BAYESIAN INFERENCE OF RISK RATIO OF TWO PROPORTIONS USING A DOUBLE SAMPLING SCHEME

Dewi Rahardja¹ and Dean M. Young²

¹Department of Clinical Sciences and Simmons Cancer Center, UT Southwestern Medical Center, Dallas, Texas, USA
²Department of Statistical Sciences, Baylor University, Waco, Texas, USA

We consider Bayesian point and interval estimation for a risk ratio of two proportion parameters using two independent samples of binary data subject to misclassification. In order to obtain model identifiability, we apply a double sampling scheme. For the identifiable model, we propose a Bayesian method for statistical inference for a two proportion risk ratio. Specifically, we derive an easy-to-implement closed-form sampling algorithm to draw from the posterior distribution of interest. We demonstrate the efficiency of our algorithm for Bayesian inference via Monte Carlo simulation studies and using a real data example.

Key Words: Bayesian credible interval; Binary data; Double sampling; Misclassification; Risk ratio.

1. INTRODUCTION

In many disciplines, such as medical research, one obtains binary data that are sometimes misclassified. For example, a healthy patient may be wrongly diagnosed as having a certain disease or vice versa. The consequence of ignoring misclassification in statistical inference from binary data was first reported by Bross (1954). He showed that classical estimators based on only data subject to misclassification can be extremely biased. Therefore, one requires additional data or information to yield model identifiability and to correct the bias. One can utilize at least two methods to achieve this goal in the Bayesian paradigm. The first method, pioneered by Tenenbein (1970), is to collect training data using a double sampling scheme; the other is to use sufficiently informative priors specified by expert opinion or previous data. The rationale of Tenenbein’s double sampling scheme is as follows. Fallible classification procedures result in misclassifications but are inexpensive, while infallible classification procedures result in errorless classification but are much more expensive. Therefore, the use of both fallible and infallible procedures yields not only model identifiability, but also economical viability. When an infallible procedure is
unavailable or prohibitively expensive, one can use sufficiently informative priors to create an identifiable model in the Bayesian paradigm.

We next review the statistical literature on statistical inference using binary data with possible misclassification. The objective of these research articles was to perform statistical inferences on the proportion parameters associated with the infallible classification method, which is also the purpose of our article.

For the single proportion problem, when data is obtained using the double sampling scheme, Tenenbein (1970) proposed a maximum likelihood estimator and derived its asymptotic variance for the proportion parameter of interest. Also, Boese et al. (2006) derived several likelihood-based confidence intervals for a proportion parameter using data subject to only false positive misclassification. For the single proportion problem using misclassified data with no training data, Gaba and Winkler (1992) and Viana et al. (1993) developed Bayesian approaches with highly informative priors. Bayesian inferences with informative priors were also developed for two-sample problems for two proportion parameters. For example, see Evans et al. (1996) for risk difference (the difference of two proportion parameters) and Gustafson et al. (2001) for odds ratios.

Although misclassified binary data are common in epidemiology, we remark that such data are also frequently seen in device testing. For example, Zhong (2002) studied the specificity and sensitivity of a fallible diagnostic test together with a gold standard. In addition, Stamey et al. (2007) developed Bayesian estimation of an intervention effect with pre- and post-misclassified binomial data. Also one can obtain clinical trial binary data containing misclassification. For example, Lyles et al. (2004) provided design and analytic considerations for single-armed studies with misclassification of a repeated binary outcome.

The risk ratio, also known as relative risk, is defined as the ratio of two proportion parameters from two groups. Risk ratios are commonly used in clinical trials and epidemiological studies to measure the relative frequency of a certain event, such as an adverse event between two groups. To date, we have found no methods for inference on risk ratios for proportions using two-sample binary data subject to misclassifications. In this article, we propose Bayesian approaches to this problem. In section 2 we describe the data, and in section 3 we develop Bayesian models and posterior sampling algorithms. In section 4 we illustrate our algorithms using real data. We then examine the performance of our Bayesian inference approaches in section 5 and conclude with a brief discussion in section 6.

