Journal of Data Science **10**(2012), 51-59

# Bayesian Credible Sets for a Binomial Proportion Based on One-Sample Binary Data Subject to One Type of Misclassification

Dewi Rahardja[1*], Yan D. Zhao[1] and Hongmei Zhang[2]
[1] *University of Texas Southwestern Medical Center and*
[2] *University of South Carolina*

*Abstract*: Interval estimation for the proportion parameter in one-sample misclassified binary data has caught much interest in the literature. Recently, an approximate Bayesian approach has been proposed. This approach is simpler to implement and performs better than existing frequentist approaches. However, because a normal approximation to the marginal posterior density was used in this Bayesian approach, some efficiency may be lost. We develop a closed-form fully Bayesian algorithm which draws a posterior sample of the proportion parameter from the exact marginal posterior distribution. We conducted simulations to show that our fully Bayesian algorithm is easier to implement and has better coverage than the approximate Bayesian approach.

*Key words*: Bayesian credible sets, binary data, double sampling, misclassification, proportion.

## 1. Introduction

In many applications, due to practical reasons such as cost-saving, fallible classifiers which are prone to error are used to classify individuals into two distinct categories. Consequently, misclassified binary data occur. Although both false-positive and false-negative errors are possible for misclassified binary data, sometimes only one type of misclassification error is present. For example, Moors *et al.* (2000) presented auditing data where only false-negative error occurred, and Perry *et al.* (2000) showed blood testing data which had only false-positive error. In this paper, without loss of generality, we assume only false-positive error exists for the data of interest.

Classical estimators of the population proportion parameter for misclassified binary data have been demonstrated to be biased (Bross, 1954). In the literature,

*Corresponding author.

several methods are popular for correcting the bias. One method is the use of multiple fallible classifiers. For example, with multiple fallible classifiers available, Lie *et al.* (1994) have used a maximum likelihood approach and York *et al.* (1995) have developed a Bayesian approach to correct false-negative error. If an infallible classifier is available, Tenenbein (1970) proposed a double-sampling scheme where an additional sample is classified using both the fallible and infallible classifiers. The rationale of Tenenbein's double-sampling scheme is very sensible. Fallible classifiers result in misclassification but are inexpensive, while infallible classifiers result in true classification but are much more expensive. Therefore, the use of both fallible and infallible classifiers not only enables the identifiability of the model, but also is economically viable.

In this paper we will focus on misclassified binary data subject to false-positive error and obtained using a double-sampling scheme. For such data, Moors *et al.* (2000) have derived a one-sided frequentist interval estimator, Raats and Moors (2003) have derived a Bayesian interval estimator, and Boese *et al.* (2006) have derived several likelihood-based confidence intervals (CIs) for the proportion parameter. To overcome the computational difficulty required by the aforementioned methods, Lee and Byun (2008) have used noninformative priors to provide Bayesian credible intervals that were easier to compute. However, because a normal approximation to the marginal posterior density was used in their Bayesian approach, some efficiency may be lost.

In this paper, we derive a closed-form fully Bayesian algorithm which draws a posterior sample of the proportion parameter from the exact marginal posterior distribution. We describe the data in Section 2 and develop Bayesian model and algorithm in Section 3. Then, we illustrate our algorithm using real data in Section 4. We compare the performance of our Bayesian algorithm with the approximate Bayesian algorithm in Section 5 and provide a discussion in Section 6.

## 2. Data

In this section we consider one-sample binary data subject to false-positive misclassification and obtained using a double-sampling scheme. The overall data set is composed of two independent data sets: original data and training data. The original data are obtained by using only the fallible classifier to classify individuals, and the training data are obtained by using both the fallible and infallible classifiers to classify individuals.

The data are displayed in Table 1. There are $M$ and $n$ individuals in the original and training data, respectively. We use $N$ to denote the total number of observations, i.e., $N = M + n$. For the original data, $X$ and $Y$ are the numbers of positive and negative observations by the fallible classifier, respectively. For

the training data, $n_{jk}$ is the number of individuals classified as $j$ and $k$ by the infallible and fallible classifiers, respectively. For example, $n_{01}$ is the number of individuals classified as negative (0) by the infallible classifier but positive (1) by the fallible classifier in the training data. Note that $n_{10}$ is not available because false negative error is assumed to be impossible.

