Statistics and Computing (1997) **7**, 253–261

The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood

MURRAY AITKIN

Department of Statistics, University of Newcastle, Newcastle-upon-Tyre, NE17RU

The posterior distribution of the likelihood is used to interpret the evidential meaning of *P*-values, posterior Bayes factors and Akaike's information criterion when comparing point null hypotheses with composite alternatives. Asymptotic arguments lead to simple re-calibrations of these criteria in terms of posterior tail probabilities of the likelihood ratio. ('Prior') Bayes factors cannot be calibrated in this way as they are model-specific.

Keywords: *P*-value, likelihood, posterior distribution, Bayes factor, fractional Bayes factor, posterior Bayes factor, AIC

1. Introduction

In a conference paper (reprinted on pp. 247–252), Dempster (1974) considered the problem of testing a simple point null hypothesis against a composite alternative. Such testing problems are universal in classical statistical theory but cause severe difficulties in Bayes theory, expressed in Lindley's paradox (Lindley, 1957; Bartlett, 1957; Berger, 1985; Aitkin, 1991). Suppose the model $p(y|\theta)$ for data y gives a likelihood function $L(\theta)$, with prior distribution $\pi(\theta)$ of θ. Suppose the null hypothesis is $H_0 : \theta = \theta_0$, and the alternative is $\bar{H}_0 : \theta \neq \theta_0$.

Dempster considered the *posterior distribution of the likelihood ratio* $L(\theta_0)/L(\theta)$. Suppose we would regard the sample evidence against H_0 as convincing if the likelihood ratio (LR) $L(\theta_0)/L(\theta_1) < k$, where θ_1 is the alternative value of θ under a specific simple hypothesis H_1, and k is a constant chosen to define 'convincing'. For example, k might in different circumstances be chosen as 0.3, or 0.1, or 0.05, or some other small value. (If H_0 and H_1 have equal prior probabilities, the posterior probability of H_0 would then be 0.231, 0.091, or 0.048. The value $k = 1$ would correspond to H_1 being 'better supported' than H_0, but it would not be regarded as convincing evidence against H_0. For unequal prior probabilities, the value of k would incorporate the prior odds.)

This inequality is equivalent to $\ell(\theta_0) - \ell(\theta_1) < \log k$, where $\ell(\theta)$ is the log-likelihood function. Under the general alternative $\bar{H}_0$, and given the data, $\ell(\theta)$ is a parametric function of θ and the observed data, and so has a posterior distribution obtainable from that of θ. Dempster proposed that the degree of certainty of the strength of evidence against H_0 be measured by the pair (k, π_k), where π_k is the posterior probability that the LR is less than k:

$$\begin{aligned}\pi_k &= \Pr(\ell(\theta_0) - \ell(\theta) < \log k|y) \\ &= \Pr(\ell(\theta) > \ell(\theta_0) - \log k|y).\end{aligned}$$

Clearly, decreasing k will decrease π_k for the same data, so if we require stronger evidence against H_0 we will have to accept a smaller posterior probability of this strength of evidence.

A simple example will illustrate this approach. We are given $n = 25$ observations on $Y \sim \mathrm{N}(\mu, \sigma^2)$ with known $\sigma = 1$, resulting in $\bar{y} = 0.4$. The null hypothesis is $H_0 : \mu = \mu_0 = 0$. The prior distribution for μ is diffuse. The likelihood ratio is

$$\frac{L(\mu_0)}{L(\mu)} = \exp\left\{-\frac{n}{2\sigma^2}(\bar{y} - \mu_0)^2\right\} \Big/ \exp\left\{-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2\right\}$$

and the posterior distribution of μ is $\mathrm{N}(\bar{y}, \sigma^2/n)$. The inequality $L(\mu_0)/L(\mu) < k$ is equivalent to

$$\sqrt{n}|\bar{y} - \mu|/\sigma < \left\{n(\bar{y} - \mu_0)^2/\sigma^2 + 2\log k\right\}^{1/2}$$

which has posterior probability

$$\pi_k = 2\Phi\left[\left\{n(\bar{y}-\mu_0)^2/\sigma^2 + 2\log k\right\}^{1/2}\right] - 1.$$

Write P for the P-value of the observed sample mean:

$$\begin{aligned} P &= \Pr(|Z| > \sqrt{n}(\bar{y}-\mu_0)/\sigma) \\ &= 2(1-\Phi(z_0)) \end{aligned}$$

where $z_0 = \sqrt{n}(\bar{y}-\mu_0)/\sigma$ is the observed standardized mean difference and Z has the standard normal distribution. Then

$$\begin{aligned} \pi_k &= 2\Phi\left(\left\{z_0^2 + 2\log k\right\}^{1/2}\right) - 1 \\ &= 2\Phi\left(\left\{\left[\Phi^{-1}(1-P/2)\right]^2 + 2\log k\right\}^{1/2}\right) - 1. \end{aligned}$$

In the example, $z_0 = 2.0$, $P = 0.0456$, and the posterior probability π_k is given for a range of values of k in Table 1.

For values of k less than $e^{-2} = 0.135$, π_k is zero, because the largest possible value of $L(\mu)$ is $L(\bar{y}) = e^{-2}$. The value $k = 1$ has a particular importance, because $\pi_1 = 1 - P$. Thus $1 - P$ *is equal to the posterior probability that the LR is less than* 1. This probability is large here (0.9544) and the correspondingly small P-value is taken in the Neyman–Pearson theory as marginally strong evidence against H_0. Viewed as a posterior probability, it gives a high degree of certainty, but the LR of less than 1 does not provide strong evidence against H_0. For convincing evidence we would need a high posterior probability that the LR is *small*, say less than 0.1, not 1. But the data cannot provide this strength of evidence: the posterior probability is only 0.350 that the LR is less than 0.15, and is *zero* that the LR is less than 0.1!