2. THE DATA

In this section we consider two-sample binary data subject to misclassification. The data are obtained using a fallible classification device that can yield both false-positive and false-negative classifications. For example, suppose a study’s objective is to assess whether a certain disease has the same prevalence rates between men and women. A positive result of an assay indicates that a subject in the study has the disease. This assay does not have perfect sensitivity and specificity, and therefore both false-positive and false-negative classifications can occur.
Table 1 Fallible data for sample \( i, \ i = 1, 2 \)

<table>
<thead>
<tr>
<th>Classification</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>( M_i - X_i )</td>
<td>( X_i )</td>
<td>( M_i )</td>
</tr>
</tbody>
</table>

To describe the data, let \( F_{ij} \) be the observed classification by the fallible method for the \( j \)th individual in the \( i \)th sample, where \( i = 1, 2, j = 1, \ldots, M_i \), and

\[
F_{ij} = \begin{cases} 
1 & \text{if the result is positive by the fallible method} \\
0 & \text{if the result is negative by the fallible method}
\end{cases}
\]

If one denotes \( X_i \) as the number of individuals with positive classifications, the observed fallible data for sample \( i, i = 1, 2 \), are displayed in Table 1.

Similarly, we define the unobserved true classification of the \( j \)th individual in the \( i \)th sample as \( T_{ij} \),

\[
T_{ij} = \begin{cases} 
1 & \text{if the result is truly positive} \\
0 & \text{otherwise}
\end{cases}
\]

Clearly, misclassification occurs when \( T_{ij} \neq F_{ij} \).

Next, we introduce the following notations for the \( i \)th sample:

\[
\begin{align*}
\pi_i & \equiv \Pr(T_{ij} = 1) \\
\pi_i & \equiv \Pr(F_{ij} = 1) \\
\phi_i & \equiv \Pr(F_{ij} = 1 | T_{ij} = 0) \\
\theta_i & \equiv \Pr(F_{ij} = 0 | T_{ij} = 1)
\end{align*}
\]

We see that \( p_i \) is the actual proportion parameter of interest, \( \pi_i \) is the proportion parameter of the fallible classification method, and \( \phi_i \) and \( \theta_i \) are the false-positive rate and the false-negative rate for the fallible classification method, respectively. Here, we allow the false-positive and false-negative rates to differ between the two samples, i.e., \( \phi_1 \neq \phi_2 \) and \( \theta_1 \neq \theta_2 \). Note that \( \pi_i \) are functions of the remaining parameters \( p_i, \theta_i, \) and \( \phi_i, i = 1, 2 \). In particular, we have

\[
\pi_i = \Pr(T_{ij} = 1) \Pr(F_{ij} = 1 | T_{ij} = 1) + \Pr(T_{ij} = 0) \Pr(F_{ij} = 1 | T_{ij} = 0)
= p_i(1 - \theta_i) + q_i \phi_i
\]

where \( q_i = 1 - p_i \).

As stated in section 1, we are interested in statistical inference on the risk ratio

\[
r \equiv p_1/p_2
\]

Because \( \pi_i \) is determined through \( p_i, \phi_i \), and \( \theta_i, i = 1, 2 \), six effective parameters exist in the model. However, the sufficient-statistic dimension is only two because \( X_i \) and
X_{2}, given in Table 1, are the sufficient statistics associated with Eq. (1). Because the dimension of the sufficient statistics is less than the parameter dimension, the model (1) is unidentifiable. Therefore, additional data or information about the parameters is needed to obtain an identifiable model. For the Bayesian paradigm, two primary methods can provide additional information: The first is to obtain training data by using a double sampling scheme, and the second is to incorporate prior knowledge about model parameters through sufficiently informative priors. In this paper we focus on data obtained using a double sampling scheme.

