the full posterior distribution of the discrepancy measures.

We generate $N = 10,000$ independent realizations p_{1j}, p_{2j} of p_1 and p_2 from their posterior distributions with flat prior distributions. For each pair j we compute the four discrepancy measures above. The empirical *cdfs* of the four measures are shown in Figs. 14–17, on log scales for the relative risk and odds ratio.

The empirical probability that the attributable risk is positive is 0.9893, with simulation standard error 0.0010, in close agreement with the theoretical value. The same probability applies to the relative risk and odds ratio being greater than 1, or the *nnt* being positive. Equal-tailed 95% credible intervals for the four measures are:

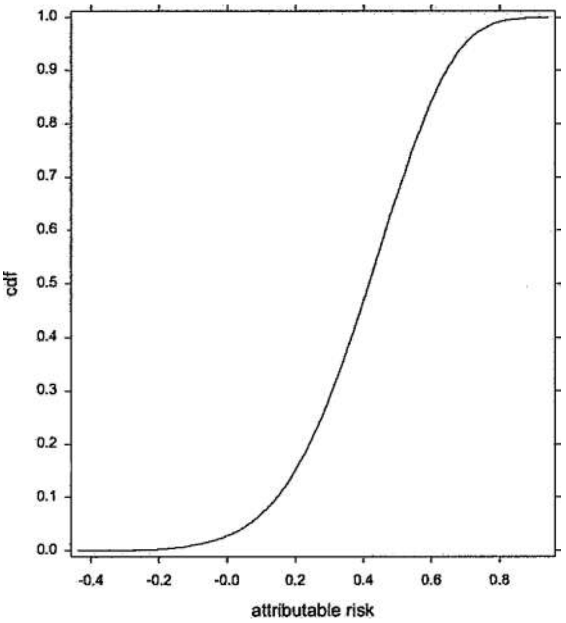

Fig. 14. *Posterior distribution of attributable risk*

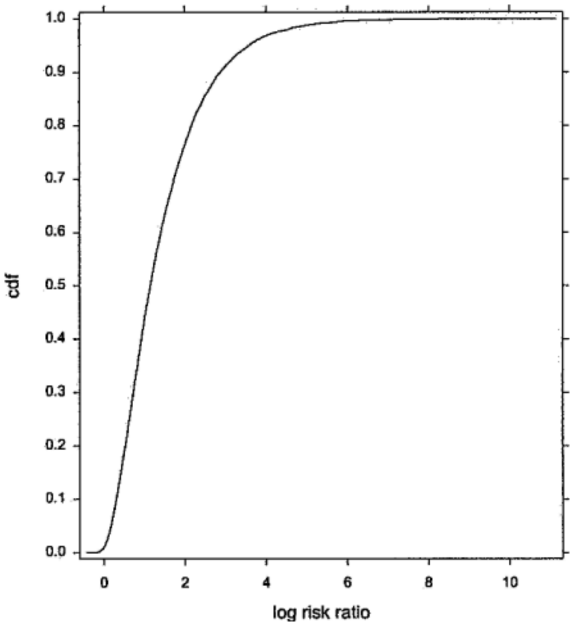

Fig. 15. *Posterior distribution of log relative risk*

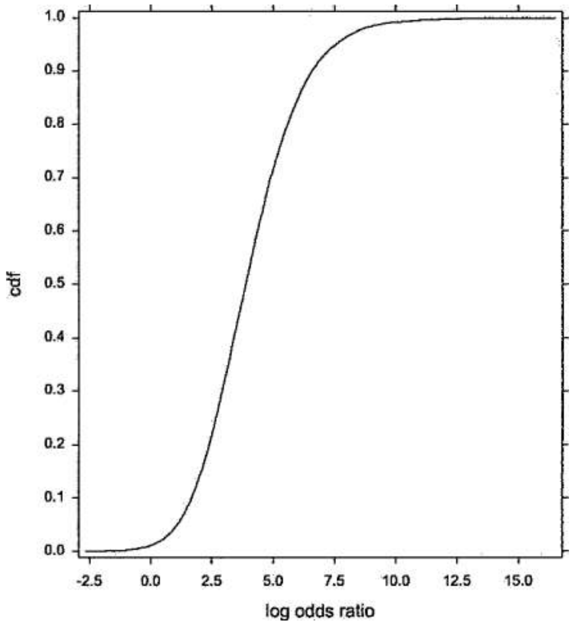

Fig. 16. *Posterior distribution of log odds ratio*

- attributable risk—(0.069, 0.948);
- relative risk—(1.09, 69.9);
- odds ratio—(1.78, 4501);
- *nnt*—(0.095,72.9).