Table 1: One-sample binary data subject to false-positive misclassification and obtained using double sampling

| | | Fallible Classifier | | |
|---|---|---|---|---|
| Data | Infallible Classifier | 0 | 1 | Total |
| Training | 0 | $n_{00}$ | $n_{01}$ | $n_{0\cdot}$ |
| | 1 | NA | $n_{11}$ | $n_{11}$ |
| | Total | $n_{00}$ | $n_{\cdot 1}$ | $n$ |
| Original | NA | $Y$ | $X$ | $M$ |

NA: Not Available

To describe the data and introduce notation, for an individual in the data, let $F$ and $T$ be the classification by the fallible classifier and infallible classifier, respectively. We define $F = 1$ if the individual is positive by the fallible classifier and $F = 0$ otherwise, and $T = 1$ if the individual is positive by the infallible classifier and $T = 0$ otherwise. Clearly, misclassification occurs when $T \neq F$.

Next, we define $p = \Pr(T = 1)$, $\pi = \Pr(F = 1)$, and $\phi = \Pr(F = 1|T = 0)$. We see that $p$ is the true proportion parameter of interest, $\pi$ is the proportion parameter of the fallible classifier, and $\phi$ is the false positive rate of the fallible classifier. Note that $\pi$ is a function of other parameters. In particular, by the law of total probability, we have

$$\begin{aligned} \pi &= \Pr(T = 1)\Pr(F = 1|T = 1) + \Pr(T = 0)\Pr(F = 1|T = 0) \\ &= p + q\phi, \end{aligned} \tag{2.1}$$

where $q = 1 - p$.

We are interested in constructing interval estimators for $p$. For easy reference, the cell probabilities of Table 1 are presented in Table 2.

Table 2: Cell probabilities

| | | Fallible Method | | |
|---|---|---|---|---|
| Data | Infallible Method | 0 | 1 | Total |
| Training | 0 | $q(1-\phi)$ | $q\phi$ | $q$ |
| | 1 | NA | $p$ | $p$ |
| Original | NA | $1-\pi$ | $\pi$ | 1 |

NA: Not Available

## 3. Model

In this section we develop Bayesian inference for the data described in the previous section. Our aim is to derive a closed-form algorithm for sampling from the exact posterior distribution of the proportion parameter $p$ given the data. Statistical inference including point and interval estimation on $p$ will be obtained using this posterior sample.

In Table 1, the observed counts $(n_{00}, n_{01}, n_{11})$ of the training data have a trinomial distribution with total size $n$ and probabilities displayed in an upper right $2 \times 2$ submatrix in Table 2, i.e.,

$$(n_{00}, n_{01}, n_{11})|p, \phi \sim \text{Trin}(n, (q(1-\phi), q\phi, p)).$$

In addition, the observed counts $(X, Y)$ have the following binomial distribution:

$$(X, Y)|p, \phi \sim \text{Bin}(M, (\pi, 1-\pi)).$$

Because $(n_{00}, n_{01}, n_{11})$ and $(X, Y)$ are independent, the sampling distribution of the vector of all data $\boldsymbol{d} = (n_{00}, n_{01}, n_{11}, X, Y)$ given the vector of all parameters $\boldsymbol{\eta} = (p, \phi)$ is

$$f(\boldsymbol{d}|\boldsymbol{\eta}) \propto [q(1-\phi)]^{n_{00}} (q\phi)^{n_{01}} p^{n_{11}} \pi^X (1-\pi)^Y. \tag{3.1}$$

To develop a Bayesian approach, a non-informative proper prior for $\boldsymbol{\eta}$ has been commonly used in the literature. Specifically, we choose a uniform prior for each component of $\boldsymbol{\eta}$ and assume that these priors are independent; i.e., the joint prior distribution is

$$f(\boldsymbol{\eta}) = 1. \tag{3.2}$$

Combining (3.1) and (3.2), we obtain the following joint posterior distribution:

$$f(\boldsymbol{\eta}|\boldsymbol{d}) \propto [q(1-\phi)]^{n_{00}} (q\phi)^{n_{01}} p^{n_{11}} \pi^X (1-\pi)^Y, \tag{3.3}$$

which has the same functional form as the sampling distribution in (3.1).

Sampling from (3.3) is not straightforward. Raats and Moors (2003) derived the posterior to be a nontrivial linear combinations of beta distributions which required heavy computation to sample from, and Lee and Byun (2008) used a normal approximation to the marginal posterior density, which may cause some loss of efficiency.

