Bayesian analysis of 2x2 contingency tables from comparative trials

Murray Aitkin and Tom Chadwick
School of Mathematics and Statistics, University of Newcastle UK

June 18, 2003

Abstract

This paper presents a Bayesian analysis of 2x2 tables arising from randomized experiments, using conjugate Beta priors, with particular emphasis on reference uniform priors for the two binomial proportions. Marginal posterior distributions for the attributable risk, the risk ratio, the odds ratio and the number needed to treat, are all very easily calculated by simulation from the respective beta posterior distributions.

The posterior distribution of the likelihood ratio between the null model and the full model is also easily calculated for any of these parametric functions, and can be expressed in a nearly-orthogonal parametrization.

1 Introduction

The Bayes analysis of contingency tables has received relatively little attention compared with that given to normal models. The first detailed analysis using the natural conjugate Beta priors for the response probabilities under the two treatment conditions was given by Altham (1969), in the context of the posterior probability that the odds ratio was less than 1, as an alternative to the Fisher exact probability of the table. Much more recently formal Bayes analyses have appeared in the medical statistics literature (Hashemi et al 1997), and the Web book by Harrell (2000) gives a detailed exposition of, and argument for, Bayesian analysis in clinical trials.

Bayesian simulation approaches to posterior density computation have become widespread with the development of Markov Chain Monte Carlo methods, so it is rather surprising that the extremely simple simulation approach to 2x2 tables seems to have been overlooked.

We give in Section 2 the conjugate analysis with general Beta priors and describe the simple simulation approach to inference about the usual measures of difference beweeen the response probabilities. We argue that in many randomized trials the uniform prior on both parameters is appropriate.

Section 3 compares this approach with the usual Fisher exact test on an extreme table from the ECMO study (Bartlett et al 1985). Despite the very small numbers in the trial and the flat priors used, the conclusion from the Bayes analysis is very clear: that the ECMO treatment is superior to the conventional treatment, though the extent of the improvement is poorly defined. This result contrasts with the Fisher "exact" probability for the table which does not reach conventional levels of significance.

Section 4 gives the posterior distribution of the likelihood ratio between the null and saturated models, discussed at length in other models in Aitkin (1997) and Aitkin, Boys and Chadwick (2003).

Section 5 gives discussion and conclusions.

2 Conjugate analysis of the 2x2 table

In the 2x2 randomized clinical trial, subjects are randomized to one of two treatment conditions, giving n_1 patients in treatment 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 sucesses. What can be said, in a Bayesian framework, about:

- the attributable risk $\Delta = p_1 - p_2$;
- the risk ratio $\rho = p_1/p_2$;
- 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 we use 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 very complex and involve sums of hypergeometric probabilities, as for the odds ratio discussed by Altham (1969). 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 function. This is an extremely simple calculation. We illustrate with the following table, from the ECMO study of Batrlett et al. This study compared the ECMO (extra corporeal membrane

Table 1: ECMO trial outcome

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

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 in the next section; 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 intially; we show the effect of non-uniform priors below, and comment in Section 4 on the general use of uniform priors.

The posterior distribution of p_1 (for ECMO) is then

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

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

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

Thus there is *very* strong evidence that ECMO is better. Altham (1969) gave this probability calculation for the general 2x2 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 $a_j = b_j = p$ for $0 \le p \le 1$, but this result need not hold for $p > 1$.

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 Figures 1–4, on log scales for the risk and odds ratios. Figure 5 gives a kernel estimate of the density of the attributable risk, with a bandwidth of 0.1. The empirical probability that the attributable risk is positive is 0.9893, in close agreement with the theoretical value. The same probability applies to the risk ratio and odds ratio being greater than 1, or the *nnt* being positive. Equal-tailed 95% credible intervals for the four measures are:

- attributable risk – (0.069, 0.948);
- risk ratio – (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 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.