For the number needed to treat, the distribution is extremely long-tailed, because of the posterior density of p_2 having its mode at zero. This is an inherent difficulty of this discrepancy measure; if *both* probabilities can be small, and especially if they can be equal, the *nnt* distribution will be extremely long-tailed in both directions and will have appreciable mass at $\pm\infty$, which

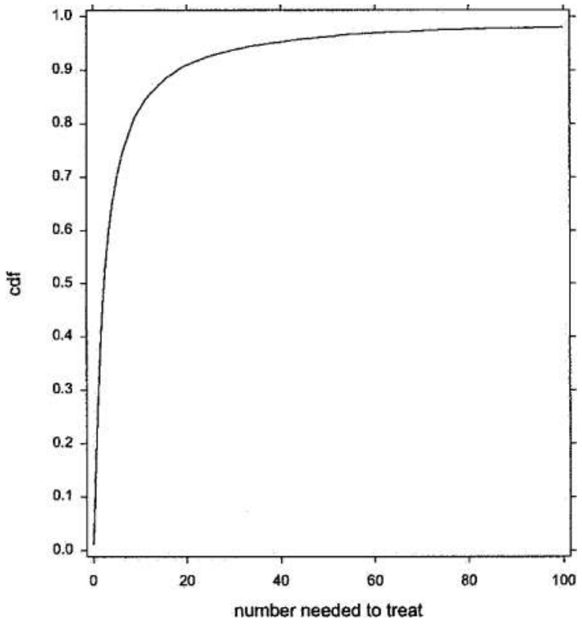

Fig. 17. *Posterior distribution of number needed to treat*

will be unhelpful for interpretation. These and other deficiencies of the *nnt* have recently been discussed by Hutton (2000).

The distributions of all the discrepancy measures are very diffuse, not surprising from the sample of one CMT baby, though they are all well away from the "null" value, as we saw above.

8.1. *Fisher's "exact" test*

The standard test for the 2 × 2 table, especially with small samples, is Fisher's "exact" test, based on the conditional hypergeometric distribution of R_1 given the marginal total $R = r$. This is

$$\begin{aligned}\Pr[R = r_1 \mid R = r] &= \Pr[R = r_1, R_2 = r_2]/\Pr[R = r]\\ &= \binom{n_1}{r_1}\binom{n_2}{r_2}\psi^{r_1}\Big/\sum_{u=u_1}^{u=u_2}\binom{n_1}{u}\binom{n_2}{r-u}\psi^{u},\end{aligned}$$

where ψ is the odds ratio, and $u_1 = \max(0, r - n_2), u_2 = \min(n_1, r)$. For the ECMO example, the conditional likelihood from the hypergeometric distribution is

$$CL(\psi) = \frac{\psi^{11}}{11\psi^{10} + \psi^{11}} = \frac{\psi}{11 + \psi}.$$

At the null hypothesis value $\psi = 1, CL(1) = 1/12 = 0.0833$. Since this table is the most extreme possible, the P-value of this observed table is 0.0833, which does not reach conventional levels of significance.

This lack of sensitivity of the "exact" test follows from the loss of information in the conditioning statistic. Although Fisher argued that the marginal total was ancillary, or at least should be treated as such, Plackett (1977) showed that the marginal total R *is* informative about ψ, though it is difficult to make use of this information in a classical framework, and as the sample sizes tend to infinity, this information becomes negligible relative to the information in the cells. However we are at the opposite extreme, where the sample sizes are very small, and here the information in the marginal total may be appreciable. This is clear from comparing the maximized conditional likelihood ratio for the null hypothesis against the alternative, of 0.0833, with the unconditional maximized likelihood ratio of $(11/12)^{11} \cdot (1/12)/1 = 0.032$ which would provide strong evidence against the null hypothesis, with a P-value of 0.0087 under the asymptotic χ_1^2 distribution, if this were valid.

