Bayesian point null hypothesis testing via the posterior likelihood ratio

Richard Boys and Tom Chadwick
Newcastle University, School of Mathematics & Statistics
Newcastle upon Tyne NE1 7RU, United Kingdom
richard.boys@ncl.ac.uk, t.j.chadwick@ncl.ac.uk

Murray Aitkin
Melbourne University, Department of Psychology
Melbourne, Australia
murray.aitkin@ncl.ac.uk

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. We follow the Bayesian approach by Dempster (1974, 1997) and Aitkin (1997) and use the likelihood ratio between the null and alternative hypotheses to provide a different evaluation of the point null hypothesis, one in which frequentist and Bayesian conclusions are much closer. In this paper, we outline the key ideas behind this approach to hypothesis testing and contrast this approach with frequentist P-values and Bayesian analyses based on usual Bayes factors.

Simple null hypotheses

In the example due to Stone (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

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

is maximized at $\theta = \widehat{\theta} = 0.201652$ $[SE(\widehat{\theta}) = 0.0005526]$. The standardized departure from the null hypothesis is $Z_1 = |\theta_1 - \widehat{\theta}|/SE(\widehat{\theta}) = 2.9895$, with a two-sided P-value of 0.0028, indicating strong evidence against the null hypothesis. The maximized likelihood ratio is $L(\theta_1)/L(\widehat{\theta}) = 0.01146$.

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

$$B = \frac{L(\theta_1)}{\int_0^1 L(\theta)\pi(\theta)\mathrm{d}\theta} \approx \frac{1}{\sqrt{2\pi}SE(\widehat{\theta})} \cdot \frac{L(\theta_1)}{L(\widehat{\theta})} = 8.27,$$

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

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

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 $\widehat{\theta} \pm 2.576 SE(\widehat{\theta}) = (0.20023, 0.20308)$ 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, $LR(\theta)$ is also unknown, but is a function of θ and so, given the data, it has a posterior distribution $\pi[LR(\theta)\,|\,\mathbf{y}]$. We may therefore find its posterior percentiles, and so can find $\Pr[LR(\theta) < 0.1\,|\,\mathbf{y}]$ 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 "deviance" as

$$D(\theta) = -2\log LR(\theta) = Z_1^2 - Z^2,$$

where $Z = [\theta - \widehat{\theta}]/SE(\widehat{\theta})$ 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

$$\Pr[LR(\theta) < 0.1\,|\,\mathbf{y}] = \Pr[D(\theta) > 4.605\,|\,\mathbf{y}] = \Pr[\chi_1^2 < 4.331] = 0.9626,$$

while

$$\Pr[LR(\theta) < 1\,|\,\mathbf{y}] = \Pr[D(\theta) > 0\,|\,\mathbf{y}] = \Pr[\chi_1^2 < 2.9895^2] = 0.9972 = 1 - P$$

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).

General point null hypothesis testing problems

Consider a family of models M, determined by a probability model $f(y\,|\,\boldsymbol{\eta})$ depending on a vector-valued parameter $\boldsymbol{\eta}^T = (\boldsymbol{\theta}^T, \boldsymbol{\phi}^T)$. It is helpful to consider the probability model in the context of a large but finite population of N members, in which $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ 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 $L(\boldsymbol{\theta}, \boldsymbol{\phi}) = f(\mathbf{y} \,|\, \boldsymbol{\theta}, \boldsymbol{\phi})$. In our analysis there are *true values* of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$; the prior distribution for these parameters represents our uncertainty about these true values. We consider a null hypothesis H_1 which specifies the value $\boldsymbol{\theta}_1$ of $\boldsymbol{\theta}$, while $\boldsymbol{\phi}$ is unspecified. The alternative hypothesis H_2 has *either* $\boldsymbol{\theta}$ completely unspecified, *or* taking a different specified value $\boldsymbol{\theta}_2$. In either case $\boldsymbol{\phi}$ is unspecified. The joint prior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is $\pi(\boldsymbol{\theta}, \boldsymbol{\phi})$. 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.

If the true value of $\boldsymbol{\phi}$, and the true value of $\boldsymbol{\theta}$ 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

$$LR = LR(\boldsymbol{\theta}, \boldsymbol{\phi}) = L(\boldsymbol{\theta}_1, \boldsymbol{\phi})/L(\boldsymbol{\theta}, \boldsymbol{\phi}),$$

where the dependence of LR on the data $\mathbf{y}$ and the known value $\boldsymbol{\theta}_1$ are suppressed, and the values of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ 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 $\boldsymbol{\phi}$, evaluated at the null hypothesis value $\boldsymbol{\theta}_1$ and at the true value of $\boldsymbol{\theta}$. Though the true values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are unknown, their posterior distribution is known and therefore so is the posterior distribution of LR. In particular, we may evaluate the posterior probability $Pr[LR < k \,|\, \mathbf{y}]$ 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" $D = -2 \log LR$.

Aitkin (1997) showed that Dempster's result for a p-parameter simple null hypothesis with normal likelihoods and flat priors, namely

$$P[LR < k \,|\, \mathbf{y}] = F_p[F_p^{-1}(1 - P) + 2 \ln k]$$

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^2_p distribution). In particular, for $k = 1$,

$$Pr[LR < 1 \,|\, \mathbf{y}] = 1 - P,$$

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. However simulation-based approaches can be used to obtain the posterior distribution of LR or D, and thereby an assessment of the null hypothesis.

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 $\overline{y} = 0.4$, 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 $t = 2.0$, with a two-sided P-value of 0.057 from the t_{24} distribution.

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

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

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

$$Pr[LR < 1 \,|\, y] = Pr[D > 0 \,|\, y] = Pr[Z^2/W < t^2 \,|\, y] = 1 - P,$$

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. However, it is easily simulated using independent random iterates for W and Z, and computing the value of $D = t^2W - Z^2$ for the observed t. For example, 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. Thus 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. Note that the Bayes factor cannot be computed here due to the diffuse prior on μ.

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.

Our approach is quite general and widely applicable; see, for example, Aitkin, Boys and Chadwick (2005) for a range of examples of the standard frequentist hypothesis testing kind. This work also extends the "nested model" approach to encompassing models, and shows that for the normal multiple regression model, straightforward posterior simulation methods give Bayesian analogues to backward elimination in frequentist theory.

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.

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., Boys, R.J. and Chadwick, T.J. (2005) Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statist. and Computing.* To appear.

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

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

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.

Stone, M. (1997) Discussion of Aitkin (1997). *Statist. and Computing* **7**, 263-264.

RÉSUMÉ