Biometrics 61, 515–523
June 2005

DOI: 10.1111/j.1541-0420.2005.031228.x

Frequentist Performance of Bayesian Confidence Intervals for Comparing Proportions in 2 × 2 Contingency Tables

Alan Agresti

Department of Statistics, University of Florida, Gainesville, Florida 32611, U.S.A.
email: aa@stat.ufl.edu

and

Yongyi Min

UN Statistics Division, Environment Statistics Section, DC2-1404,
2 United Nations Plaza, New York, New York 10017, U.S.A.

Summary. This article investigates the performance, in a frequentist sense, of Bayesian confidence intervals (CIs) for the difference of proportions, relative risk, and odds ratio in 2 × 2 contingency tables. We consider beta priors, logit-normal priors, and related correlated priors for the two binomial parameters. The goal was to analyze whether certain settings for prior parameters tend to provide good coverage performance regardless of the true association parameter values. For the relative risk and odds ratio, we recommend tail intervals over highest posterior density (HPD) intervals, for invariance reasons. To protect against potentially very poor coverage probabilities when the effect is large, it is best to use a diffuse prior, and we recommend the Jeffreys prior. Otherwise, with relatively small samples, Bayesian CIs using more informative (even uniform) priors tend to have poorer performance than the frequentist CIs based on inverting score tests, which perform uniformly quite well for these parameters.

Key words: Binomial distribution; Difference of proportions; Odds ratio; Relative risk; Score confidence interval; Small sample.

1. Introduction

The methodology using the Bayesian paradigm has advanced tremendously in the past decade. New computational methods such as Markov chain Monte Carlo (MCMC) make it easier to evaluate posterior distributions for model parameters. However, Bayesian inference does not seem to be used much yet in practice for basic inference in 2 × 2 contingency tables, which are ubiquitous in biometric applications. For such data, computations are not especially difficult and do not require new methods.

Very little of the literature on Bayesian inference for 2 × 2 contingency tables refers to interval estimation. One of the authors of this article is preparing a survey paper on Bayesian inference for contingency tables; of a few hundred papers on this topic, only a handful (mentioned below) focus on interval estimation of parameters for 2 × 2 tables. Some of the most commonly cited articles (e.g., Good, 1956; Altham, 1969) on Bayesian inference for contingency tables deal with point estimation or significance testing and connections between the Bayesian results and the results based on a frequentist approach.

For instance, consider two independent samples, with X_i a binomial bin(n_i, p_i) variate, $i = 1, 2$. Altham (1969) discussed inference for the odds ratio with a multinomial sample over the four cells of a 2 × 2 table. In the context of comparing parameters for two independent binomial samples, Altham's results correspond to testing of $\mathrm{H}_0: p_1 \leq p_2$ against $p_1 > p_2$ using independent beta(a_i, b_i) priors for p_1 and p_2. She showed that the posterior probability that $p_1 \leq p_2$ equals the one-sided P-value for Fisher's exact test when one uses improper prior distributions $(a_1, b_1) = (1, 0)$ and $(a_2, b_2) = (0, 1)$ favoring the null. Howard (1998) showed that with independent beta(0.5, 0.5) priors the posterior probability that $p_1 \leq p_2$ approximately equals the one-sided P-value for the large-sample z-test using pooled variance for testing $\mathrm{H}_0: p_1 = p_2$ against $\mathrm{H}_a: p_1 > p_2$.

This article investigates the performance, in a frequentist sense, of Bayesian confidence intervals (CIs) for three of the most common measures used in biometrics—the difference of proportions $p_1 - p_2$, the ratio p_1/p_2 (the "relative risk"), and the odds ratio, $[p_1/(1 - p_1)]/[p_2/(1 - p_2)]$. In many applications, there are no obvious prior distributions to use for p_1 and p_2. The main goal of this article is to consider whether certain prior distributions tend to work well regardless of the true association parameter values. Hence, the corresponding CIs might be "default" ones that software could report unless the user prefers to select particular prior distributions. According to this criterion, our results show that it is best to use highly diffuse priors, perhaps even more diffuse than the uniform prior. In addition, we argue that CIs using the tail method are more sensible than using highest posterior density (HPD) intervals. We have prepared software (R functions), available at a website, for constructing tail-method intervals with independent beta priors.

2. Prior Distributions for Binomial Probabilities

2.1 *Beta Prior Distributions*

The family of beta densities is conjugate for the binomial parameter and has received by far the most attention for this problem. It provides a flexible family of priors, as prior parameters can be selected to give various shapes with various degrees of skew. The beta priors with parameters $a = b = 0.5$, 1, 2.0 are symmetric about 0.5, with U shape, uniform, and bell shape, and standard deviations 0.35, 0.29, and 0.22.

Complete prior ignorance might suggest a uniform prior distribution, $a = b = 1$. Alternatively, a popular prior with Bayesians is the Jeffreys prior. Unlike a uniform prior, it is still the appropriate prior for a one-to-one transformation of the parameter space (e.g., Box and Tiao, 1973, p. 32, 41–42). This prior is proportional to the square root of the determinant of the Fisher information matrix for the parameters of interest, for a single observation. For a binomial parameter, the Jeffreys prior equals the beta prior with $a = b = 0.5$ (Box and Tiao, 1973, p. 34–38). This prior is also the reference prior, being approximately noninformative in the sense of Bernardo's reference analysis approach (Bernardo and Smith, 1994, p. 315). Brown, Cai, and DasGupta (2001) showed that the posterior distribution generated by this prior yields a CI for a single binomial parameter p that performs well.