The posterior distribution of the likelihood ratio requires a choice of parameterization for the nuisance intercept parameter in the regression model for the 2 × 2 table. For the normal regression model for a two-group structure with group sample sizes n_1 and n_2, the dummy variable coding giving information-orthogonal parameters is $(-n_2/n, n_1/n)$. We adopt this parametrization for the identity link probability model for the attributable risk, though the resulting information matrix is not quite orthogonal because of the iterative weights in the generalized linear model analysis. The parameters transform to

$$p_1 = \beta_0 - \frac{n_2}{n}\beta_1, \quad p_2 = \beta_0 + \frac{n_1}{n}\beta_1,$$

with

$$\beta_0 = (n_1 p_1 + n_2 p_2)/n = \bar{p}.$$

Note that this form of the nuisance parameter is *exactly* information-orthogonal to the log-odds ratio parameter, see Cox in the discussion of Yates (1984). However there is no analytic relation between these two parametrizations and so we use the simpler linear model parametrization. The likelihood in the regression parameters is

$$L(\beta_0, \beta_1) = \left(\beta_0 - \frac{n_2}{n}\beta_1\right)^{r_1} \left(1 - \beta_0 + \frac{n_2}{n}\beta_1\right)^{n_1 - r_1} \times \left(\beta_0 + \frac{n_1}{n}\beta_1\right)^{r_2} \left(1 - \beta_0 - \frac{n_1}{n}\beta_1\right)^{n_2 - r_2},$$

and under the null hypothesis,

$$L(\beta_0, 0) = \beta_0^r (1 - \beta_0)^{n-r}.$$

The likelihood ratio is, in the p_1, p_2 parametrization,

$$LR = \frac{\bar{p}^r (1 - \bar{p})^{n-r}}{p_1^{r_1}(1 - p_1)^{n_1 - r_1} p_2^{r_2}(1 - p_2)^{n_2 - r_2}}.$$

Figures 18 and 19 show the empirical *cdf* of the likelihood ratio and the corresponding deviance from the 10000 simulations above.

The empirical probability that $LR < 1$ is 0.9893, the same value as the posterior probability that the attributable risk is positive; the simulated posterior probability that $LR > 1$ of 0.0107, with simulation standard error 0.0010, is substantially below the Fisher P-value, but greater than the P-value from the unconditional LR test using the asymptotic χ_1^2 distribution.

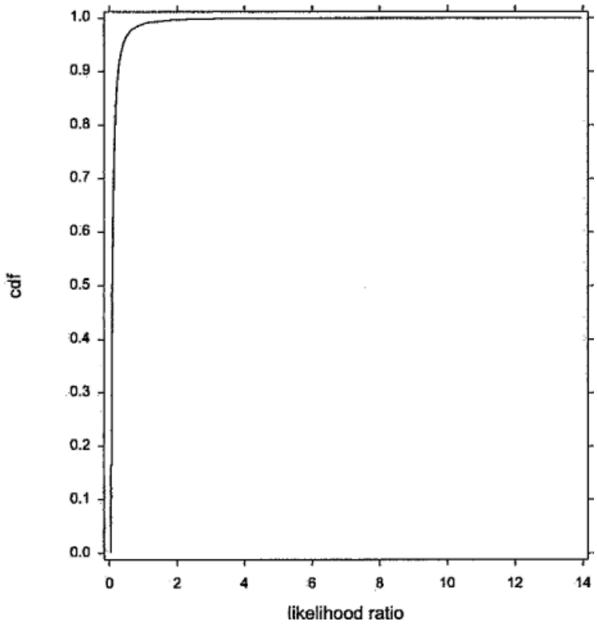

Fig. 18. *Posterior distribution of likelihood ratio*

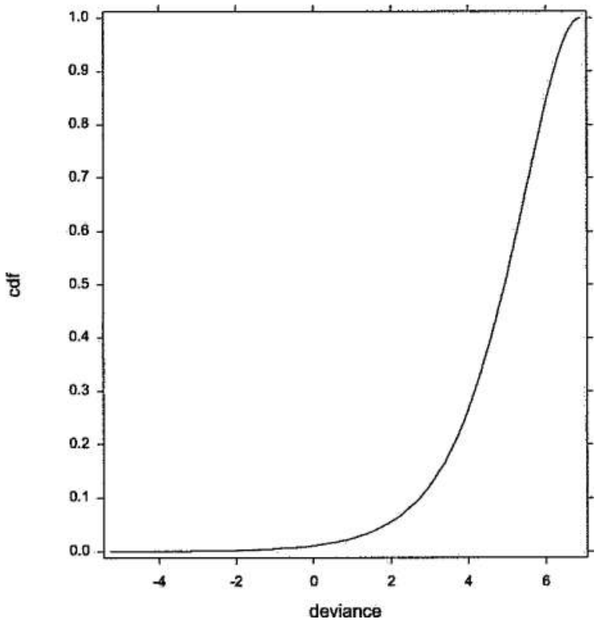

Fig. 19. *Posterior distribution of deviance*

The difficulty of calibrating the unconditional test, and the dependence of its size on the true response probabilities, is resolved by the Bayes analysis.

