Credible sets for risk ratios in over-reported two-sample binomial data using the double-sampling scheme

Dewi Rahardja *, Dean M. Young

Department of Statistical Sciences, Baylor University, Waco, TX 76798-7140, USA

ABSTRACT

We consider point and interval estimation for risk ratios based on two independent samples of binomial data subject to false positive misclassification. For such data it is well known that the model is unidentifiable. We consider incorporating training data obtained by using a double-sampling scheme to make the model identifiable. In this identifiable model, we propose a Bayesian method to make statistical inferences. In particular, we derive an easy-to-implement closed-form algorithm for drawing from the posterior distributions. The algorithm is illustrated using a real data example and further examined via Monte Carlo simulation studies.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Due to reasons such as imprecise diagnostic procedures or human errors, misclassifications can occur in binomial data. Usually misclassifications of both types (false positive or false negative errors) are possible. For example, with an imperfect blood test, a healthy patient may be incorrectly diagnosed to be sick and vice versa. In some cases, only one type of misclassification is present. For instance, Moors et al. (2000) presented auditing data where only false negative (under-reported) errors occurred. Perry et al. (2000) showed blood testing data which had only false positive (over-reported) errors.

Among others, Bross (1954) reported that classical estimators ignoring misclassifications can be extremely biased when applied to misclassified binomial data. Therefore, one needs to incorporate additional information or data in order to make the model identifiable and to correct the bias. Several methods are popular in the literature for this purpose. One is to include training data by using a double-sampling scheme suggested by Tenenbein (1970); the other is to use informative priors specified by expert opinions or previous data in the Bayesian paradigm. The rationale of Tenenbein’s double-sampling scheme is very sensible. Fallible classification procedures result in misclassifications but are inexpensive, while infallible (true) classification procedures result in true classifications but are much more expensive. Therefore, the use of both fallible and infallible procedures not only enables the identifiability of the model, but also is economically viable. In some cases, an infallible procedure is unavailable or prohibitively expensive; informative priors can then be used to make the model identifiable in the Bayesian paradigm. Another method is to use multiple fallible classifiers to make valid inferences. In what follows we review the literature of research on binomial data with misclassifications. The objectives of these studies are to draw statistical inferences on the proportion parameters associated with the infallible classification method, which is also the objective of this article.

For one-sample problems, several researchers have considered the case where only false negative errors were present. Lie et al. (1994) used a maximum likelihood approach to the problem where false negative errors are corrected using multiple fallible classifiers. York et al. (1995) considered this same problem from the Bayesian perspective. For when data are obtained using a double-sampling scheme, Moors et al. (2000) discussed maximum likelihood estimation and one-sided
interval estimation and applied their methods to auditing data. Boese et al. (2006) derived several likelihood-based CIs for the proportion parameter.

In addition, several authors also discussed one-sample problems with misclassification errors of both types. For when double sampling is used, Tenenbein (1970) proposed a maximum likelihood estimator along with its asymptotic variance for the proportion parameter. Although inexplicitly presented in his article, a Wald-type confidence interval (CI) for the proportion parameter can be readily constructed using Tenenbein’s formulas. For when training data are unavailable in one-sample problems, Gaba and Winkler (1992) and Viana et al. (1993) developed Bayesian approaches with informative priors.

For two-sample problems with misclassification errors of both types, Bayesian inferences with informative priors were also developed for when training data were unavailable. For example, see Evans et al. (1996) for risk difference (the difference of two proportion parameters) and Gustafson et al. (2001) for odds ratios.

To date, no methods for inference on risk ratio (the ratio of two proportion parameters) have been developed for two-sample binomial data subject to misclassifications. In this article, we limit our scope to data subject only to one error type. Without loss of generality, we consider data with only false positive errors. We propose Bayesian approaches for constructing point and interval estimations for risk ratios. In Section 2 we describe the data. In Section 3 we develop Bayesian models and algorithms for data with or without training data. In Section 4 we illustrate our algorithms using real data. The performances of our approaches are examined in Section 5 and a discussion can be found in Section 6.

### 2. Data

In this section we consider two-sample binomial data subject to misclassifications. The data are obtained using a fallible device or method which can yield false positive but not false negative classification. For example, a study is conducted to assess whether a certain infection has the same prevalence rates for men and women, and a positive result of a blood test is used to determine whether a subject in the study has the infection. This blood test is not perfect and only false positive classification can occur.