For two independent binomial samples, we consider a beta$(a_i,\ b_i)$ prior for p_i, $i = 1, 2$. For simplicity, we shall treat p_1 and p_2 as independent with the same beta priors. There are corresponding priors for the measures themselves. For instance, with uniform priors for p_1 and p_2, $p_1 - p_2$ has a triangular density over $(-1, +1)$, $r = p_1/p_2$ has density $g(r) = 1/2$ for $0 \leq r \leq 1$ and $g(r) = 1/2r^2$ for $r > 1$, and the log odds ratio has the Laplace density (Nurminen and Mutanen, 1987). The independent posterior distributions for p_1 and p_2 induce posterior distributions for their difference, ratio, and the odds ratio.

2.2 *Normal Prior Distributions for Logits*

An alternative approach specifies priors for the logit, the natural parameter in the exponential family representation of the binomial distribution. With an $N(0, \sigma^2)$ prior distribution for $\log[p_i/(1 - p_i)]$, on the p_i scale the shape of this density is symmetric, being unimodal when $\sigma^2 \leq 2$ and bimodal when $\sigma^2 > 2$, but always tapering off toward 0 as p_i approaches 0 or 1. Specifically, it is mound-shaped for $\sigma = 1$, roughly uniform except near the boundaries when $\sigma \approx 1.5$, and with more pronounced peaks for the modes when σ is about 2 or larger. The peaks for the modes get closer to 0 and 1 as σ increases further, and the curve has appearance that is essentially U-shaped when $\sigma = 3$ and similar to that of a beta(0.5, 0.5) prior. For $\sigma = 1, 2, 3$, the standard deviations on the p_i scale of these priors are 0.21, 0.31, and 0.37, similar to the values of 0.22, 0.29, and 0.35 for the beta priors mentioned above with $a = b = 2.0, 1, 0.5$. The logit-normal prior with $\sigma = 2.67$ matches the Jeffreys prior in the first two moments (on the probability scale), and the logit-normal prior with $\sigma = 1.69$ matches the uniform prior in the first two moments.

Using a logit-normal prior connects this Bayesian approach with models for the log odds ratio that use normal priors for the parameters of the saturated log-linear model (e.g., Leonard, 1975). This builds on the work of Lindley (1964), who had considered approximations for the posterior distribution of contrasts of log probabilities, such as the log odds ratio. A hierarchical version puts second-stage priors on the parameters of the prior distribution (Leonard, 1972, 1975). The logit-normal prior also has a natural connection with Bayesian approaches for logistic regression (e.g., Wong and Mason, 1985).

3. Evaluating Posterior Distributions and Confidence Intervals for Measures of Association

Suppose that X is a binomial variate for n trials with parameter p. When p has a beta prior distribution with parameters a and b, then given $X = x$, the posterior distribution of:

- p is beta with parameters $x + a$ and $n - x + b$, or equivalently the same as the distribution of
$$\frac{\left(\dfrac{x+a}{n-x+b}\right)F}{1+\left(\dfrac{x+a}{n-x+b}\right)F},$$
where F is an F random variable with $df_1 = 2(x + a)$ and $df_2 = 2(n - x + b)$;
- $[(n-x+b)/(x+a)][p/(1-p)]$ is the F distribution with $df_1 = 2(x + a)$ and $df_2 = 2(n - x + b)$.

With independent beta priors, these posterior distributions induce posterior distributions for the difference of proportions, odds ratio, and relative risk. Hashemi, Nandrum, and Goldberg (1997) and Nurminen and Mutanen (1987) gave integral expressions for these posterior distributions. Equivalent expressions using finite sums were provided by Latorre (1982) for the odds ratio, Hora and Kelley (1983) for the odds ratio and relative risk, Weisberg (1972), Aitchison and Bacon-Shone (1981), and Gupta et al. (1997) for the relative risk (the latter authors also used uniform priors over restricted ranges), and Pham-Gia and Turkkan (1993) for the difference of proportions (see also Walters, 1986 for uniform priors). It is relatively simple to evaluate numerically these posterior distributions.

For the logit-normal prior with mean 0, the prior density function for p is

$$f(p) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(\log\frac{p}{1-p}\right)^2\right\}\frac{1}{p(1-p)},\qquad 0 < p < 1.$$

The posterior density function for p, given the outcome x, is

$$f(p\,|\,x) = \frac{\exp\left\{-\dfrac{1}{2\sigma^2}\left(\log\dfrac{p}{1-p}\right)^2\right\}p^{x-1}(1-p)^{n-x-1}}{\displaystyle\int_0^1 \exp\left\{-\dfrac{1}{2\sigma^2}\left(\log\dfrac{p}{1-p}\right)^2\right\}p^{x-1}(1-p)^{n-x-1}\,dp}.$$

Because the posterior distribution of p does not have closed-form in this case, the posterior distributions of the summary measures of interest are more complex to evaluate than with beta priors. However, they are not so difficult as to require MCMC methods.

A Bayesian CI is a region (sometimes called a *credible set*) that has posterior probability equal to the desired confidence level. One approach is to construct HPD CIs. Such intervals have posterior probability equal to the desired confidence level and have higher posterior density for every value inside the interval than for every value outside of it. This approach was used by Hashemi et al. (1997). An alternative confidence interval uses the "tail method," by which a $100(1-\alpha)\%$ interval for a parameter consists of values between the $\alpha/2$ and $(1-\alpha/2)$ quantiles.

In our opinion, a fatal disadvantage of the HPD interval for the odds ratio and relative risk is its lack of invariance under parameter transformation. For instance, if $(L,\ U)$ is a 95% HPD interval using the posterior distribution of the odds ratio, then the 95% HPD interval using the posterior distribution of the inverse of the odds ratio (which is relevant if we reverse the labeling of the two groups being compared) is not $(1/U,\ 1/L)$. In fact, it can be considerably different. This is not surprising when one realizes that the 95% region of highest density for a random variable X is not the inverse of the 95% region of highest density for $1/X$.