To create model identifiability for one-sample binary data subject to misclassification, Tenenbein (1970) used additional training data obtained by double sampling. Specifically, in addition to the original fallible data classified only by the fallible method, he also used a smaller training data sample obtained by classifying each individual in these training data by both the fallible and the infallible classification methods. The additional data enable the assessment of the false-positive and false-negative rates of the fallible method. Other applications of double sampling schemes can be found in Tenenbein (1972), Hochberg (1977), and Boese et al. (2006).

We apply the same double sampling scheme to our two proportion risk ratio problem to obtain \( n_i \) training data in addition to the original \( M_i \) fallible data for the \( i \)th sample, \( i = 1, 2 \). The combined data are presented in Table 2. In this table we use \( n_{ijk} \) to denote the number of individuals classified as \( j \) and \( k \) by the infallible and fallible labeling methods, respectively, where \( j = 0, 1 \) and \( k = 0, 1 \). For example, \( n_{01} \) is the number of individuals classified as negative by the infallible method but positive by the fallible method in the \( i \)th sample. One can see that the dimension of sufficient statistics \( (X_1, X_2, n_{000}, n_{101}, n_{110}, n_{200}, n_{201}, n_{210})' \) for the combined data

\[
\begin{array}{cccc}
\text{Data} & \text{Infallible method} & 0 & 1 & \text{Total} \\
\hline
\text{Training} & 0 & n_{00} & n_{01} & n_0 \\
& 1 & n_{10} & n_{11} & n_1 \\
\text{Total} & & n_0 & n_1 & n_i \\
\text{Original} & NA & M_i - X_i & X_i & M_i \\
\end{array}
\]

Note. NA, not available.

\[
\begin{array}{cccc}
\text{Data} & \text{Infallible method} & 0 & 1 & \text{Total} \\
\hline
\text{Training} & 0 & q_i (1 - \phi_i) & q_i \phi_i & q_i \\
& 1 & p_i \theta_i & p_i (1 - \theta_i) & p_i \\
\text{Original} & NA & 1 - \pi_i & \pi_i & 1 \\
\end{array}
\]

Note. NA, not available.
is now greater than the number of parameters in Eq. (1). Therefore, the model is now identifiable. For easy reference, we present the cell probabilities for Table 2 in Table 3.

3. THE MODEL

We develop Bayesian inference techniques for the data described in the previous section. In particular, we derive algorithms to draw from the posterior distributions of each of the parameters. Once a posterior sample is drawn for $p_1$ and $p_2$, a posterior sample for risk ratio $r$ is readily obtained from Eq. (2). We perform statistical inference for $r$, including determining a credible interval for $r$, based on the sample median from the posterior of $r$. In what follows, we give explicit algorithms for drawing from the posterior distributions of interest.

For sample $i$, $i = 1, 2$, in Table 2, the observed counts $(n_{i00}, n_{i01}, n_{i10}, n_{i11})'$ of the training data have a quadrinomial distribution with total size $n_i$ and cell probabilities displayed in an upper right $2 \times 2$ submatrix in Table 3, i.e.,

$$(n_{i00}, n_{i01}, n_{i10}, n_{i11})' | p_i, \phi_i, \theta_i \sim \text{Quad}[n_i, (q_i(1 - \phi_i), q_i \phi_i, p_i \theta_i, p_i(1 - \theta_i))]$$

In addition,

$$X_i | M_i, p_i, \phi_i, \theta_i \sim \text{Bin}(M_i, \pi_i), i = 1, 2$$

Because $(n_{i00}, n_{i01}, n_{i10}, n_{i11})'$ and $X_i$ are independent for sample $i$ and samples 1 and 2 are independent, the sampling distribution of the data is

$$f(d | \eta) \propto \prod_{i=1}^{2} \left[ \left[ q_i(1 - \phi_i) \right]^{n_{i00}} (q_i \phi_i)^{n_{i01}} (p_i \theta_i)^{n_{i10}} [p_i(1 - \theta_i)]^{n_{i11}} \pi_i^{X_i} (1 - \pi_i)^{M_i - X_i} \right]$$

where

$$d \equiv (n_{100}, n_{101}, n_{110}, n_{111}, X_1, n_{200}, n_{201}, n_{210}, n_{211}, X_2)'$$