If we accept a LR of 0.1 as the evidence necessary to reject H_0, then we do not have that evidence here. However, if we were willing to reject H_0 on the basis of a *larger* LR – say 0.3, for example – the posterior probability is 0.8 that the LR is less than this value.

These results are generally in accord with Bayesian analyses of this model (Berger and Sellke, 1987) which conclude that the P-value overstates the strength of the evidence against H_0, but are more equivocal since we have two complementary measures (k and π_k) of evidential meaning rather than one.

A striking feature of the relation between P and π_k is that it is *independent of n*, depending only on k. We return to this point below.

2. Posterior Bayes factors

Aitkin (1991) proposed to correct the overstatement of strength of evidence by replacing the unknown likelihood under the alternative hypothesis by its *posterior mean* L^A, defined by

$$L^A = \int L(\theta)\pi(\theta|y)\mathrm{d}\theta.$$

Table 1. *Posterior probability π_k that the LR $< k$ when $P = 0.0456$*

k	0.1	0.135	0.15	0.2	0.3	0.4	0.5	1.0
π_k	0	0	0.350	0.673	0.793	0.859	0.894	0.954

He proposed that the *posterior Bayes factor* (PBF) – the likelihood ratio $L(\theta_0)/L^A$ – should be interpreted like a likelihood ratio from two simple hypotheses (parallelling the usual Bayes factor interpretation – see Section 4 below). The posterior mean is in many models a *penalized* form of the maximized likelihood $L(\hat{\theta})$.

By considering the posterior distribution of the LR $L(\theta_0)/L(\theta)$, we can calibrate the PBF in a similar way. Specifically, if the PBF has the value k, what is the posterior probability that the LR is less than k?

For the normal example above, Aitkin (1991) showed that $L^A = L(\hat{\mu})/\sqrt{2}$, and so

$$L(\mu_0)/L^A = \sqrt{2}L(\mu_0)/L(\hat{\mu}).$$

If the PBF equals k, then $L(\mu_0)/L(\bar{y}) = k/\sqrt{2}$, or equivalently

$$L(\mu_0)/L(\mu) = (k/\sqrt{2})L(\bar{y})/L(\mu).$$

The posterior probability that $L(\mu_0)/L(\mu) < k$, when the PBF equals k, is then

$$\begin{aligned} \pi_k &= \Pr((k/\sqrt{2})L(\bar{y})/L(\mu) < k|y) \\ &= \Pr(L(\mu)/L(\bar{y}) > 1/\sqrt{2}|y) \\ &= \Pr(Z^2 < \log 2) \\ &= 0.595, \end{aligned}$$

independent of k. This posterior probability is around 0.5, as might be expected: although the maximized likelihood is mean-corrected, the posterior variance of the likelihood around its mean means that the posterior probability that the true likelihood falls below or above its posterior mean will be around 0.5, the difference from this value being due to the severe skewness of the $e^{-1/2\chi_1^2}$ distribution.

If we want a *high* posterior probability that the LR is less than k, the 'critical value' of the PBF will have to be reduced. In general, if the PBF equals k^*, what is the posterior probability that the LR is less than k? As above, this is

$$\begin{aligned} \pi_{k,k^*} &= \Pr((k^*/\sqrt{2})L(\bar{y})/L(\mu) < k|y) \\ &= \Pr(Z^2 < \log 2 - 2\log(k^*/k)) \\ &= 2\Phi\left(\left\{\log 2 - 2\log(k^*/k)\right\}^{1/2}\right) - 1. \end{aligned}$$

Suppose we want $\pi_{0.1,k^*} = 0.9$, i.e. a posterior probability of 0.9 that the LR is less than 0.1. Then k^* has to be 0.0369, much smaller than the conventional values suggested by Aitkin (1991).

We now generalize these results to the large-sample multiparameter case.

3. The multiparameter case

Given a v-dimensional parameter θ in the model $p(y|\theta)$ for data y, we assume the sample size is large and the usual regularity conditions hold. We assume the prior $\pi(\theta)$ is locally diffuse in the neighbourhood of the MLE $\hat{\theta}$. Then the log-likelihood has the quadratic expansion

$$\ell(\theta) = \ell(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})' I (\theta - \hat{\theta})$$

where I is the observed information matrix. A null hypothesis H_0 specifies $\theta = \theta_0$, the alternative $\bar{H}_0$ is the negation of H_0. Our interest is in the posterior distribution of the likelihood ratio $L(\theta_0)/L(\theta)$, or the log-likelihood difference $\ell(\theta_0) - \ell(\theta)$. It follows immediately that

$$\ell(\theta_0) - \ell(\theta) = \frac{1}{2}(\theta - \hat{\theta})' I (\theta - \hat{\theta}) - \frac{1}{2}X^2$$

where X^2 is the usual large-sample test statistic $(\theta_0 - \hat{\theta})' I (\theta_0 - \hat{\theta})$. So

$$2(\ell(\theta_0) - \ell(\theta)) + X^2 \sim \chi^2_v,$$

and the (posterior) probability that the LR $L(\theta_0)/L(\theta)$ is less than k is

$$\pi_{k,v} = \Pr(\chi^2_v < X^2 + 2 \log k).$$

For $k = 1$, this probability is $1 - P$, where P is the P-value of the observed X^2. So again $1 - P$ is the probability that the LR is less than 1, and a typically small P-value cannot be interpreted as convincing evidence against H_0. Table 2 gives $\pi_{k,v}$ for two values of k, 1 and 1/9, for $v = 1(1)5$, and a range 3(1)12 of values of the X^2 test statistic. This table shows an interesting feature. For a fixed P-value P, the posterior probability $\pi_{k,v}$ that the LR $< 1/9$ *increases* steadily with v, as shown in Table 3 for $P = 0.05$, and $v = 1(1)10$. Thus the P-value is a more reliable indicator of the strength of evidence against H_0 for large values of v than for small.