For example, consider the case of uniform prior densities when $n_1 = n_2 = 10$. When $x_1 = 1$ and $x_2 = 5$, the sample odds ratio = 1/9 and the Bayes' 95% HPD confidence interval for the true odds ratio is (0.0006, 0.82); when $x_1 = 5$ and $x_2 = 1$, the odds ratio = 9 and the Bayes' 95% HPD CI is (0.17, 38.23), which is very different from (1/0.82, 1/0.0006). By contrast, the 95% tail CIs for the odds ratios when $n_1 = n_2 = 10$ with uniform priors are (0.017, 1.10) when $x_1 = 1$ and $x_2 = 5$ and (0.91, 57.93) when $x_1 = 5$ and $x_2 = 1$. In another example, for tables with sample odds ratio equal to 0, the HPD interval with diffuse priors is typically of the form $(0,\ U)$, but when rows are interchanged so that the sample odds ratio $= \infty$, the HPD interval has a finite upper bound.

HPD invariance to group labeling does occur on the log scale for the odds ratio and relative risk, for which the relevant parameter is a difference (i.e., log odds ratio = difference of log odds and log relative risk = difference of log probabilities). For example, when $x_1 = 1$ and $x_2 = 5$ with $n_1 = n_2 = 10$, the Bayes' 95% HPD CI for the log odds ratio is (−3.94, 0.19), and when $x_1 = 5$ and $x_2 = 1$, it is (−0.19, 3.94). The corresponding intervals for the odds ratio are (0.019, 1.21) and (0.82, 51.43), but of course, although these match appropriately, they are not HPD intervals on that scale.

As a referee suggested to us, the HPD method is one that first requires a firm commitment to a "preferred" scale of measurement. For the odds ratio, these invariance considerations suggest the log scale. However, users interpret the magnitude of the odds ratio on its original scale rather than the log scale, and the lack of invariance when constructing HPD intervals on the original scale is to us a compelling reason not to use the HPD approach for this measure.

Likewise, although the HPD approach is invariant for the difference of proportions and the log relative risk, the HPD interval formed directly for the relative risk is not invariant, as may not be HPD intervals formed for measures derived from the difference of proportions such as the number needed to treat (NNT). Thus, we used the tail method for these measures as well. The tail-method approach is invariant on any scale. A disadvantage is that it is longer than the HPD interval.

In the evaluations reported below, the only case for which we used the HPD interval was for the difference of proportions when the sample measure takes its boundary values of +1 and −1 (i.e., when $x_1 = n_1$ and $x_2 = 0$ or when $x_1 = 0$ and $x_2 = n_2$). We observed that the posterior density is monotone in such cases with the Jeffreys prior or more diffuse priors, and close to monotone for priors that are more informative than the Jeffreys prior. With the Jeffreys prior the CI then has the form $(L,\ 1)$ or $(-1,\ U)$. With a monotone posterior, excluding both upper and lower tails of the posterior distribution with the tail method seems inappropriate. We did conduct numerical evaluations of the HPD interval for the log odds ratio, but results are not shown in the tables discussed below. The coverage probabilities tended to be slightly worse, on the average, than the tail intervals (in terms of the criteria reported in those tables). When considered on the odds ratio scale, they were shorter than the tail-method CIs in only a slight majority of the cases considered.

With beta prior distributions, the computations for the tail interval are not difficult. Let $F_\omega(t)$ denote the cumulative distribution function for the posterior distribution of a generic measure of association ω. Then,

$$F_\omega(t) = \iint_{S_t} f(p_1 \,|\, x_1) f(p_2 \,|\, x_2)\ dp_1\, dp_2,$$

where $S_t = \{(p_1,\ p_2) : \omega \le t,\ 0 < p_1,\ p_2 < 1\}$. The tail CI $(L,\ U)$ satisfies $F_\omega(L) = \alpha/2$ and $F_\omega(U) = 1 - \alpha/2$. For numerical evaluations we used the free software `R`, combining the `R` function "integrate" to perform the numerical integration which yields F_ω (using the integral expressions in Hashemi et al., 1997) and the `R` function "optim" to perform the Nelder–Mead algorithm. That algorithm minimizes the function

$$G(L,U) = |\, F_\omega(U) - (1-\alpha/2)\,| + |\, F_\omega(L) - \alpha/2\,|$$

with respect to L and U. We used simple Monte Carlo to obtain the starting values for L and U for this iterative method, by randomly generating 100,000 values from the posterior distribution and choosing the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ percentiles.

Computations with the logit-normal prior are somewhat more complex, because the posterior distribution of p itself involves an integral. We used the `R` function "integrate" with an `R` function modified based on one subroutine of the `FORTRAN` code `QUADPACK` (Piessens et al., 1983) to perform the numerical integration to calculate $F_\omega(t)$. We then used the `R` function "optimize" to minimize $|F_\omega(U) - (1-\alpha/2)|$ and $|F_\omega(L) - \alpha/2|$ separately to determine U and L.

4. Evaluations of Coverage Probabilities

With both the beta and logit-normal priors, we restricted our attention to cases in which the prior distributions of p_1 and p_2 are the same. We considered various cases, including beta priors with $a = b = 0.5$, 1, 1.5, 2.0 and logit-normal priors with $\sigma = 1.0$, 1.5, 2.0, 3.0. These give priors for each probability that are symmetric about 0.5, either U-shaped or uniform (or roughly so in the logit-normal case) or bell-shaped. Such prior specification would not be appropriate when one has strong prior belief that the p_i are both near 0 or near 1, or when there is prior belief that one parameter is larger than the other. However, as mentioned above, our goal was to find a "default" CI that performs well in a wide variety of cases, so we felt that the prior should be symmetric in the identification of "success" and "failure."