3 Fisher exact test

The standard test for the 2x2 table, especially with small samples, is Fisher's "exact" test, based on the conditional hypergeometric distribution of R_1 given the marginal total $R = r = r_1 + r_2$. This is

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

where

$$u_1 = \max(0, r - n_2), u_2 = \min(n_1, r).$$

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 (0.0833/1), 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^2_1 distribution, if this were valid.

4 Posterior distribution of the likelihood ratio

The posterior distribution of the parameters p_1, p_2 maps directly into that of the likelihood ratio between the null hypothesis model and the alternative hypothesis model (Aitkin 1997, Aitkin, Boys and Chadwick 2003). The mapping however requires a choice of parameterization for the nuisance intercept parameter in the regression model for the 2×2 table. Aitkin, Boys and Chadwick discuss the importance of the information-orthogonal parametrization, giving a diagonal (observed or expected) information matrix.

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

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

with

$$\beta_0 = (n_1 p_1 + n_2 p_2)/n.$$

The likelihood in the regression parameters is

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

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,

$$LR = \frac{(n_1 p_1 + n_2 p_2)^r [n_1(1-p_1) + n_2(1-p_2)]^{n-r}/n^n}{p_1^{r_1}(1-p_1)^{n_1-r_1} p_2^{r_2}(1-p_2)^{n_2-r_2}}.$$

Figures 6 and 7 show the empirical cdf of the likelihood ratio and the corresponding deviance $(-2\log LR)$ 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 posterior probability that $LR > 1$ of 0.0107 is substantially below the Fisher P-value, but greater than the P-value from the unconditional LR test using the asymptotic χ^2_1 distribution.

5 Discussion and conclusions

The difficulty of calibrating the unconditional test, and its size-dependence 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 6 (at least) p-values which have been proposed for this table, ranging from 0.001 to 0.62 (see Begg 1990 and the discussion, and Ware 1989 and the discussion for the range of p-values and arguments for them).

It will 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? 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, or not 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 biases the result of the randomized trial – in such trials it seems to us critical to "let the data speak" without changing its information content by introducing informative priors.

We illustrate this point by a second analysis of the 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 – $(0.097, 0.993)$;
- risk ratio – $(1.11, 2452)$;
- 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 posteriors have infinite spikes at their former modes of 1 and 0, accentuating the information in the likelihood, which makes the Jeffreys prior results hard to justify.

6 References

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

Aitkin, M., Boys, R.J. and Chadwick, T.J. (2003) An ABC(D) of Bayesian point null hypothesis testing. Submitted.

Altham, P.M.E. (1969) Exact Bayesian analysis of a 2x2 contingency table, and Fisher's "exact" test. *J. Roy. Statis. 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.

Harrell, F. (2000) *Practical Data Analysis from a Former Frequentist* 120pp. http://hesweb1.med.virginia.edu/biostat/teaching/bayes.short.course.pdf

Hashemi, L., Balgobin, N, and Goldberg, R. (1997) Bayesian analysis for a single 2x2 table. *Statist. in Med. 16*, 1311-1328.

Hutton, J.L. (2000) Number needed to treat: properties and problems (with comments). *J. Roy. Statist. Soc. A, 163*, 403-419.

Plackett, R.L. (1977) The marginal totals of a 2×2 table. *Biometrika 64*, 37-42.