This analysis also resolves the "reference set" difficulties of the ECMO study, in which the play-the winner randomization rule used different assignment probabilities of babies to the ECMO and CMT conditions. The stopping rule for the study was not well-defined, and this makes it very difficult to determine the reference set of tables against which this one should be compared, leading to the six (at least) P-values which have been proposed for this table, ranging from 0.001 to 0.62; see Ware

Table 3. *Second ECMO trial outcome*

	ECMO	CMT	Total
Recover	9	6	15
Died	0	4	4
Total	9	10	19

(1989) and Begg (1990) and their discussions for the range of P-values, and the arguments for them.

A second randomized trial of ECMO, described in Ware, was carried out because of the inconclusive P-value results from the first trial due to the single death under CMT. The second trial used a different stopping rule and resulted in the outcome shown in Table 3. Using the same flat priors as for the first study, we have

$$\pi(p_1 \mid y) = 10p_1^9,$$
$$\pi(p_2 \mid y) = p_2^6(1-p_2)^4/B(7,5)$$

and hence

$$\begin{aligned}\Pr[p_1 > p_2 \mid y] &= \int_0^1 p_2^6(1-p_2)^4 \mathrm{d}p_2 \int_{p_2}^1 10p_1^9 \mathrm{d}p_1 \\ &= 2\int_0^1 p_2^6(1-p_2)^4\left(1-p_2^{10}\right)\mathrm{d}p_2/B(7,5) \\ &= [B(7,5)-B(17,5)]/B(7,5) \\ &= 1-\frac{22}{969} = 0.977.\end{aligned}$$

The larger study provided less persuasive evidence against the null hypothesis, because of the better-defined and much higher recovery rate under CMT.

8.2. *Choice of priors*

It may be argued that the Bayes analysis above has arbitrary assumptions of its own, in the choice of priors. If a reference prior is to be used, why not use the Jeffreys prior—why is the uniform prior appropriate? Should we not in any case use informative priors, based on previous experience with both treatments, especially when the sample sizes are so small? Since changes in priors affect the conclusions, should we not report a sensitivity analysis over a range of priors?

We argue that, in studies of this kind involving randomized trials to establish the value of a new treatment, informative priors, and the Jeffreys prior, should not be used without a *reference* analysis with uniform priors. The uniform prior has a unique position in binomial experiments, since for the (large but finite) conceptual population of N individuals to whom the treatments are to be applied, the population number of successes R is necessarily an integer, and so the population proportion of successes takes values on an equally-spaced grid of values R/N. In the absence of experimental information, the possible values of this proportion are equally well supported on this grid, and so p should be given a uniform prior distribution.

Incorporating the information from previous non-randomized studies in an informative prior affects the inferences from the randomized trial—in such a trial it seems to us critical to "let the data speak" through uniform priors before changing its information content by introducing informative priors.

We illustrate this point by a second analysis of the first ECMO table with the Jeffreys prior

$$\pi_j(p_j) = p_j^{-0.5}(1-p_j)^{-0.5}/B(0.5, 0.5).$$

The empirical probability of a positive attributable risk now changes to 0.9954, and the equal-tailed 95% credible intervals become

- attributable risk—(0.097, 0.993);
- relative risk—(1.11, 2452);
- odds ratio—(3.27, 1.14×10^6);
- *nnt*—(0.120, 2540).

The apparent strength of evidence against the null hypothesis has increased, while the credible intervals have become even more diffuse. Both response probability posteriors have infinite spikes at their former finite modes of 1 and 0, the priors accentuating the information in the likelihood, which makes the Jeffreys prior choice hard to justify.