In the summary of results below, we mainly emphasize the beta priors, as similar results occurred with logit-normal priors.

4.1 *Difference of Proportions*

For the difference of proportions, $p_1 - p_2$, we now discuss the performance of the Bayesian tail CI (suitably modified at the boundary), based on independent beta priors. Denote an interval based on priors with $a_i = b_i = a$ by $I_a(n_1, x_1; n_2, x_2)$, or I_a for short. Our discussion refers mainly to the 0.95 confidence level, but the evaluations also studied the 0.99 level. Let $C_a(n_1, p_1; n_2, p_2)$, or C_a for short, denote the true coverage probability of a nominal 95% CI I_a. We investigated whether there is a value of a for which $|C_a(n_1, p_1; n_2, p_2) - 0.95|$ tends to be small for most (p_1, p_2), even with small n_1 and n_2, with C_a rarely very far (say 0.02) below 0.95.

We explored the performance of I_a for various combinations of p_1 and p_2 and for various fixed (n_1, n_2) combinations. Table 1 summarizes some characteristics in small-sample cases $(n_1, n_2) = (10, 10), (20, 20), (30, 30), (30, 10)$, in an average sense based on taking (p_1, p_2) uniform from the unit square. It tabulates the average coverage probability, the average distance between the actual coverage probability and the nominal level of 0.95, and the proportion of the parameter space for which the coverage probability is less than 0.93. Average interval lengths are not reported, as they were similar for the various intervals considered, although naturally longer for the more diffuse priors. By virtue of the uniform averaging, the interval I_1 (i.e., uniform prior) matches the nominal coverage probability. However, the Jeffreys prior also does well in this sense, and it tends to be better than the uniform prior in terms of distance of actual coverage probability from the nominal level and the incidence of low coverage probabilities. The more informative priors have much poorer performance in terms of these other criteria.

Table 1

Summary of performance of tail 95% confidence intervals for $p_1 - p_2$ using beta(a, a) priors, averaging with respect to a uniform distribution for (p_1, p_2)

		Prior parameter a				
Characteristic	n_i	0.5	1	1.5	2	Score
Coverage	10	0.946	0.950	0.939	0.919	0.954
	20	0.947	0.950	0.944	0.932	0.949
	30	0.948	0.950	0.946	0.937	0.949
	30, 10	0.948	0.950	0.939	0.923	0.955
Distance	10	0.012	0.017	0.035	0.060	0.012
	20	0.006	0.010	0.021	0.037	0.007
	30	0.004	0.007	0.015	0.027	0.004
	30, 10	0.008	0.012	0.028	0.047	0.008
Cov. prob. < 0.93	10	0.089	0.120	0.233	0.308	0.016
	20	0.020	0.065	0.169	0.259	0.008
	30	0.007	0.037	0.131	0.223	0.001
	30, 10	0.021	0.088	0.221	0.312	0.005

Table 2

Summary of performance of tail 95% confidence intervals for $p_1 - p_2$ using beta(a, a) priors, averaging with respect to (p_1, p_2) uniform over the region $|p_1 - p_2| < 0.1$

		Prior parameter a				
Characteristic	n_i	0.5	1	1.5	2	Score
Coverage	10	0.950	0.969	0.978	0.984	0.958
	20	0.948	0.961	0.969	0.974	0.954
	30	0.948	0.958	0.964	0.969	0.952
	30, 10	0.949	0.962	0.967	0.971	0.954
Distance	10	0.015	0.019	0.028	0.034	0.016
	20	0.009	0.012	0.019	0.024	0.008
	30	0.005	0.008	0.014	0.019	0.005
	30, 10	0.010	0.012	0.018	0.021	0.007
Cov. prob. < 0.93	10	0.093	0	0	0	0.037
	20	0.011	0	0	0	0
	30	0.004	0	0	0	0
	30, 10	0.007	0.003	0.003	0.002	0.010

Similar results occurred under other averaging with respect to beta priors with $a = b$. The intervals $I_{0.5}$ and I_1 have similar performance, better than the more informative priors in terms of having average coverage probability near the nominal level and a small proportion of low coverage probabilities. Averaging performance over the unit square with respect to relatively diffuse priors can mask poor behavior in certain regions, and in practice certain pairings (e.g., $|p_1 - p_2|$ small) are often more common or more important than others. Table 2 shows results for a somewhat different averaging, taking (p_1, p_2) uniform over the region $|p_1 - p_2| < 0.1$. This may be more realistic in giving more weight to similar parameter values. In this case the more informative priors tend to be overly conservative. Similar results (not shown here) were obtained using averaging with respect to beta priors that tended to give similar p_i values, such as using beta(3, 12) priors for each p_i, for which the prior mean is 0.2 and standard deviation is 0.1.

Besides studying these summary expectations, we plotted C_a as a function of p_1 for various fixed values of p_2 and of $p_1 - p_2$. Figure 1 illustrates, plotting $C_{0.5}$ and C_1 as a function of p_1 when $p_2 = 0.1$ and 0.5, for the case $n_1 = n_2 = 20$. Performance of these intervals is adequate except for I_1 when the parameters are far apart. Not surprisingly, coverage probabilities tended to be very low for I_2 when the true parameter values were far apart. For small samples, using such informative priors can significantly reduce the chance that the interval contains certain parameter values (see Carlin and Louis, 2000, p. 98–103, for a discussion of this in the single binomial parameter case).