is the data vector and

$$\eta \equiv (p_1, \phi_1, \theta_1, p_2, \phi_2, \theta_2)'$$

is the parameter vector. For our Bayesian framework, we chose a uniform prior for each component of $\eta$ and assumed these priors are independent; therefore, the joint prior distribution is the noninformative prior,

$$p(\eta) = 1$$

Combining Eqs. (3) and (5), we obtain the joint posterior distribution

$$f(\eta | d) \propto \prod_{i=1}^{2} \left[ \left[ q_i(1 - \phi_i) \right]^{n_{i00}} (q_i \phi_i)^{n_{i01}} (p_i \theta_i)^{n_{i10}} [p_i(1 - \theta_i)]^{n_{i11}} \prod_i X_i (1 - \pi_i)^{M_i - X_i} \right]$$

which has the same functional form as Eq. (3).
To sample from the posterior density in Eq. (6), we perform a transformation of the parameters $\eta$ and obtain a closed-form posterior sampling algorithm. Specifically, we define

$$\hat{\lambda}_{i1} = p_i(1 - \theta_i)/\pi_i$$

and

$$\hat{\lambda}_{i2} = p_i\theta_i/(1 - \pi_i)$$

Using Eqs. (7) and (8), the posterior density in Eq. (6) becomes

$$f(\eta^* | d) \propto \prod_{i=1}^{2} \lambda_{i1}^{n_{i1}} (1 - \lambda_{i1})^{n_{i01}} \lambda_{i2}^{n_{i20}} (1 - \lambda_{i2})^{n_{i02}} \pi_i^{X_i + n_{i1}} (1 - \pi_i)^{M_i - X_i + n_{i0}}$$

where $d$ is defined in Eq. (4) and

$$\eta^* = (\hat{\lambda}_{i1}, \hat{\lambda}_{i2}, \pi_1, \hat{\lambda}_{21}, \hat{\lambda}_{22}, \pi_2)'$$

is the transformed parameter vector. Because the transformed parameters of Eq. (10) are now separable, we can straightforwardly draw $\hat{\lambda}_{i1}$, $\hat{\lambda}_{i2}$, and $\pi_i$ from the posterior Eq. (9) by using

$$\hat{\lambda}_{i1} \sim \text{Beta}(n_{i1} + 1, n_{i01} + 1)$$

$$\hat{\lambda}_{i2} \sim \text{Beta}(n_{i20} + 1, n_{i00} + 1)$$

$$\pi_i \sim \text{Beta}(X_i + n_{i1} + 1, M_i - X_i + n_{i0} + 1)$$

with $i = 1, 2$. Once $\hat{\lambda}_{i1}$, $\hat{\lambda}_{i2}$, and $\pi_i$ are available, we can obtain $p_i$, $\phi_i$, and $\theta_i$ by solving Eqs. (1), (7), and (8) so that

$$p_i = \pi_i \hat{\lambda}_{i1} + (1 - \pi_i)\hat{\lambda}_{i2}$$

$$\phi_i = (1 - \hat{\lambda}_{i2})\pi_i/q_i$$

and

$$\theta_i = \hat{\lambda}_{i2}(1 - \pi_i)/p_i$$

In summary, the following is the closed-form algorithm to sample from the posterior density of Eq. (6). First, choose a large number $J$, say 10,000, for the posterior draw sample size. For $i = 1, 2$:

1. Obtain $J$ samples of $\hat{\lambda}_{i1}$, $\hat{\lambda}_{i2}$, and $\pi_i$ using Eqs. (11)–(13).
2. Obtain $J$ samples of $p_i$, $\phi_i$, and $\theta_i$ using Eqs. (14)–(16).
3. Obtain $J$ samples of the risk ratio $r$ using Eq. (2).