To improve upon existing algorithms, we propose to use a reparameterization and then derive a closed-form algorithm. Specifically, define $\boldsymbol{\eta}^* = (\lambda, \pi)$, where

$$\begin{aligned} \lambda &= p/\pi, \\ \pi &= \pi. \end{aligned}$$

The reparameterization from $\boldsymbol{\eta}$ to $\boldsymbol{\eta}^*$ retains the number of unique parameters and is also invertible:

$$p = \lambda\pi, \tag{3.4}$$
$$\phi = (1-\lambda)\pi/q. \tag{3.5}$$

We remark that $\boldsymbol{\eta}^*$ are interpretable parameters. In fact, by Table 2, we have

$$\lambda = \frac{\Pr(T=1, F=1)}{\Pr(F=1)} = \Pr(T=1|F=1),$$
$$\pi = \Pr(F=1).$$

Therefore, $\lambda$ is the conditional probability that an individual is truly classified as positive, given that the fallible classification is already positive. Clearly, $\lambda$ and $\pi$ are quantities between 0 and 1.

We now develop a Bayesian model based on the new parameters $\boldsymbol{\eta}^*$. The sampling function of $\boldsymbol{d}$ given $\boldsymbol{\eta}^*$ is

$$f(\boldsymbol{d}|\boldsymbol{\eta}^*) \propto \lambda^{n_{11}}(1-\lambda)^{n_{01}}\pi^{X+n_{\cdot 1}}(1-\pi)^{Y+n_{00}}. \tag{3.6}$$

We specify the joint prior density for $\boldsymbol{\eta}^*$ such that the components of $\boldsymbol{\eta}^*$ have independent beta distributions:

$$f(\boldsymbol{\eta}^*) \propto \lambda^a(1-\lambda)^b\pi^c(1-\pi)^d. \tag{3.7}$$

Combining (3.6) and (3.7), we obtain the following joint posterior density:

$$f(\boldsymbol{\eta}^*|\boldsymbol{d}) \propto \lambda^{n_{11}+a}(1-\lambda)^{n_{01}+b}\pi^{X+n_{\cdot 1}+c}(1-\pi)^{Y+n_{00}+d}, \tag{3.8}$$

where hyper-parameters $a$, $b$, $c$, and $d$ are specified based on prior information. Note that if we set these four hyper-parameters to be zero, then we have a non-informative uniform prior, which will be used in Section 4 and 5 for the example and the simulation study.

Because the new parameters $\lambda$ and $\pi$ are now independent given data, it is straightforward to draw $\lambda$ and $\pi$ from (3.8) by using the following closed-form algorithm:

$$\lambda \sim \text{Beta}(n_{11}+a+1, n_{01}+b+1), \tag{3.9}$$
$$\pi \sim \text{Beta}(X+n_{\cdot 1}+c+1, Y+n_{00}+d+1). \tag{3.10}$$

Once $\lambda$ and $\pi$ are available, we can obtain $p$ and $\phi$ by (3.4) and (3.5).

In summary, the following is the closed-form algorithm for sampling from the posterior density in (3.8). First, choose a large number $J$ (say, 10,000) for the posterior draw sample size. Then,

1. Obtain size-$J$ samples of $\lambda$ and $\pi$ using (3.9) and (3.10).

2. Obtain size-$J$ samples of $p$ and $\phi$ using (3.4) and (3.5).

Then, we use the median of the sample of $p$ as a point estimator for $p$. We choose the median because the distribution of the posterior sample of $p$ may be skewed. Finally, we obtain a $100(1-\alpha)\%$ credible set for $p$ by using the highest posterior density method.

## 4. Example

In this section we apply our closed-form Bayesian algorithm to social security payment data described in Raats and Moors (2003). In Netherlands, six companies are responsible for social security payment and can make mistakes due to the complexity of the rules and regulations. To assess the error rate, an internal auditor of a company checked a random sample of 500 payments and reported that there were 16 errors. The internal auditor might also make mistakes and hence was considered a fallible classifier. Then, a supervising institution (infallible classifier) double-checked a subsample of 53 payments. Analogous to Table 1, the original data sample size is $M = 447$ and the training sample size is $n = 55$. Finally, the classification result can be summarized as $n_{00} = 50$, $n_{01} = 1$, $n_{10} = 0$, $n_{11} = 2$, $X = 14$, and $Y = 433$. Because $n_{10} = 0$, the false-negative rate is likely 0. Therefore, we assume the data follow our model.

Using the algorithm developed in the previous section with posterior sample size $J$=10,000, the posterior median for $p$ is 0.0222 and a 95% Bayesian credible interval is $(0.0047, 0.0396)$. These results are the same as those reported by Lee and Byum (2008) up to the fourth decimal point.