Tables 1 and 2 also summarize performance of a good frequentist method for interval estimation of the difference of proportions, namely the interval based on inverting the score test (Mee, 1984). The 95% confidence interval consists of the set of Δ values for which $|z| \leq 1.96$ with

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \Delta}{\sqrt{\tilde{\pi}_1(1 - \tilde{\pi}_1)/n_1 + \tilde{\pi}_2(1 - \tilde{\pi}_2)/n_2}},$$

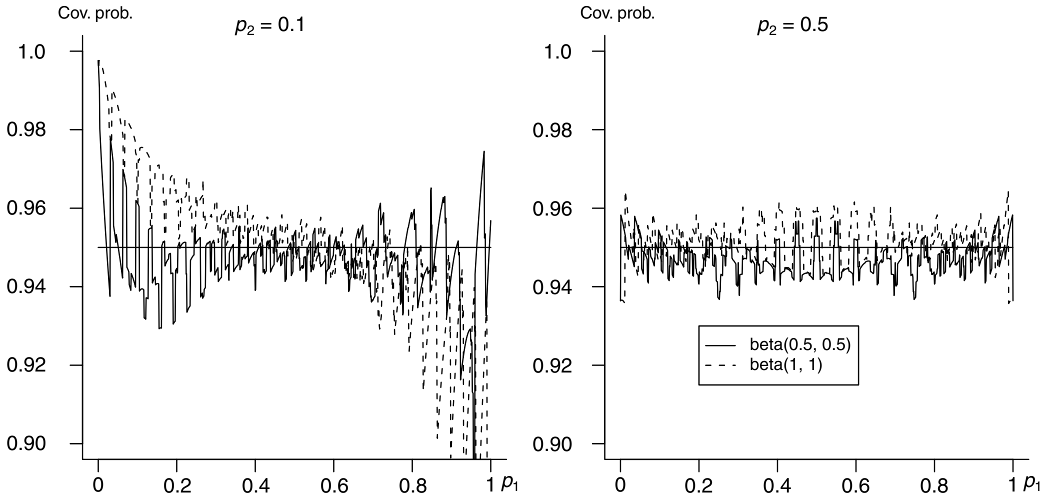

Figure 1. Coverage probabilities for nominal 95% confidence intervals for $p_1 - p_2$ plotted as function of p_1 when $p_2 = 0.1$ and 0.5, using independent beta(0.5, 0.5) or beta(1, 1) priors for each p_i, with $n_1 = n_2 = 20$.

Table 3

Summary of performance of tail 95% *confidence intervals for odds ratio when* $n_1 = n_2 = 20$ *using independent beta*$(a,\ a)$ *priors for* p_1 *and* p_2*, averaging with respect to beta priors with* $a = b = 0.5,\ 1,\ 2,$ *and* $a = 3,\ b = 12$

	Averaging	Prior parameter a				
Characteristic	a, b	0.5	1	1.5	2	Score
Coverage	0.5	0.949	0.878	0.792	0.717	0.959
	1	0.951	0.951	0.931	0.899	0.953
	2	0.946	0.955	0.955	0.950	0.950
	3, 12	0.947	0.961	0.965	0.965	0.954
Distance	0.5	0.028	0.086	0.170	0.245	0.014
	1	0.011	0.012	0.033	0.066	0.008
	2	0.007	0.006	0.010	0.017	0.005
	3, 12	0.009	0.012	0.018	0.023	0.006
Cov. prob. < 0.93	0.5	0.032	0.234	0.443	0.565	0.041
	1	0.014	0.048	0.185	0.318	0.012
	2	0.009	0.002	0.031	0.110	0.001
	3, 12	0.019	0.001	0.018	0.051	0.001

where $\tilde{\pi}_1$ and $\tilde{\pi}_2$ denote the maximum likelihood estimates of π_1 and π_2 under the constraint $\pi_1 - \pi_2 = \Delta$. (A slightly more conservative version, discussed by Miettinen and Nurminen, 1985 and Nurminen, 1986, has variance that differs by a factor of $(n_1 + n_2)/(n_1 + n_2 - 1)$.) Based on our evaluations here and in another project, this method performs well in a broad variety of conditions, even with small samples. Here, we found that it performs at least as well as the Bayesian intervals (and usually better) in terms of the prevalence of undercoverage probabilities when the sample sizes are small. It does, however, tend to be wider than the Bayesian intervals.

Similar results were obtained with logit-normal priors as with the beta priors. The intervals based on diffuse priors similar in shape to Jeffreys prior ($\sigma = 2$ and 3) performed considerably better than more informative priors ($\sigma = 1$ or 1.5). Using higher σ was more advantageous as p_1 and p_2 were potentially more different. Generally, considering a wide variety of cases, the Bayesian intervals based on logit-normal priors did not tend to perform as well as the score interval in terms of the proportion of cases in which the actual coverage probability was unacceptably low. Based on all the evaluations we conducted about the difference of proportions, our preference is to use the Jeffreys prior as a default prior.

4.2 *Odds Ratio and Relative Risk*

For the odds ratio using beta priors, the interval performed much better using the Jeffreys prior than the other beta priors. Table 3 shows some illustrative results, focusing on the case $n_1 = n_2 = 20$ and averaging with respect to various independent beta priors. In this case, the interval I_1 can have poor performance when the parameters may be far apart (as illustrated by results for the averaging with $a = 0.5$). The $I_{0.5}$ interval was the only one that had overall coverage proportion close to the nominal level for each case.