We then use the median of the $J$ observations from $p(r | d)$ as a point estimator of $r$ because the distribution of the posterior sample of $r$ is skewed. Finally, we obtain
4. AN EXAMPLE

In this section we apply our Bayesian inference algorithm to real data that were first described in Hildesheim et al. (1991) and later were used in Boese (2003). This study explored the association between exposure to herpes simplex virus (HSV) and having invasive cervical cancer (ICC). The data are displayed in Table 4. In this example we define HSV = 1 for women exposed to HSV and HSV = 0 otherwise. In addition, we define ICC = 1 for women in the first group or the case group (having ICC) and ICC = 2 for women in the second group or the control group (not having ICC). This data indicate a total of 2044 women, with 732 women in the case group and 1312 women in the control group. The Western blot procedure is a fallible detector of HSV. A subsample of the women was also tested with the refined Western blot procedure, which is a relatively accurate procedure and thus is treated as an infallible classification. We regard this subsample as the training data in the double sampling scheme. Both false-positive and false-negative misclassification errors of HSV using the Western blot procedure occurred in this study. Here, \( p_1 = \Pr(\text{exposed to HSV} \mid \text{having ICC}) \) is the probability that a patient has been exposed to the HSV given she has ICC (case group), \( p_2 = \Pr(\text{exposed to HSV} \mid \text{not having ICC}) \) is the probability that a patient has been exposed to the HSV given she does not have ICC (control group), and \( \lambda \) is given in Eq. (2). Using the sampling algorithm developed in the previous section with a posterior sample size \( J = 10,000 \), we determine that \( \hat{\lambda}_{\text{Bayes}} = 1.34 \), and a 90% Bayesian credible interval for \( \lambda \) is

\[
[1.01, 1.75]
\]

We interpret Eq. (17) to mean that the probability that the random variable \( \lambda \) is in interval (17) is approximately 90%. Because the lower limit of interval (17) exceeds 1.0, we believe that strong statistical evidence indicates that a larger proportion of women have been exposed to HSV in the case group than in the control group. Thus, conceivably an association could exist between exposure to HSV and actually having ICC.

<table>
<thead>
<tr>
<th>Group</th>
<th>Data</th>
<th>Fallible method</th>
<th>Infallible method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Case</td>
<td>Training</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>NA</td>
<td>318</td>
</tr>
<tr>
<td>Control</td>
<td>Training</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>NA</td>
<td>701</td>
</tr>
</tbody>
</table>

Note: NA, not available.
5. MONTE CARLO SIMULATIONS

In this section we conduct Monte Carlo simulation studies to examine the performance of our sampling algorithms and to evaluate the accuracy of the credible intervals for $r$ for varying sample sizes, false-positive rates, and false-negative rates. For the sake of simplifying the presentation of simulation results, we assumed the sample size and parameter configurations of the form $N = N_1 = N_2$, $n = n_1 = n_2$, $\phi = \phi_1 = \phi_2$, and $\theta = \theta_1 = \theta_2$. We remark that these assumptions are not required by our posterior sampling algorithms.

We considered 32 simulation scenarios resulting from combinations of the following configurations:

1. True proportion parameters of interest $(p_1, p_2)$: (.1, .2), (.4, .6).
2. False-positive and false-negative rates $(\phi, \theta)$: (.1, .1), (.2, .2).
3. Ratio of training-sample size versus the total sample size $(n/N)$: 0.2, 0.4.
4. Total sample sizes $(N)$: 100, 200, 300, 400.

For each simulation scenario, we simulated $K = 10,000$ data sets. For each data set, we drew $J = 10,000$ samples of $r$ from the posterior density using the algorithm described in section 3. We then computed the posterior sample median as a point estimator of $r$ and a 90% credible interval for $r$. Next, we generated boxplots of the $K$ posterior medians of $r$ to examine their behavior as a point estimator of $r$. In addition, we calculated the coverage probability and the average length of the $K$ credible intervals.