9. Conclusion

The possible inconsistency between the conclusions from posterior distributions of "null hypothesis" parameters and those from Bayes factors for testing the hypotheses can be avoided by retaining the full posterior distribution of the alternative model parameters and transforming from this distribution to that of the likelihood ratio between the models. The resulting inferences are consistent between "hypothesis testing" and "estimation", as they are in frequentist theory, and are closely related to frequentist P-value conclusions, though these need to be recalibrated.

Parametrization issues have to be considered carefully in this approach, as they do in other Bayesian analyses and in frequentist analyses of models with nuisance parameters. A particular strength of this analysis is the freedom to use flat, non-informative or other reference priors in the comparisons of models in the same way they are used in posterior densities for individual model parameters.

Acknowledgments

We appreciate the careful reading of the manuscript and suggestions by the Editor, Associate Editor and referees which have greatly improved the paper. We thank Dr Gary Phillips for the initial stimulus for some of this work, and we are grateful to the UK Engineering and Physical Sciences Research Council for support for Tom Chadwick, and to the American Institutes for Research and the National Center for Education Statistics, US Department of Education for support for Tom Chadwick and Murray Aitkin. This project has been funded at least in part with Federal funds from the U.S. Department of Education, National Center for Education Statistics under contract number RN95127001. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education, National Center for Education Statistics, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government.

References

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

Aitkin M., Anderson D.A., Francis B.J. and Hinde J.P. 1989. Statistical Modelling in GLIM. Oxford University Press.

Aitkin, M. and Stasinopoulos, M. 1989. Likelihood analysis of a binomial sample size problem. In: Gleser L.J., Perlman M.D., Press S.J. and Sampson A.R. (Eds.), Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin. Springer-Verlag, New York (1989).

Altham P.M.E. 1969. Exact Bayesian analysis of a 2 × 2 contingency table, and Fisher's "exact" test. J. Roy. Statist. Soc. B 31: 261–269.

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

Begg C.B. 1990. On inferences from Wei's biased coin design for clinical trials (with discussion). Biometrika 77: 467–484.

Berger J. and Bernardo J.M. 1989. Estimating a product of means: Bayesian analysis with reference priors. J. Amer. Statist. Assoc. 84: 200–207.

Berger J., Liseo B. and Wolpert R.L. 1999. Integrated likelihood methods for eliminating nuisance parameters. Statist. Science 14: 1–28.

Chadwick T.J. 2002. A general Bayes theory of nested model comparisons. Unpublished PhD thesis, University of Newcastle upon Tyne.

Cox D.R. and Reid N. 1987. Parameter orthogonality and approximate conditional inference (with Discussion). J. Roy. Statist. Soc. B 49: 1–39.

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

Dempster A.P. 1997. The direct use of likelihood in significance testing. Statist. and Computing 7: 247–252.

Geisser S. 1993. Predictive Inference: An Introduction. CRC Press, Boca Raton.

Hashemi L., Balgobin N., and Goldberg R. 1997. Bayesian analysis for a single 2 × 2 table. Statist. in Med. 16: 1311–1328.

Henderson H.V. and Velleman P.F. 1981. Building multiple regression models interactively. Biometrics 29: 391–411.

Hutton J.L. 2000. Number needed to treat: Properties and problems (with comments). J. Roy. Statist. Soc. A 163: 403–419.

Kahn W.D. 1987. A cautionary note for Bayesian estimation of the binomial parameter n. Amer. Statist. 41: 38–39.

Kass R.E. and Greenhouse, J.B. 1989. Comment: A Bayesian perspective. Statist. Science 4: 310–317.

Kass R.E. and Raftery, A.E. 1995. Bayes factors. J. Amer. Statist. Assoc. 90: 773–795.

Olkin I., Petkau, A.J. and Zidek, J.V. 1981. A comparison of n estimators for the binomial distribution. J. Amer. Statist. Assoc. 76: 637–642.

Plackett R.L. 1977. The marginal totals of a 2 × 2 table. Biometrika 64: 37–42.

Stone M. 1997. Discussion of Aitkin (1997). Statist. and Computing 7: 263–264.

Tanner M.A. 1996. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edn., Springer-Verlag, New York.

Ware J.H. 1989. Investigating therapies of potentially great benefit: ECMO (with discussion). Statist. Science 4: 298–340.

Yates F. 1984. Tests of significance for 2 × 2 tables (with Discussion). J. Roy. Statist. Soc. A 147: 426–463.