Similar results occurred for the relative risk. Table 4 shows an analogous table for it. Coverage probabilities tended to be closer to the nominal level and tended to have fewer cases of unacceptably low coverages for $I_{0.5}$ than I_1. For the odds ratio and the relative risk, as with the difference of proportions, the more informative priors tend to be quite conservative when the actual parameters are close, as shown by results averaged with respect to the beta prior with $a = 3$ and $b = 12$.

Tables 3 and 4 also show results of forming the frequentist CI based on inverting the score test. See Cornfield (1956) and Miettinen and Nurminen (1985) for the score interval for the odds ratio and Koopman (1984), Miettinen and Nurminen (1985), and Nurminen (1986) for the score interval for the relative risk. This method tends to perform uniformly well in a wide variety of cases. It tends to be better than the Bayesian intervals in terms of closeness of the actual coverage probability to the nominal level. It also tends to be better

Table 4

Summary of performance of tail 95% *confidence intervals for relative risk when* $n_1 = n_2 = 20$ *using independent beta*$(a,\ a)$ *priors for* p_1 *and* p_2*, averaging with respect to beta priors with* $a = b = 0.5,\ 1,\ 2$ *and* $a = 3,\ b = 12$

	Averaging	Prior parameter a				
Characteristic	$a,\ b$	0.5	1	1.5	2	Score
Coverage	0.5	0.950	0.912	0.863	0.818	0.954
	1	0.950	0.951	0.939	0.919	0.951
	2	0.947	0.953	0.953	0.950	0.950
	3, 12	0.946	0.960	0.969	0.975	0.949
Distance	0.5	0.020	0.053	0.103	0.151	0.012
	1	0.009	0.009	0.025	0.048	0.006
	2	0.006	0.004	0.009	0.016	0.004
	3, 12	0.006	0.010	0.019	0.025	0.005
Cov. prob. < 0.93	0.5	0.041	0.139	0.314	0.422	0.049
	1	0.019	0.037	0.147	0.263	0.014
	2	0.010	0.002	0.037	0.109	0.001
	3, 12	0.019	0	0	0	0.005

than the informative Bayesian intervals in terms of how often the actual coverage probability is unacceptably low.

With logit-normal priors, again the intervals based on diffuse priors similar in shape to Jeffreys prior ($\sigma = 2$ and 3) performed considerably better than more informative priors ($\sigma = 1$ or 1.5). The case that encountered the least behavior of very low coverage probabilities was $\sigma = 2$. Its performance was similar to that of the score interval, but the score interval tended to have actual coverage probability closer to the nominal level.

5. Using Priors with Correlated Probabilities

The above evaluations used independent priors for the two probabilities. In practice, it may sometimes be sensible to treat these parameters as dependent, a priori. For instance, with little prior information one might be content subjectively to treat each p_i as uniform. However, if one were told that $p_1 = 0.05$, then conditionally in many applications this would induce the subjective belief that p_2 is also close to 0.

We also considered CIs constructed using dependent priors, focusing mainly on bivariate normal priors for the logits. Results were consistent with independent priors, in the sense that more diffuse priors provided more protection over a broader range of parameter values. As one would expect, the greater the positive correlation in the prior, the poorer the coverage probabilities tended to be when the parameters were actually quite different. If one uses a bivariate logit-normal prior with a moderate correlation, we recommend taking a large value of σ (around 3) for a default value if one wants good coverage protection over a relatively broad range. To illustrate, Figure 2 shows coverage probabilities for the 95% tail interval for $p_1 - p_2$ when $n_1 = n_2 = 20$, plotted as a function of p_1 when $p_2 = 0.1$ and 0.5, using a bivariate logit-normal prior distribution with correlation 0.5 and $\sigma = 2$ or 3. These correlated priors gave sufficient smoothing for the sample sizes considered that the posterior density was not monotone, so we used the tail interval for all cases.

When p_1 and p_2 are truly close, one would expect benefits to using positively correlated priors with relatively small σ. Table 5 investigates this. It considers the performance of a bivariate logit-normal prior with correlation 0.5 and $\sigma = 1$, 2, 3, for constructing 95% CIs for $p_1 - p_2$, when results are averaged with respect to $(p_1,\ p_2)$ being uniform over the region $|p_1 - p_2| < 0.1$. For $\sigma = 1$, the coverage probabilities tend to be too high. However, the interval has the benefit of shorter average length compared to using larger σ or the score interval. These are also reported in Table 5.

Howard (1998) suggested an alternative correlated prior. He amended the independent beta priors and used prior density function proportional to