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

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

Denoting as \( X_i \) and \( Y_i \) the numbers of individuals with positive and negative classifications, the observed data obtained by the fallible method for Sample \( i \) 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 notation for the \( i \)th sample:

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

We see that \( p_i \) is the true proportion parameter of interest, \( \pi_i \) is the proportion parameter of the fallible method, and \( \phi_i \) is the false positive rate of the fallible method. Here we allow the false positive rates to be different for the two samples, i.e., \( \phi_1 \neq \phi_2 \). Note that \( \pi_1 \) and \( \pi_2 \) are functions of other parameters. In particular, by the law of total probability, we have

\[
\begin{align*}
\pi_i &= \Pr(T_i = 1) \Pr(F_i = 1 | T_i = 1) + \Pr(T_i = 0) \Pr(F_i = 1 | T_i = 0) \\
&= p_i + q_i \phi_i.
\end{align*}
\]

where \( q_i = 1 - p_i \).

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

\[
r = \frac{p_1}{p_2}.
\]

Because \( \pi_i \) is determined through \( p_i \) and \( \phi_i \), effectively there are four parameters in the model: \( p_1, \phi_1, p_2, \phi_2 \). However, the dimension of the sufficient statistics is only 2. For example, the vector \((X_1, X_2)\) is a vector of sufficient statistics for this

<table>
<thead>
<tr>
<th>Classification</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Y_i )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_i )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_i )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
model. Because the dimension of the sufficient statistics is less than the number of parameters, the model is unidentifiable. Therefore, additional data or information are needed to make the model identifiable. In this paper we consider the double-sampling scheme with training data to achieve this goal.

Tenenbein (1970) used additional training data obtained by double sampling in an effort to make one-sample binomial data subject to misclassification become identifiable. Specifically, in additional to the original fallible data classified only by the fallible method, a new but smaller training data set is obtained by classifying each individual in the training data by both the fallible method and the infallible method. Doing so enables the assessment of the misclassification rates of the fallible method. Other applications of double-sampling schemes include Tenenbein (1972), Hochberg (1977) and Boese et al. (2006).

We apply the same double-sampling scheme to our two-sample problem and obtain size-\( n_i \) training data in addition to the original size-\( M_i \) fallible data for the \( i \)th sample. 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 methods, respectively. For example, \( n_{001} \) is the number of individuals classified as negative by the infallible method but positive by the fallible method in the \( i \)th sample. Note that \( n_{100} \) is not available because false negative errors are assumed to be impossible. One can see that the dimension of sufficient statistics for the combined data is 6, which is greater than the number of parameters. For example, the vector \((X_1, X_2, n_{100}, n_{101}, n_{200}, n_{201})\) is a set of sufficient statistics. Therefore, the model is now identifiable. For easy reference, the cell probabilities of Table 2 are presented in Table 3.

### Table 2

Data for Sample \( i \).

<table>
<thead>
<tr>
<th>Data</th>
<th>Infallible method</th>
<th>Fallible method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Training</td>
<td>( n_{00} )</td>
<td>( n_{01} )</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>( n_{11} )</td>
</tr>
<tr>
<td>Total</td>
<td>( n_{00} )</td>
<td>( n_{*1} )</td>
</tr>
<tr>
<td>Original</td>
<td>NA</td>
<td>( Y_i )</td>
</tr>
</tbody>
</table>

NA: Not available.

### Table 3

Cell probabilities for Sample \( i \).

<table>
<thead>
<tr>
<th>Data</th>
<th>Infallible method</th>
<th>Fallible method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Training</td>
<td>( q_i(1 - \phi_i) )</td>
<td>( q_i\phi_i )</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>( p_i )</td>
</tr>
<tr>
<td>Original</td>
<td>NA</td>
<td>( 1 - \pi_i )</td>
</tr>
<tr>
<td></td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
</tbody>
</table>

NA: Not available.

3. The model

In this section we develop Bayesian inferences for the data described in Section 2. Our aim is to derive explicit algorithms for drawing from the posterior distributions of all the parameters given the data. After posterior samples are drawn for both \( p_1 \) and \( p_2 \), a posterior sample for risk ratio \( r \) is readily obtained by dividing the sample of \( p_1 \) by the sample of \( p_2 \) elementwise. Statistical inferences including point and interval estimation on \( r \) can be obtained on the basis of this sample drawing.