We can express this differently: for what P-value is the posterior probability equal to π that the LR $< k$? This is easily calculated:

$$\begin{aligned}\pi_{k,v} &= \Pr(\chi^2_v < X^2 + 2 \log k) \\ &= F_v(X^2 + 2 \log k) \\ &= F_v(F_v^{-1}(1 - P) + 2 \log k)\end{aligned}$$

where $F_v(x)$ is the c.d.f. of χ^2_v at x, and P is the P-value of X^2. Then

$$P = 1 - F_v(F_v^{-1}(\pi_{kv}) - 2 \log k).$$

Table 4 gives the P-value required to give a posterior probability of 0.9 that the LR $< k$, for a range of k and v.

A similar relation can be established between the PBF and the posterior probability. For the large-sample case above, Aitkin (1991) gave the PBF

$$A = L(\theta_0)/L^A = L(\theta_0)/(2^{-v/2} L(\hat{\theta})) = 2^{v/2} L(\theta_0)/L(\hat{\theta})$$

and so

$$X^2 = -2 \log[L(\theta_0)/L(\hat{\theta})] = v \log 2 - 2 \log A.$$

If the PBF $A = k$, then

$$\begin{aligned}\pi_{kv} &= \Pr(\chi^2_v < X^2 + 2 \log k) \\ &= \Pr(\chi^2_v < v \log 2).\end{aligned}$$

Table 2. *Posterior probability π_{kv} that the LR $< k$ given X^2 and v*

v	$k \backslash X^2$	3	4	5	6	7	8	9	10	11	12
1	1	0.917	0.954	0.975	0.986	0.992	0.995	0.997	0.998	0.999	0.999
	1/9	0	0	0.564	0.795	0.894	0.942	0.968	0.982	0.990	0.994
2	1	0.777	0.865	0.918	0.950	0.970	0.982	0.989	0.993	0.996	0.997
	1/9	0	0	0.261	0.552	0.728	0.835	0.900	0.939	0.963	0.978
3	1	0.608	0.738	0.828	0.888	0.928	0.954	0.971	0.981	0.988	0.993
	1/9	0	0	0.105	0.342	0.543	0.693	0.797	0.867	0.914	0.945
4	1	0.442	0.594	0.713	0.801	0.864	0.908	0.939	0.960	0.973	0.983
	1/9	0	0	0.038	0.192	0.374	0.538	0.670	0.769	0.842	0.993
5	1	0.300	0.451	0.584	0.694	0.779	0.844	0.891	0.925	0.949	0.965
	1/9	0	0	0.012	0.099	0.239	0.393	0.534	0.654	0.748	0.821

Table 3. *Posterior probability that the LR $<$ 1/9, given $P = 0.05$ and v*

v	1	2	3	4	5	6	7	8	9	10
$X^2(P = 0.05)$	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31
$\pi_{1/9,v}$	0	0.550	0.668	0.722	0.754	0.776	0.792	0.805	0.815	0.823

Table 4. *P-value required to give $\pi_{k,v} = 0.9$, for k and v*

$k\backslash v$	1	2	3	4	5	6	7	8	9	10
1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
1/3	0.027	0.033	0.038	0.041	0.043	0.046	0.048	0.049	0.051	0.052
1/9	0.0077	0.011	0.014	0.016	0.018	0.020	0.022	0.023	0.025	0.026
1/27	0.0023	0.0037	0.0050	0.0062	0.0074	0.0085	0.0095	0.0105	0.0115	0.0124

This probability *decreases* steadily with v, as shown in Table 5. Thus the PBF is a *less* reliable indicator of the strength of evidence against H_0 for large values of v than for small. This calibration may be corrected, as for the P-value, by requiring that the PBF should be k^* to give a high posterior probability that the LR $< k$. Then

$$\begin{aligned}\pi_{kv} &= \Pr(\chi^2_v < X^2 + 2\log k)\\ &= \Pr(\chi^2_v < -2\log k^* + v\log 2 + 2\log k)\\ &= \Pr(\chi^2_v < 2\log(k/k^*) + v\log 2).\end{aligned}$$

Table 6 gives the value of $\log(k/k^*)$ required to achieve a posterior probability of 0.9.

Thus for $v = 1$, k^* needs to be e^{-1} times k to give a posterior probability of 0.9, and for $v = 10$ this factor is $e^{-4.5} \simeq 10^{-2}$. Posterior Bayes factors of k need very substantial recalibration in large dimension models to give a high posterior probability that the LR $< k$.

A similar recalibration may be used for other penalized LRT statistics, discussed in Aitkin (1991). For example, Akaike's information criterion (Akaike, 1973) is AIC $= X^2 - 2v$. If the alternative hypothesis model is chosen when AIC ≥ 0, then given AIC ≥ 0, the posterior probability that the LR $< k$ is

$$\begin{aligned}\pi_{kv} &= \Pr(\chi^2_v < X^2 + 2\log k)\\ &\geq \Pr(\chi^2_v < 2v + 2\log k).\end{aligned}$$

The RHS approaches 1 as $v \to \infty$ for any fixed k. Table 7 gives the posterior probability for small v and several values of k.

For $k = 1$ and any v, an AIC of zero or more gives a very high posterior probability that the alternative hypothesis is better supported than the null. For small values of k, AIC is less successful for small v.

Recalibration of the AIC may be simply obtained by choosing the alternative hypothesis model when AIC $\geq -2\log k^*$ where k^* is chosen so that the posterior probability that the LR $< k$ is at least (say) 0.9. Table 8 gives the value of $\log(k/k^*)$ required to achieve a posterior probability of 0.9.