$$e^{-(1/2)u^2} p_1^{a-1}(1-p_1)^{b-1} p_2^{c-1}(1-p_2)^{d-1},$$

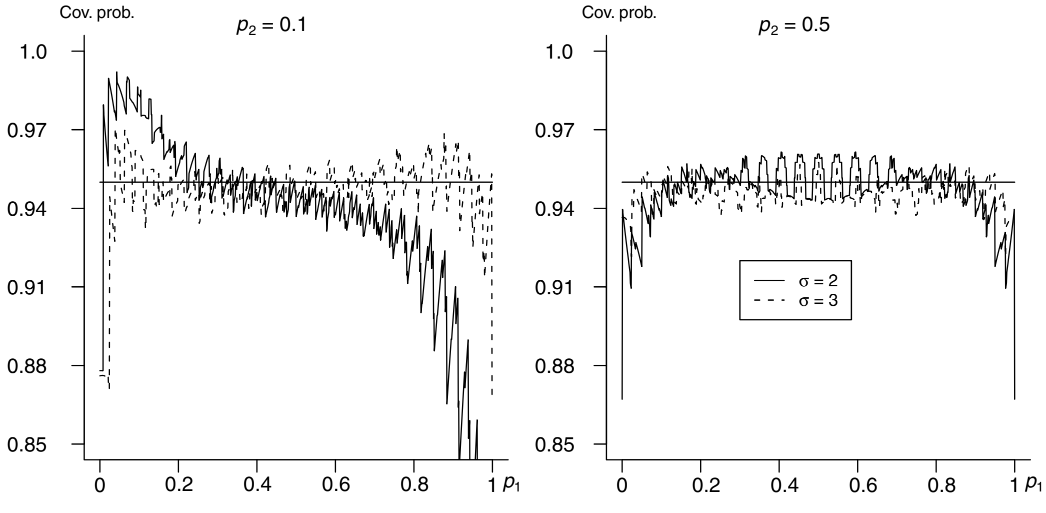

Figure 2. Coverage probabilities for nominal 95% confidence intervals for $p_1 - p_2$ plotted as function of p_1 when $p_2 = 0.1$ and 0.5, using bivariate normal prior with correlation 0.5 and $\sigma = 2$ or 3, when $n_1 = n_2 = 20$.

Table 5

Summary of performance of tail 95% confidence intervals for $p_1 - p_2$ using bivariate normal prior for logits with correlation 0.5 and $\sigma = 1, 2, 3$, averaging with respect to (p_1, p_2) uniform over the region $|p_1 - p_2| < 0.1$

		Standard deviation σ			
Characteristic	n	1	2	3	Score
Coverage	10	0.994	0.972	0.953	0.958
	20	0.985	0.961	0.947	0.954
	30, 10	0.986	0.965	0.950	0.954
Distance	10	0.044	0.022	0.015	0.016
	20	0.035	0.012	0.010	0.008
	30, 10	0.036	0.015	0.012	0.007
Length	10	0.535	0.613	0.630	0.707
	20	0.423	0.462	0.470	0.506
	30, 10	0.460	0.513	0.523	0.565
Cov. prob. < 0.93	10	0	0	0.046	0.063
	20	0	0.006	0.026	0
	30, 10	0.003	0.007	0.011	0.009

where

$$u = \frac{1}{\sigma} \log \left(\frac{p_1(1 - p_2)}{p_2(1 - p_1)} \right).$$

Howard suggested using $\sigma = 1$ for a standard form of this prior. However, this is a relatively strong dependence. For the amendment of the Jeffreys priors ($a = b = c = d = 0.5$) the correlation is 0.84 when $\sigma = 1$, 0.59 when $\sigma = 2$, and 0.41 when $\sigma = 3$. By contrast, a correlation of 0.5 in a bivariate logit-normal prior corresponds to a correlation between the binomial parameters of 0.49 when $\sigma = 1$, 0.47 when $\sigma = 2$, and 0.45 when $\sigma = 3$.

As with the bivariate logit-normal prior, coverages with Howard's prior can be highly dependent on the choice of σ for the parameter values that are relatively far apart, with small σ appropriate only when the parameter values are relatively close. We considered $\sigma = 1$, 2, and 3 in evaluations of coverage probabilities. Overall, for the adapted Jeffreys prior ($a = b = c = d = 0.5$), $\sigma = 3$ performed best in terms of protecting against overly low coverage probabilities. The case $\sigma = 3$ gave similar results as the bivariate logit-normal prior with correlation 0.5 and $\sigma = 3$. The logit-normal approach did slightly better when the proportions were highly divergent (e.g., 0.1 and 0.9).

For the prior to be less informative, Howard suggested taking $a = b = c = d$ equal to some small ϵ close to 0 (positive values ensure a proper posterior). However, the correlation between the proportions is then very strong. Our evaluations with this prior showed more cases of relatively low coverage probabilities.

Overall, neither correlated prior gave coverage performance uniformly as good as that provided by the score CI. An alternative way of inducing dependence is to use a hierarchical prior, but we did not consider that. Also, we have not considered the possibility of using matching priors (e.g., Rousseau, 2000). We decided to put main emphasis on the method (beta priors) that receives primary attention in the current texts and research literature on Bayesian inference. In addition, the model with beta priors is most transparent to applied statisticians.

6. Discussion

Why would a frequentist consider using a Bayesian CI? One reason might be to have some of the benefits that a Bayesian approach can have for small sample sizes, such as shrinkage relative to simple Wald intervals that use the maximum likelihood point estimate as the center of the interval. Ordinary frequentist methods that perform relatively well, such as the score interval and adjusted Wald intervals that add pseudo observations to the sample before forming ordinary Wald intervals (e.g., Agresti and Caffo, 2000), use such shrinkage.

The results of this article suggest that if one uses a Bayesian approach but worries about frequentist performance, specifically maintaining good coverage performance over the entire parameter space, it is best to use quite diffuse priors. Even uniform priors are often too informative. Our recommendation, in agreement with Brown et al. (2001) in the single binomial case, is to use independent Jeffreys priors for the binomial parameters. On the other hand, if a Bayesian is unconcerned about coverage probabilities deteriorating as parameter values move farther away from the region in which the prior density places most of its mass, then there are benefits in expected length to using more informative priors, as Table 5 illustrated. If one prefers the frequentist paradigm, it seems adequate to use the ordinary score CI. It performed well for all cases considered in this article—including the three parameters, different sample sizes, and different confidence levels.

The conclusion that it is "safest" to use a diffuse prior will not surprise most readers. What did surprise us was how much the coverage probabilities could vary according to the choice of the prior, even for moderate sample sizes such as $n_1 = n_2 = 30$. An implication is that careful selection of prior distributions is crucial in the much more complex, often hierarchical, models to which the Bayesian methods are being routinely applied these days.