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

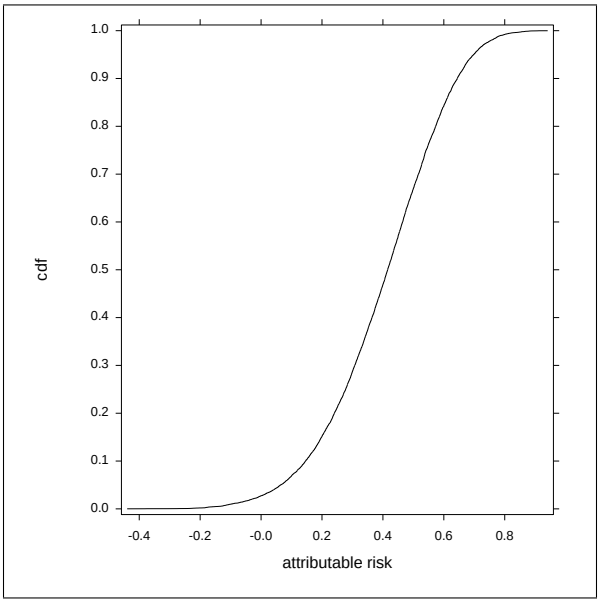

Figure 1: Posterior distribution of attributable risk

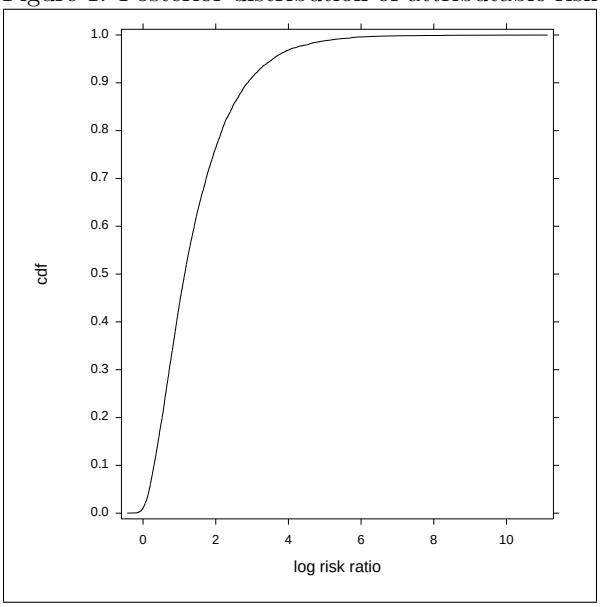

Figure 2: Posterior distribution of log relative risk

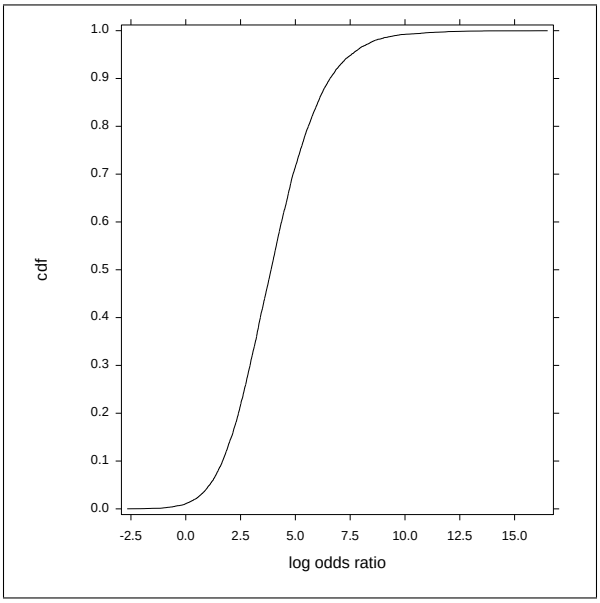

Figure 3: Posterior distribution of log odds ratio