For small v (say $v \leq 5$) the AIC may be used with only small correction, but as v increases it becomes increasingly conservative, requiring unduly large values of X^2 as evidence against H_0. Equivalently, it requires posterior probabilities of almost 1 as the measure of certainty.

4. Bayes factors

It might be expected that Bayes factors could be calibrated similarly. The Bayes factor B for the point null hypothesis $H_0 : \theta = \theta_0$ is $B = L(\theta_0)/\int L(\theta)\pi(\theta)\mathrm{d}\theta$. Here the prior distribution must be proper: the denominator is undefined if

Table 5. *Posterior probability that the LR $< k$, given PBF $= k$*

v	1	2	3	4	5	6	7	8	9	10
$\pi_{k,v}$	0.595	0.500	0.444	0.403	0.371	0.345	0.322	0.302	0.284	0.268

Table 6. *Value of $\log(k/k^*)$ required to give $\pi_{kv} = 0.9$ when PBF $= k^*$*

v	1	2	3	4	5	6	7	8	9	10
$\log(k/k^*)$	1.01	1.61	2.09	2.50	2.89	3.24	3.58	3.91	4.22	4.53

Table 7. *Posterior probability π_{kv} when AIC ≥ 0*

$k\backslash v$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.843	0.865	0.888	0.908	0.925	0.938	0.949	0.958	0.965	0.971	0.976	0.980	0.983	0.986	0.988
1/3	0	0.594	0.716	0.786	0.833	0.867	0.893	0.913	0.929	0.942	0.952	0.960	0.967	0.973	0.977
1/9	0	0	0.342	0.538	0.654	0.732	0.788	0.830	0.863	0.889	0.909	0.925	0.938	0.949	0.958
1/27	0	0	0	0.157	0.363	0.507	0.612	0.691	0.751	0.798	0.835	0.865	0.889	0.908	0.924

Table 8. *Value of* $\log(k/k^*)$ *required to give* $\pi_{kv} \geq 0.9$, *when AIC* $\geq -2\log k^*$

v	1	2	3	4	5	6	7	8	9	10
$\log(k/k^*)$	0.353	0.303	0.126	−0.110	−0.382	−0.678	−0.992	−1.32	−1.66	−2.01
v	11	12	13	14	15	16	17	18	19	20
$\log(k/k^*)$	−2.36	−2.73	−3.09	−3.47	−3.85	−4.23	−4.62	−5.01	−5.40	−5.79

the prior is improper. For the large-sample normal likelihood and the locally diffuse prior, we have

$$\int L(\theta)\pi(\theta)\mathrm{d}\theta = \pi(\hat{\theta})L(\hat{\theta})\int \exp\left\{-\frac{1}{2}(\theta-\hat{\theta})'I(\theta-\hat{\theta})\right\}\mathrm{d}\theta$$
$$= (2\pi)^{v/2}|I|^{-1/2}\pi(\hat{\theta})L(\hat{\theta}).$$

The Bayes factor is then

$$B = \frac{(2\pi)^{-v/2}|I|^{1/2}}{\pi(\hat{\theta})}\frac{L(\theta_0)}{L(\hat{\theta})}.$$

The first factor depends on the sample size through the information matrix I. Assuming that y consists of n independent observations, write $I = nI_0$, where I_0 depends only on the design configuration and not on n. Then

$$B = \frac{(2\pi)^{-v/2}|I_0|^{1/2}}{\pi(\hat{\theta})}n^{v/2}\frac{L(\theta_0)}{L(\hat{\theta})}$$
$$= \frac{(2\pi)^{-v/2}|I_0|^{1/2}}{\pi(\hat{\theta})}n^{v/2}\exp\left(-\frac{1}{2}X^2\right).$$

As $n \to \infty$, if H_0 is true, then X^2 behaves stochastically like χ^2_v, and so $B \to \infty$, while if H_1 is true then X^2 behaves like non-central χ'^2_v with non-centrality parameter proportional to n, and so $B \to 0$ in probability. Thus the Bayes factor has the appealing feature in large samples of correctly identifying the true model (Schwarz, 1978), unlike the PBF, the P-value or the AIC.

This feature is greatly complicated however by the dependence of B on the prior ordinate $\pi(\hat{\theta})$ and the design configuration through $|I_0|$. For example, for a p-variable normal regression model with known variance σ^2 and a locally diffuse prior on β, I is the raw (uncorrected) SSP matrix of the predictors divided by σ^2, and $|I_0|$ is the determinant of the mean-corrected predictor SSP matrix, divided by $\sigma^{2(p+1)}$. So the value of the Bayes factor depends on the measurement units of the response and predictor variables: a rescaling of any predictor variable will scale the Bayes factor proportionately unless the prior distribution is correspondingly rescaled.

These difficulties are often avoided in practice by ignoring the contribution of the first term to B, since it is independent of n. This is equivalent to assuming that $\pi(\hat{\theta}) = |I_0|^{1/2}/(2\pi)^{v/2}$ which requires a peculiar data-based prior, whose precision depends on the experiment to be performed. (Other attempts to avoid Lindley's paradox with the diffuse prior are briefly reviewed in Aitkin, 1991.)

We may ask: if the Bayes factor B equals k, what is the posterior probability that the LR $< k$? We have

$$\pi_{kv} = \Pr(\chi^2_v < X^2 + 2\log k)$$
$$= \Pr(\chi^2_v < \log|I_0| - 2\log\pi(\hat{\theta}) + v\log(n/2\pi)),$$

but to give a numerical value for the posterior probability requires the specification of the information matrix and the (proper) prior density ordinate at $\hat{\theta}$ for the model. However it is clear that for any fixed k, the posterior probability *increases with* n, illustrating again the different behaviour of the Bayes factor from that of the PBF, P-value or AIC.