For a concrete example of how results for actual data can depend strongly on the prior, we consider an interesting example from a clinical trial discussed by Begg (1990). For an urn-sampling method to allocate patients to treatments, the 11 patients allocated to the experimental treatment were all successes and the only patient allocated to the control treatment was a failure. That is, the table has rows (11, 0) and (0, 1). The 95% tail CI was (1.16, 218.4) for the beta(2, 2) prior, (1.71, 4677.2) for the uniform prior, and ($3.28, 1.36 \times 10^6$) for the Jeffreys prior. By contrast, the score interval is ($4.49, \infty$).

It would be of interest to extend the present investigation to consider analogous CIs for stratified 2 × 2 tables such as occur with meta analyses (e.g., Warn, Thompson, and Spiegelhalter, 2002). While a relatively informative prior may be fine for representing the subjective beliefs of a researcher, it may result in poor performance in terms of ordinary frequentist criteria (for instance, if the researcher's prior beliefs are incorrect and the true proportions are actually far apart).

Many Bayesians may consider such criteria irrelevant, but it is worthy of attention to those who traditionally take a frequentist approach but find the Bayesian approach appealing for certain types of modeling. In addition, regardless of one's philosophical approach, for standard models for categorical data such as logistic regression and log-linear models we believe it is inappropriate to form HPD CIs on the odds ratio scale. The lack of invariance is severe rather than a minor inconvenience.

At the start of the article, we mentioned that Bayesian interval estimation does not seem to be used much in practice with these parameters. If one does want to use the Bayesian tail intervals with independent beta priors, it is quite simple. From the simple expression of the posterior distribution of the binomial probability or the odds in terms of an F or beta random variable, one can quickly simulate the posterior distributions of the three measures considered in this article by simulating values from the F or beta distribution. Thus, it is simple to simulate reasonable approximations for the endpoints of CIs for the "tail method," for instance forming the 95% CI by the values between the simulated 2.5 percentile and 97.5 percentile of the appropriate posterior distribution.

Finding more precise intervals or HPD intervals requires better approximations. We have constructed functions using the free software R for the tail intervals for the three parameters discussed in this article, using independent beta priors. These R functions are available at **http://www.stat.ufl.edu/∼aa/cda/R/bayes/index.html**.

Of course, one could also use general-purpose Bayesian software such as BUGS. For simulation, however, it is adequate to use the direct approach mentioned above, and for these simple cases numerical integration very quickly gives excellent accuracy.

Acknowledgements

This research was partially supported by grants from the National Institute of Health and the National Science Foundation.

References

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *American Statistician* **54,** 280–288.

Aitchison, J. and Bacon-Shone, J. (1981). Bayesian relative risk analysis. *American Statistician* **35,** 254–257.

Altham, P. M. E. (1969). Exact Bayesian analysis of a 2×2 contingency table, and Fisher's "exact" significance test. *Journal of the Royal Statistical Society, Series B* **31,** 261–269.

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

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* New York: Wiley.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Massachusetts: Addison-Wesley.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* **16,** 101–117.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edition. London: Chapman and Hall.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman (ed), Volume 4, 135–148.

Good, I. J. (1956). On the estimation of small frequencies in contingency tables. *Journal of the Royal Statistical Society, Series B* **18,** 113–124.

Gupta, R. C., Albanese, R. A., Penn, J. W., and White, T. J. (1997). Bayesian estimation of relative risk in biomedical research. *Environmetrics* **8,** 133–143.

Hashemi, L., Nandram, B., and Goldberg, R. (1997). Bayesian analysis for a single 2 × 2 table. *Statistics in Medicine* **16,** 1311–1328.

Hora, S. C. and Kelley, G. D. (1983). Bayesian inference on the odds and risk ratios. *Communications in Statistics—Theory and Methods* **12,** 725–738.

Howard, J. V. (1998). The 2 × 2 table: A discussion from a Bayesian viewpoint. *Statistical Science* **13,** 351–367.

Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics* **40,** 513–517.

Latorre, G. (1982). The exact posterior distribution of the cross-ratio of a 2 × 2 contingency table. *Journal of Statistical Computation and Simulation* **16,** 19–24.

Leonard, T. (1972). Bayesian methods for binomial data. *Biometrika* **59,** 581–589.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B* **37,** 23–37.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *Annals of Mathematical Statistics* **35,** 1622–1643.

Mee, R. W. (1984). Confidence bounds for the difference between two probabilities (letter). *Biometrics* **40,** 1175–1176.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4,** 213–226.

Nurminen, M. (1986). Confidence intervals for the ratio and difference of two binomial proportions. *Biometrics* **42,** 675–676.

Nurminen, M. and Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics* **14,** 67–77.

Pham-Gia, T. and Turkkan, N. (1993). Bayesian analysis of the difference of two proportions. *Communications in Statistics—Theory and Methods* **22,** 1755–1771.

Piessens, R., deDoncker-Kapenga, E., Uberhuber, C., and Kahaner, D. K. (1983). *Quadpack: A Subroutine Package for Automatic Integration.* Berlin: Springer-Verlag.

Rousseau, J. (2000). Coverage properties of one-sided intervals in the discrete case and application to matching priors. *Annals of the Institute of Statistical Mathematics* **52,** 28–42.

Walters, D. E. (1986). On the reliability of Bayesian confidence limits for a difference of two proportions. *Biometrical Journal* **28,** 337–346.

Warn, D. E., Thompson, S. G., and Spiegelhalter, D. J. (2002). Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Statistics in Medicine* **21,** 1601–1623.

Weisberg, H. I. (1972). Bayesian comparison of two ordered multinomial populations. *Biometrics* **28,** 859–867.

Wong, G. Y. and Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association* **80,** 513–524.

Received December 2003. *Revised June* 2004.
Accepted August 2004.