For Sample \( i \), in Table 2, the observed counts \( (n_{00}, n_{01}, n_{11}) \) of the training data have a trinomial distribution with total size \( n_i \) and the probabilities displayed in an upper right 2 \( \times \) 2 submatrix in Table 3, i.e.,

\[
(n_{00}, n_{01}, n_{11}) | p_i, \phi_i \sim \text{Trin}(n_i, (q_i(1 - \phi_i), q_i\phi_i, p_i)).
\]

In addition, the observed counts \( (X_i, Y_i) \) have the following binomial distribution:

\[
(X_i, Y_i) | p_i, \phi_i \sim \text{Bin}(M_i, (\pi_i, 1 - \pi_i)).
\]

Because \( (n_{00}, n_{01}, n_{11}) \) and \( (X_i, Y_i) \) are independent for Sample \( i \) and Sample 1 is independent of Sample 2, the sampling distribution of the vector of all data

\[
d = (n_{00}, n_{01}, n_{11}, X_1, Y_1, n_{200}, n_{201}, n_{211}, X_2, Y_2)
\]

given the vector of all parameters

\[
\eta = (p_1, \phi_1, p_2, \phi_2)
\]

is

\[
f(d | \eta) \propto \prod_{i=1}^{2} \left\{ [q_i(1 - \phi_i)]^{n_{00}} (q_i\phi_i)^{n_{01}} P_i^{n_{11}} \pi_i^{X_i} (1 - \pi_i)^{Y_i} \right\}. \tag{3}
\]
To put this into a Bayesian framework, we choose a non-informative proper prior for $\eta$. Specifically, we choose a uniform prior for each component of $\eta$ and assume that these priors are independent; i.e., the joint prior distribution is
\[ p(\eta) = 1. \tag{4} \]
Combining Eqs. (3) and (4), we obtain the following joint posterior distribution:
\[ f(\eta|d) \propto \prod_{i=1}^{2} \left[ \frac{q_i(1 - \phi_i)^{n_{i00}} (q_i \phi_i)^{n_{i01}} \pi_i^{n_{i11}} (1 - \pi_i)^{Y_i}}{q_i} \right], \tag{5} \]
which has the same functional form as the sampling distribution in Eq. (3).
In order to draw from the posterior density in Eq. (5), we perform a transformation of parameters $\eta$ and derive a closed-form algorithm. First note that
\[ 1 - \pi_i = q_i(1 - \phi_i). \tag{6} \]
Then, we define
\[ \lambda_i = p_i/\pi_i. \tag{7} \]
With Eqs. (6) and (7), the posterior density in Eq. (5) becomes
\[ f(\eta|d) \propto \prod_{i=1}^{2} \lambda_i^{n_{i11}} (1 - \lambda_i)^{n_{i01}} \pi_i^{n_{i11}} (1 - \pi_i)^{Y_i+n_{i00}}. \tag{8} \]
Because the transformed parameters
\[(\lambda_1, \pi_1, \lambda_2, \pi_2)\]
are now separable, it is straightforward to draw $\lambda_i$ and $\pi_i$ from Eq. (8) by using the following closed-form algorithm:
\[
\begin{align*}
\lambda_i & \sim \text{Beta}(n_{i11} + 1, n_{i01} + 1) \tag{9} \\
\pi_i & \sim \text{Beta}(X_i + n_{i1} + 1, Y_i + n_{i00} + 1). \tag{10}
\end{align*}
\]
Once $\lambda_i$ and $\pi_i$ are available, we can obtain $p_i$ and $\phi_i$ by solving Eqs. (1) and (7):
\[
\begin{align*}
p_i &= \pi_i \lambda_i \tag{11} \\
\phi_i &= (1 - \lambda_i) \pi_i / q_i. \tag{12}
\end{align*}
\]
In summary, the following is the closed-form algorithm for drawing from the posterior density in Eq. (5). First, choose a large number $J$ (say, 10,000) for the posterior draw sample size. For $i = 1, 2$:
1. Obtain size-$J$ samples of $\lambda_i$ and $\pi_i$ using Eqs. (9) and (10).
2. Obtain size-$J$ samples of $p_i$ and $\phi_i$ using Eqs. (11) and (12).
3. Obtain a size-$J$ sample of the risk ratio $r$ using Eq. (2).

Then, we use the median of the sample of $r$ as a point estimator of $r$. We choose the median because the distribution of the posterior sample of $r$ is skewed. Finally, we obtain a $100(1 - \alpha)\%$ CI for $r$ by using the lower and upper ($\alpha/2$)th percentiles of the sample of $r$. A histogram of the sample of $r$ can be plotted to help us to understand the posterior distribution of $r$.