Some of the difficulties with the Bayes factor may be avoided by the use of the *fractional Bayes factor* (FBF; O'Hagan, 1995). The FBF for the simple null hypothesis is defined by FBF $= L(\theta_0)/q_b$ where

$$q_b = \int L^{1-b}(\theta)\pi_b(\theta\,|\,y)\mathrm{d}\theta$$
$$= \int L^{1-b}(\theta)L^b(\theta)\pi(\theta)\mathrm{d}\theta \Big/ \int L^b(\theta)\pi(\theta)\mathrm{d}\theta$$
$$= \int L(\theta)\pi(\theta)\mathrm{d}\theta \Big/ \int L^b(\theta)\pi(\theta)\mathrm{d}\theta$$

where b is the fraction of data used to convert the prior to a proper posterior, and $1-b$ is the remaining fraction used to define the likelihood; the data for 'prior' and 'likelihood' are assumed to be the same. The FBF does not depend on the information matrix or prior ordinate: as in O'Hagan, for the normal likelihood and diffuse prior,

$$\mathrm{FBF} = \frac{L(\theta_0)}{b^{v/2}L^{1-b}(\hat{\theta})}$$
$$= \frac{L^b(\hat{\theta})}{b^{v/2}}\exp\left(-\frac{1}{2}X^2\right).$$

If the FBF $= k$, the posterior probability that the LR $< k$ is

$$\pi_{kv} = \Pr(\chi^2_v < X^2 + 2\log k)$$
$$= \Pr(\chi^2_v < -v\log b + 2b\ell(\hat{\theta})).$$

O'Hagan recommends that b be chosen as $O(n^{-1})$; if $b = n^{-1}$ then

$$\pi_{kv} = \Pr(\chi^2_v < v\log n + 2\ell(\hat{\theta})/n).$$

This tends to 1 as $n \to \infty$ like the Bayes factor, but its actual value depends on the maximized log-likelihood for the model, so is again model-specific. This value will be affected by multiplicative constants (like $\sqrt{2\pi}$) which might

be omitted from the likelihood, and in the normal regression model by arbitrary re-scaling of the response.

We now extend Dempster's results to composite null hypotheses.

5. Composite hypotheses

In the model $p(y|\theta), \theta$ is now partitioned into the parameter of interest ψ, of dimension v_1, and the nuisance parameter λ of dimension v_2. The composite null hypothesis is $H_0 : \psi = \psi_0$, the alternative being $\bar{H}_0 : \psi \neq \psi_0$. We assume as before a prior for (ψ, λ) which is diffuse in the region of the joint MLE $(\hat{\psi}, \hat{\lambda})$. Our interest is in the likelihood ratio LR, now defined to be $L(\psi_0, \lambda)/L(\psi, \lambda)$.

Let $\hat{\lambda}_0$ be the MLE of λ given $\psi = \psi_0$. Expand the log-likelihoods about $\lambda = \hat{\lambda}_0$ and $(\psi, \lambda) = (\hat{\psi}, \hat{\lambda})$ respectively:

$$\ell(\psi_0, \lambda) = \ell(\psi_0, \hat{\lambda}_0) - \frac{1}{2}(\lambda - \hat{\lambda}_0)' I_0 (\lambda - \hat{\lambda}_0)$$

$$\ell(\theta) = \ell(\psi, \lambda) = \ell(\hat{\psi}, \hat{\lambda}) - \frac{1}{2}(\theta - \hat{\theta})' I (\theta - \hat{\theta})$$

where I_0 is the information

$$I_0 = -\frac{\partial^2 \ell(\psi_0, \lambda)}{\partial \lambda \partial \lambda'}\bigg|_{\lambda = \hat{\lambda}_0}.$$

Then

$$\begin{aligned}\ell(\psi_0, \lambda) - \ell(\psi, \lambda) &= \ell(\psi_0, \hat{\lambda}_0) - \frac{1}{2}(\lambda - \hat{\lambda}_0)' I_0 (\lambda - \hat{\lambda}_0) \\ &\quad - \ell(\hat{\psi}, \hat{\lambda}) + \frac{1}{2}(\theta - \hat{\theta})' I (\theta - \hat{\theta}) \\ &= \frac{1}{2}\left[(\theta - \hat{\theta})' I (\theta - \hat{\theta}) - (\lambda - \hat{\lambda}_0)' I_0 (\lambda - \hat{\lambda}_0)\right] - \frac{1}{2}X^2,\end{aligned}$$

where X^2 is the usual LRTS test statistic for H_0. From the usual regression sum of squares partitioning we have immediately that the posterior distribution of $2(\ell(\psi_0, \lambda) - \ell(\psi, \lambda)) + X^2$ is $\chi^2_{v_1}$ as for the simple null hypothesis, and so all the asymptotic results of Section 3 apply to the composite null hypothesis as well.

Small-sample results are more difficult to establish. We illustrate with the *t*-test.

6. The *t*-test

Given a sample of n observations from $\mathrm{N}(\mu, \sigma^2)$, consider the test of the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $\bar{H}_0 : \mu \neq \mu_0$. The LR is defined by

$$L(\mu_0, \sigma)/L(\mu, \sigma) = \exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{y} - \mu_0)^2 - (\bar{y} - \mu)^2\right]\right\}.$$