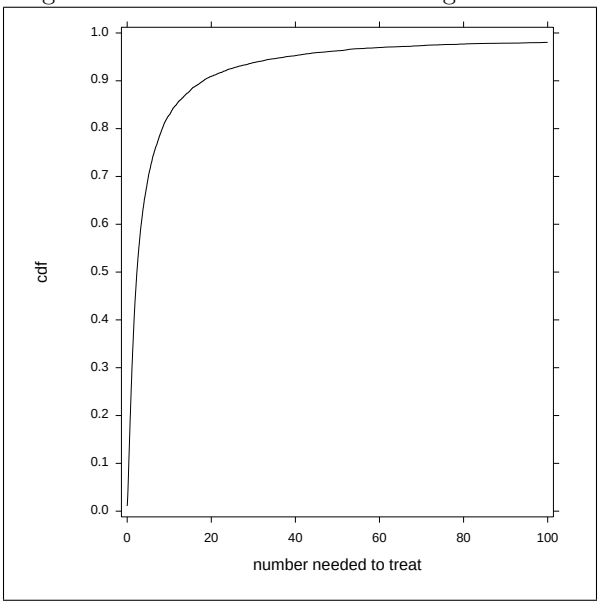

Figure 4: Posterior distribution of number needed to treat

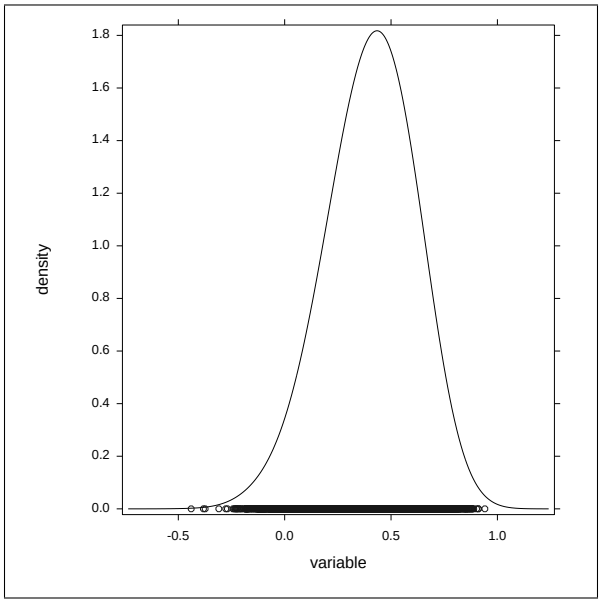

Figure 5: Kernel density for attributable risk

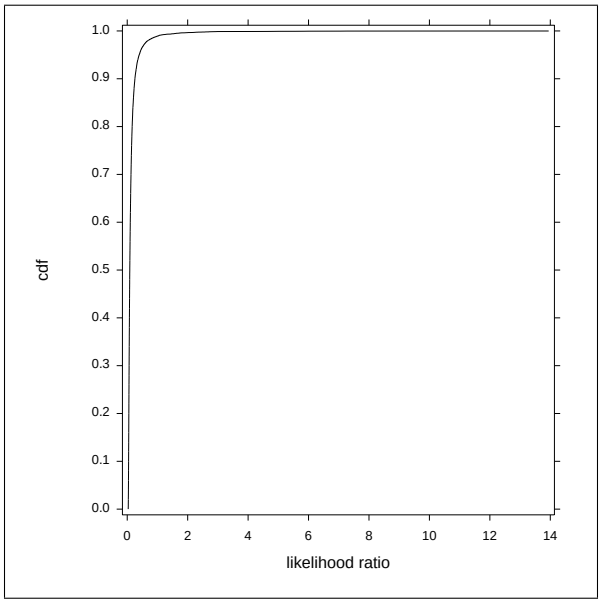

Figure 6: Posterior distribution of likelihood ratio

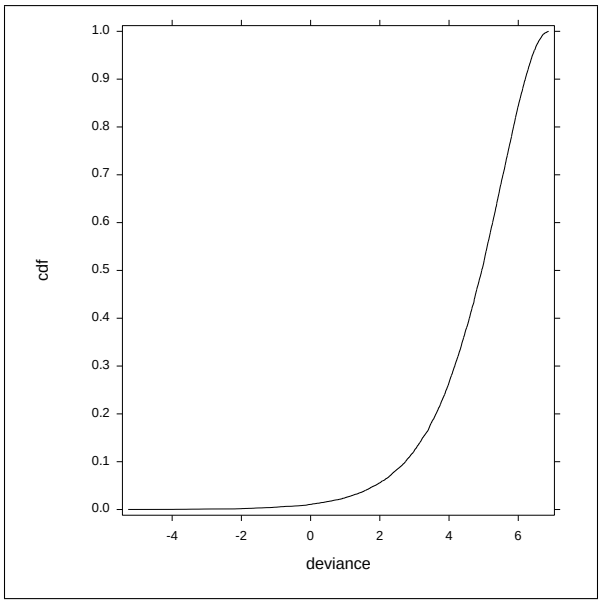

Figure 7: Posterior distribution of deviance