4. An example

In this section we apply our Bayesian inference algorithm to real data first described in Hildesheim et al. (1991) and later used in Boese (2003). The study examined the relationship between exposure to the herpes simplex virus (HSV) and invasive cervical cancer. The western blot procedure is a fallible detector of HSV. A sub-sample of the women is also tested with the refined western blot procedure, which is a relatively accurate procedure and thus is treated as infallible. We regard this sub-sample as the training data in the double-sampling scheme. Both false positive and negative misclassification errors occurred in this study. However, for the sake of illustration, we consider only the false positives and absorb the $n_{i10}$ (false negative) observations into the $n_{i11}$ observations. We display the data in Table 4. Using the algorithm developed in Section 3 with posterior sample size $J = 10,000$, the posterior median for $r$ is 1.45 and a 90% Bayesian credible interval is (1.18, 1.80).
Table 4
Hildesheim example data.

<table>
<thead>
<tr>
<th>Group</th>
<th>Data</th>
<th>Infallible method</th>
<th>Fallible method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>Case</td>
<td>0</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>NA</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>318</td>
<td>375</td>
</tr>
<tr>
<td>Control</td>
<td>Original</td>
<td>NA</td>
<td>701</td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td>33</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>NA</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>32</td>
<td>375</td>
</tr>
</tbody>
</table>

NA: Not available.

Fig. 1. Boxplots of posterior medians versus total sample sizes $N$ where $(p_1, p_2) = (0.1, 0.2)$. The top two panels have $\phi = 0.1$; the left two panels have $n/N = 0.2$ and the right two panels have $n/N = 0.4$.

5. Simulations

We conduct simulation studies in this section to examine the performance of our algorithms and to evaluate the impact on the credible intervals of various true proportion parameter values, sample sizes, and false positive rates. Although not required by our algorithms, for the sake of simplifying the conduct and presentation of simulation results, we let $N = N_1 = N_2, n = n_1 = n_2$, and $\phi = \phi_1 = \phi_2$.

We consider 32 simulation scenarios resulting from combinations of the following values:

1. True proportion parameters of interest $(p_1, p_2)$: $(0.1, 0.2), (0.4, 0.6)$.
2. False positive rates $\phi$: 0.1, 0.2.
3. Ratios 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 simulate $K = 10,000$ data sets. Within each data set, we draw a size-$J = 10,000$ posterior sample of $r$ according to the algorithms in Section 3. Then, the posterior median and a 90% credible interval are computed. Finally, we generate boxplots of the $K$ posterior medians (point estimator) of $r$ to examine the behavior around the true $r$. In addition, we calculate the coverage probability and the average length of the $K$ credible intervals.

In Figs. 1–2 we present the boxplots of $K$ posterior medians of $r$ against total sample size $N$. The true proportion parameters of $(p_1, p_2)$ are $(0.1, 0.2)$ and $(0.4, 0.6)$ for Figs. 1 and 2, respectively. In each figure, the top two panels have $\phi = 0.2$ and the bottom two panels have $\phi = 0.1$. In addition, the left two panels have $n/N = 0.2$ and the right two panels have $n/N = 0.4$.

From the 32 simulation scenarios for both figures, the posterior medians are centered around the true $r$ and therefore is a good point estimator. Moreover, we have the following observations:

1. For each panel of four 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 $\phi$ is greater than that of the bottom two panels with smaller $\phi$.
3. For each figure, the variation of the posterior medians of the left two panels with smaller $n/N$ is greater than that of the right two panels with larger $n/N$.
4. For the same $\phi, n/N, N$, the boxplot in Fig. 1 ($(p_1, p_2)$ away from $(0.5, 0.5)$) has greater variation than the boxplot in Fig. 2 ($(p_1, p_2)$ close to $(0.5, 0.5)$).
In this paper we proposed Bayesian credible interval estimation for risk ratios for binomial data subject to false positive misclassification. Simulations showed that our algorithms produced credible intervals with the nominal coverage probabilities. The posterior median as a point estimator was also demonstrated to be good.

There are many advantages of our closed-form algorithms for drawing from the full posterior distributions:

1. Because we draw directly from the posterior distributions, there is no need to specify initial values and there are no burn-in period or convergence issues.
2. Because posterior draws are available for each parameter, inferences on risk difference, odds ratio, and other functions of $p_1$ and $p_2$ are also straightforward.
3. As shown in Eqs. (9)–(10), the algorithms can handle zero counts.
4. No asymptotic theory is involved.
5. Our algorithms can be generalized to data with more than two samples.
In future research we will generalize our results to binomial data with both error types. In addition, we will explore likelihood-based methods for making statistical inferences on $r$.

Acknowledgements

We thank the editor and referee for their constructive comments which have improved the presentation of this paper.

References