The joint prior distribution for $(\mu, \log \sigma)$ is taken to be diffuse; the joint posterior can then be expressed as

$$\mu|\sigma \sim \mathrm{N}(\bar{y}, \sigma^2/n),\ T/\sigma^2 \sim \chi^2_{n-1},$$

where $T = \sum (y_i - \bar{y})^2$. The inequality LR $< k$ is equivalent to

$$n\left[(\bar{y} - \mu_0)^2 - (\bar{y} - \mu)^2\right]/\sigma^2 > -2 \log k.$$

Write $s^2 = T/(n-1)$, $t = \sqrt{n}(\bar{y} - \mu_0)/s$. Then the inequality is

$$\left[s^2 t^2 - n(\bar{y} - \mu)^2\right]/\sigma^2 > -2 \log k$$

or equivalently

$$n(\bar{y} - \mu)^2/\sigma^2 < s^2 t^2/\sigma^2 + 2 \log k.$$

Now $\sqrt{n}(\bar{y} - \mu)/\sigma|\sigma \sim \mathrm{N}(0, 1)$, independently of σ, and so $X_1 = n(\bar{y} - \mu)^2/\sigma^2$ and $X_2 = T/\sigma^2$ are *a posteriori* independently χ^2_1 and χ^2_{n-1}. So

$$\begin{aligned}\pi_k &= \Pr(\mathrm{LR} < k|y) \\ &= \Pr(X_1 < t^2 X_2/(n-1) + 2 \log k).\end{aligned}$$

When $k = 1$, this probability is

$$\begin{aligned}\pi_1 &= \Pr(F_{1,n-1} < t^2) \\ &= 1 - P\end{aligned}$$

where P is the P-value of the observed t, the same result as for the normal case. However for $k \neq 1$, the posterior probability has to be evaluated by numerical integration over the joint density of X_1 and X_2.

We consider finally a very small-sample case in which the effect of the prior cannot be ignored.

7. Risk assessment

Aitkin (1992) and Lindley (1993) discussed a court case brought against the Ministry of Health for performing a vaccination with allegedly defective vaccine, leading to irreversible brain damage in the baby vaccinated (Geisser, 1992 also commented on this case). A statistician expert witness was required to state the evidence for an increase in the risk of brain damage, based on the occurrence of four cases of brain damage.

The analysis was based on a Poisson model: four events had been observed from a Poisson distribution with mean $\lambda = \lambda_0 = 0.9687$ under the null hypothesis H_0 of the standard risk; the alternative was an unspecified higher risk. The P-value of four or more events under H_0 is 0.0171.

The likelihood function is, apart from a multiplicative constant,

$$L(\lambda) = \mathrm{e}^{-\lambda}\lambda^4$$

and the LR is $L(\lambda_0)/L(\lambda)$. The maximized LR is $L(\lambda_0)/L(4) = 0.0713$, which clearly overstates the evidence against H_0 : whatever the true value of λ, the LR must be greater than this.

For the one-sided null hypothesis $H_0^* : \lambda \leq \lambda_0$ against the one-sided alternative $H_1^* : \lambda > \lambda_0$ and the diffuse prior $\mathrm{d}\lambda/\lambda$ the Bayes factor is 0.0174, equal to $P/(1-P)$ (Aitkin,

Table 9. *Gamma prior mean and (variance)*

		v							
		0	0.5		1		1.5		2	
	0	*	*		*		*		*	
	0.1	*	5	(50)	10	(100)	15	(150)	20	(200)
	0.2	*	2.5	(12.5)	5	(25)	7.5	(37.5)	10	(50)
	0.3	*	1.67	(5.56)	3.33	(11.1)	5	(16.7)	6.67	(22.2)
ϕ	0.4	*	1.25	(3.13)	2.5	(6.25)	4.75	(11.9)	5	(12.5)
	0.5	*	1	(2)	2	(4)	3	(6)	4	(8)
	1	*	0.5	(0.5)	1	(1)	1.5	(1.5)	2	(2)
	1.5	*	0.33	(0.22)	0.67	(0.44)	1	(0.67)	1.33	(0.89)
	2	*	0.25	(0.13)	0.5	(0.25)	0.75	(0.38)	1	(0.5)

*, improper prior.

1992), but for the uniform prior $d\lambda$ this Bayes factor is 0.0032, giving a vastly different impression of the strength of evidence against H_0. For the point null hypothesis these improper priors cannot be used as the denominator is undefined. For the conjugate gamma (ϕ, v) prior

$$\pi(\lambda) = \phi^v e^{-\phi\lambda}\lambda^{v-1}/\Gamma(v)$$

the prior mean v/ϕ and variance v/ϕ^2 are given in Table 9 for a small range of values of ϕ and v varying from weak to sharp prior information. The Bayes factor for the point null hypothesis H_0 against the general alternative $\bar{H}_0 : \lambda \neq \lambda_0$ is given in Table 10 for the same range of values of ϕ and v. The Bayes factor is very sensitive to the choice of the hyperparameters, not surprising in view of the small number of events: over the range of hyperparameter values tabled the Bayes factor varies from 0.127 to 5.053. (Lindley, 1993 objected to the use of formal gamma priors in Aitkin, 1992, though he did not say why gamma priors should not be able to represent prior information.) At either of the improper limits $\phi \to 0$ or $v \to 0$ the Bayes factor $\to \infty$. (For the one-sided alternative H_1^* the results are very similar and equally sensitive.)

Lindley (1993) noted that this sensitivity is to be expected in Bayes factors: indeed, we should be concerned if it was absent. Lindley proposed to resolve this dependence by elicitation of the court's prior: the statistician should

Table 10. *Bayes factor for H_0 against $\bar{H}_0$*

		v				
		0	0.5	1	1.5	2
	0	*	*	*	*	*
	0.1	*	0.247	0.224	0.302	0.493
	0.2	*	0.259	0.173	0.173	0.208
	0.3	*	0.303	0.172	0.146	0.149
ϕ	0.4	*	0.366	0.187	0.142	0.131
	0.5	*	0.447	0.212	0.149	0.127
	1	*	1.152	0.446	0.256	0.178
	1.5	*	2.568	0.907	0.476	0.302
	2	*	5.053	1.692	0.842	0.508

'judge the opinion of the court as to how far the Ministry could have erred' in departing from the null λ_0. 'Alternatively, [the statistician] could ask the court's opinion' on this matter. Judges no doubt have an opinion on such matters prior to the data being considered in the case. But should this opinion be so influential in the statistician's formulation of the evidence which the judges are then to assess? Might not the judges want to know what the data say to the statistician, separately from their own opinion?