In Figures 1 and 2 we present the boxplots of $K$ posterior sample medians of $r$ versus the total sample size $N$. The actual proportion parameters are $(p_1, p_2) = (1.1, 2), (4, 6)$ for Figures 1 and 2, respectively. In each figure, the top two panels

![Boxplots of posterior medians versus total sample sizes N where (p1, p2) = (1, 2). The top two panels have (phi, theta) = (2, 2) and the bottom two panels have (phi, theta) = (1, 1); the left two panels have n/N = .2 and the right two panels have n/N = .4.](image)
BAYESIAN INFERENCE OF RISK RATIO

Figure 2  Boxplots of posterior medians versus total sample sizes \( N \) where \((p_1, p_2) = (0.4, 0.6)\). The top two panels have \((\phi, \theta) = (0.2, 0.2)\) and the bottom two panels have \((\phi, \theta) = (0.1, 0.1)\); the left two panels have \(n/N = 0.2\) and the right two panels have \(n/N = 0.4\).

have misclassification probabilities \((\phi, \theta) = (0.2, 0.2)\), and the bottom two panels have the misclassification probabilities \((\phi, \theta) = (0.1, 0.1)\). In addition, for the left two panels we use \(n/N = 0.2\), and for the right two panels we use \(n/N = 0.4\).

For the 32 simulation configurations for both figures, we have the following observations:

1. For each panel of 4 boxplots, the variation of the posterior medians decreases as \( N \) increases.
2. For each figure, the variation of the posterior medians of the top two panels with larger misclassification probabilities \( \phi \) and \( \theta \) is greater than that of the bottom two panels with smaller misclassification probabilities.
3. For each figure, the variation of the posterior medians shown in the left two panels for smaller \( n/N \) is greater than the variation of the posterior medians depicted in the right two panels for larger \( n/N \).
4. For the same values of \( \phi \), \( \theta \), \( n/N \), and \( N \), the boxplots in Figure 1 have greater variation than the corresponding boxplots in Figure 2.

In Table 5 we present the coverage probabilities, average lengths, and standard deviations of the \( K \) credible intervals for \( r \) under each simulation scenario. The coverage probabilities are all close to the nominal 90% level, thus indicating that our credible interval estimator performs well. Also, we have the following observations concerning Table 5:

1. For fixed \( p_1, p_2, \phi, \theta \), and \( n/N \), the average length and standard deviation of the credible intervals decrease as \( N \) increases.
2. For fixed \( p_1, p_2, n/N \), and \( N \), the average length and standard deviation of the credible intervals decrease as \( \phi \) and \( \theta \) decrease.
For fixed $p_1, p_2, \phi, \theta$, and $N$, the average length and standard deviation of the credible intervals decrease as $n/N$ increases.

4. For the same $\phi, \theta$, $n/N$, and $N$, the average length and standard deviation of the credible intervals decrease when $(p_1, p_2)$ approaches $(.5, .5)$.

**6. DISCUSSION**

In this paper we proposed Bayesian credible interval estimation for risk ratios of two proportions using binary data subject to both false-positive and false-negative misclassification. Simulations showed that our algorithms produced credible intervals with the nominal coverage probabilities. The posterior median performed well as a point estimator of $r$.

Several advantages of our closed-form algorithms are apparent for sampling from the full posterior distributions:

- Because we draw directly from the posterior distributions, we need not specify initial values and no burn-in period or convergence issues occur.
- Because posterior draws are available for each parameter, inferences on the risk difference, the odds ratio, and other functions of $p_1$ and $p_2$ are also straightforward.
As shown in Eqs. (11)-(13), the posterior sampling algorithms can handle zero counts.
No asymptotic theory is involved.
Our posterior sampling algorithms can be generalized to data from more than two groups.

Applications exist where no training data is available. In this case, if sufficiently informative priors are obtainable, then more sophisticated Monte Carlo Markov chain (MCMC) methods may need to be used to sample the posterior of interest, instead of the more straightforward Monte Carlo method that we propose in this paper.

For future research, we plan to develop frequentist methods for constructing confidence intervals for the risk ratio \( r \) using double sampling.

ACKNOWLEDGMENTS

We thank the associate editor and anonymous referees for their suggestions that have significantly improved the presentation of this article. We also thank Julie Beyrer for her careful editing of this article.

REFERENCES