Table 11 gives the posterior mean and variance, and Table 12 gives the PDF for the same range of hyper-

Table 11. *Gamma posterior mean and (variance)*

		v								
		0		0.5		1		1.5		2	
	0	4	(4)	4.5	(4.5)	5	(5)	5.5	(5.5)	6	(6)
	0.1	3.64	(3.31)	4.09	(3.71)	4.55	(4.14)	5	(4.55)	5.45	(4.95)
	0.2	3.33	(2.78)	3.75	(3.13)	4.17	(3.47)	4.58	(3.82)	5	(4.17)
	0.3	3.08	(2.37)	3.46	(2.66)	3.85	(2.96)	4.23	(3.25)	4.62	(3.55)
ϕ	0.4	2.86	(2.04)	3.21	(2.29)	3.57	(2.55)	3.93	(2.81)	4.29	(3.06)
	0.5	2.67	(1.78)	3	(2)	3.33	(2.22)	3.67	(2.44)	4	(2.67)
	1	2	(1)	2.25	(1.13)	2.5	(1.25)	2.75	(1.38)	3	(1.5)
	1.5	1.6	(0.64)	1.8	(0.72)	2	(0.8)	2.2	(0.88)	2.4	(0.96)
	2	1.33	(0.44)	1.5	(0.5)	1.67	(0.56)	1.83	(0.61)	2	(0.67)

Table 12. *Posterior Bayes factor for H_0 against $\bar{H}_0$*

		v				
		0	0.5	1	1.5	2
	0	0.102	0.100	0.102	0.106	0.113
	0.1	0.103	0.099	0.098	0.100	0.104
	0.2	0.105	0.099	0.097	0.096	0.098
	0.3	0.109	0.101	0.097	0.095	0.095
ϕ	0.4	0.114	0.104	0.098	0.094	0.093
	0.5	0.120	0.108	0.100	0.095	0.093
	1	0.163	0.139	0.122	0.111	0.102
	1.5	0.229	0.189	0.161	0.140	0.125
	2	0.322	0.259	0.215	0.183	0.159

parameter values. It is well-defined for the improper prior limits; for the two improper priors $d\lambda/\lambda$ and $d\lambda$ the PBF is 0.102. The PBF for this example is much more stable than the Bayes factor, with a range of 0.093 to 0.322 over the same hyperparameter values. The PBF is always substantially smaller than the Bayes factor, especially for the informative priors with large ϕ and small v, where the Bayes factor is greater than 1, while the PBF is less than 0.25.

To compute the posterior probability that the LR $< k$, there is a technical difficulty because the LR is not an analytically invertible function of λ. This difficulty can be avoided by using the likelihood-normalizing cube-root transformation of λ (Anscombe, 1964). Write $\phi = \lambda^{1/3}$; then the likelihood can be approximated to a very high accuracy by

$$L(\lambda) \doteq L(\hat{\lambda}) \exp\left\{-\frac{(\phi - \hat{\phi})^2}{2\sigma^2}\right\}$$

where

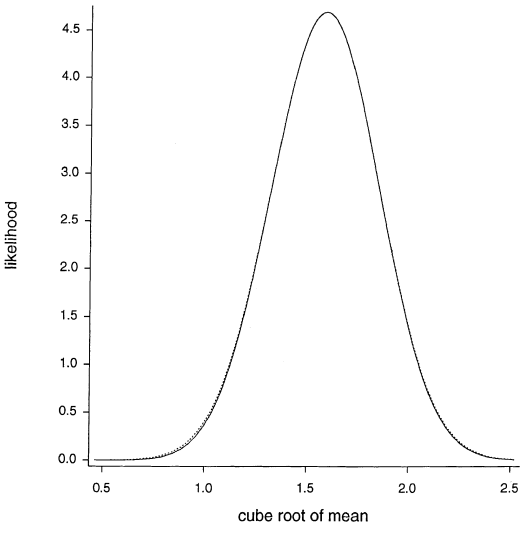

Fig. 1. *Likelihood (solid line) and normal approximation (dotted line) for four events*

$$\hat{\phi} = \hat{\lambda}^{1/3} = 4^{1/3} = 1.587,$$
$$\sigma^2 = \left[6\hat{\lambda}^{1/3}(1 + 2/\hat{\lambda})\right]^{-1} = (0.265)^2.$$

The exact and approximating likelihoods are shown on the ϕ scale in Figure 1.

The posterior distribution of λ is gamma $(1 + \phi, 4 + v)$ and the likelihood ratio $L(\lambda_0)/L(\lambda)$ is approximated by

$$\frac{L(\lambda_0)}{L(\hat{\lambda})} \exp\left\{\frac{(\lambda^{1/3} - 1.587)^2}{2 \times 0.265^2}\right\}.$$

The posterior probability that the LR $< k$ is then approximately

$$\begin{aligned}\pi_k &= \Pr\left(\frac{(\lambda^{1/3} - 1.587)^2}{0.265^2} < 2 \log \frac{kL(\hat{\lambda})}{L(\lambda_0)}\right) \\ &= \Pr\left(\frac{|\lambda^{1/3} - 1.587|}{0.265} < [2 \log k + 5.282]^{1/2}\right) \\ &= \Pr\Big(\left(1.587 - 0.265[2 \log k + 5.282]^{1/2}\right)^3 \\ &< \lambda < \left(1.587 + 0.265[2 \log k + 5.282]^{1/2}\right)^3\Big)\end{aligned}$$

which can be calculated directly from the gamma cdf. Tables 13 and 14 show the posterior probability π_k for

Table 13. *Posterior probability that the LR < 1*

		v				
		0	0.5	1	1.5	2
	0	0.978	0.981	0.977	0.967	0.952
	0.1	0.976	0.985	0.986	0.982	0.974
	0.2	0.971	0.984	0.989	0.989	0.986
	0.3	0.964	0.981	0.990	0.992	0.992
ϕ	0.4	0.955	0.977	0.988	0.993	0.995
	0.5	0.945	0.971	0.985	0.992	0.995
	1	0.878	0.927	0.958	0.977	0.987
	1.5	0.790	0.860	0.911	0.945	0.967
	2	0.688	0.776	0.845	0.897	0.933

Table 14. *Posterior probability that the LR < 0.1*

		v				
		0	0.5	1	1.5	2
	0	0.581	0.595	0.580	0.543	0.488
	0.1	0.572	0.609	0.616	0.598	0.559
	0.2	0.550	0.605	0.633	0.635	0.614
	0.3	0.518	0.587	0.633	0.655	0.652
ϕ	0.4	0.480	0.559	0.620	0.658	0.674
	0.5	0.438	0.524	0.596	0.649	0.681
	1	0.243	0.323	0.407	0.491	0.568
	1.5	0.117	0.169	0.231	0.301	0.377
	2	0.052	0.080	0.117	0.164	0.219

$k = 1$ and $k = 0.1$ for the same range of values of ϕ and v. The values $\phi = v = 0$ correspond to the diffuse prior $d\lambda/\lambda$ and the values $\phi = 0$, $v = 1$ to the prior $d\lambda$.

There is strong evidence (a high posterior probability) that the LR is < 1 except for the strongly informative priors with ϕ large and v small. But the evidence that the LR < 0.1 is weak, with a posterior probability of at most 0.68 over the table values. Interestingly, the two diffuse priors give the same posterior probability: 0.98 for $k = 1$ and 0.58 for $k = 0.1$.

Over a large range of hyperparameter values, the posterior probability that the LR $< k$ is around 0.595, which is the posterior probability that for a normal likelihood the LR $< k$ when the PBF $= k$. Here the PBF $\doteq 0.1$ for a large range of hyperparameter values. If we take $k = 0.1$ as a convincing LR, then the strength of the evidence in support of this or a smaller value is rather weak. This is not surprising as the maximized likelihood ratio is 0.0713.

8. Discussion

Dempster's approach to posterior inference about a point null hypothesis can be straightforwardly extended to composite (point) null hypotheses. Regarding the likelihood ratio as the inferential construct of interest leads to a simple interpretation in large samples of 1 minus the *P*-value as the posterior probability that the likelihood ratio is less than 1. Since a likelihood ratio of less than 1 does not constitute convincing evidence against a simple null hypothesis in favour of a simple alternative, it is clear that *P*-values need recalibration as measures of strength of evidence. In large samples the calibration is quite straightforward, leading to the inferential pairs (or function) (k, π_k) as the summary of the evidence, as proposed by Dempster for simple null hypotheses. This recalibration is relatively less severe in high-dimensional than in low-dimensional models.

Posterior Bayes factors can be recalibrated similarly, as can any other penalized LR statistic, such as the AIC. Recalibration of the posterior Bayes factor is relatively more severe in high-dimensional than in low-dimensional models. Recalibration of the AIC makes it more *generous* in high-dimensional models. On recalibration, *all* the penalized LR criteria lead to the same conclusions about the LR as the *P*-value, that is, they are (asymptotically) equivalent in the information they provide about the LR.

Bayes factors (or the Bayesian information criterion, BIC) and fractional Bayes factors cannot be calibrated in this way as they depend on specific features of the prior and the model information matrix, or the scaling of the variables.

The sensitivity of Bayes factors to hyperparameter variations means that a routine sensitivity analysis is always required as part of their use, and a corresponding serious effort to 'get the prior right', apparently by eliciting judges' opinions in the court case example. The use of posterior tail probabilities of the likelihood ratio is, by contrast, quite robust to hyperparameter variations, even in this very small sample, over the range $0 \leq \phi \leq 0.5$, $0 \leq v \leq 2$. Only for strongly informative priors do they vary substantially. Thus 'non-informative' reference prior analyses using the posterior distribution of the likelihood ratio do not suffer from the well-known difficulties of Bayes factors, while still providing a fully Bayesian analysis.

Acknowledgements

I am grateful to Peter Walley, Ken Brewer and Jay Kadane for helpful discussions.

References

Aitkin, M. (1991) Posterior Bayes factors (with Discussion). *Journal of the Royal Statistical Society*, B **53**, 111–42.

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

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. 267–81. Akademiai Kiado, Budapest.

Anscombe, F. J. (1964) Normal likelihood functions. *Annals of the Institute of Statistical Mathematics*, **26**, 1–19.

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

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Springer-Verlag, New York.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of *P*-values and evidence. *Journal of the American Statistical Association*, **82**, 112–22.

Dempster, A. P. (1974) The direct use of likelihood for significance testing. In *Proc. Conf. Foundational Questions in Statistical Inference* (eds O. Barndorff-Nielsen, P. Blaesild and G. Sihon), pp. 335–52. University of Aarhus.

Geisser, S. (1992) Some statistical issues in medicine and forensics. *Journal of the American Statistical Association*, **87**, 607–14.

Lindley, D. V. (1957) A statistical paradox. *Biometrika*, **44**, 187–92.

Lindley, D. V. (1993) On the presentation of evidence. *Mathematical Scientist*, **18**, 60–63.

O'Hagan, A. (1995) Fractional Bayes factors for model comparisons (with Discussion). *Journal of the Royal Statistical Society*, B **57**, 99–138.

Schwartz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 461–64.