## 5. Simulations

We conduct two simulation studies and report results in this section to examine and compare the performance of our closed-form Bayesian algorithm (Fully) and the approximate Bayesian algorithm (CIBL) by Lee and Byun (2008). We make the comparison using coverage probability (CP) and average length (AL) of the Bayesian credible intervals. We consider 95% confidence limit and fix the false-positive rate $\phi$ at .1 for all the simulations. For each simulation configuration, we generate 10,000 data sets based on which the CP and AL are computed. For our fully Bayesian algorithm, we use posterior samples of size 1,000 for computing point and interval estimators.

In the first simulation study, we fix $p$ at .1 and the proportion of the training data $n/N$ at .1. Then, we choose the total sample size $N$ to range from 30 to 400 with an increment of 10. This simulation setup is similar to what's been done

in Lee and Byun (2008). In Figure 1 we plot the CP and AL versus $N$ for both our fully Bayesian algorithm and the approximate Bayesian algorithm. Figure 1 shows that our fully Bayesian algorithm performs better than the approximate Bayesian algorithm with narrower intervals.

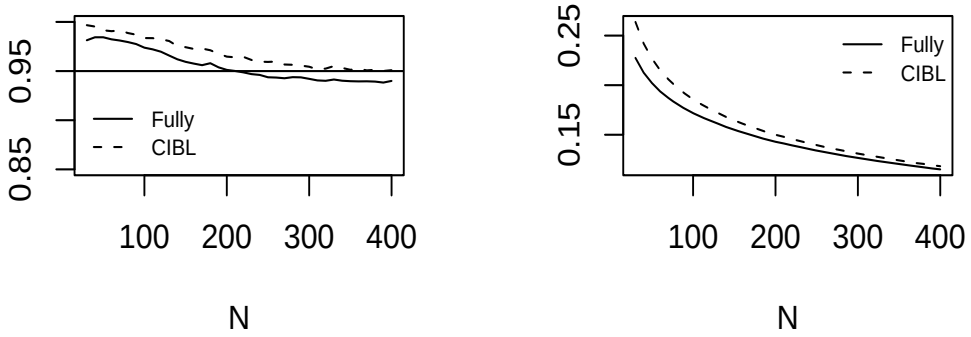


Figure 1: Coverage probabilities (left panel) and average lengths (right panel) of 95% credible intervals versus $N$, where $p = \phi = .1$ and $n = 0.1N$

In the second simulation study, we use $\phi = .1$, $N = 400$, and $n = 40$. Then, we choose the proportion parameter $p$ to range from .01 to .99 with an increment of .01. In Figure 2 we plot the CP and AL versus $p$ for both our fully Bayesian algorithm and the approximate Bayesian algorithm. Similar to Figure 1, Figure 2 shows that our fully Bayesian algorithm performs better that the approximate Bayesian algorithm with narrower intervals. We note that when $p$ is extremely close to 1, the approximate Bayesian intervals have over-coverage and the fully Bayesian intervals have under-coverage. It is hard to tell which one is better in this case; therefore, we recommend that the overall sample size and the proportion for the training data need to be large for valid statistical inference when $p$ is close to 0 or 1.

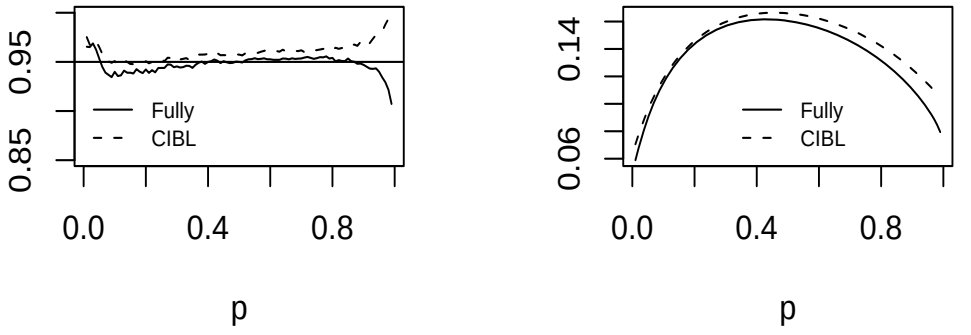


Figure 2: Coverage probabilities (left panel) and average lengths (right panel) of 95% credible intervals versus $p$, where $\phi = .1$, $N = 400$, and $n = 40$

## 6. Discussion

In this paper we derived a closed-form fully Bayesian credible interval for proportion parameter $p$ for one-sample binary data subject to false-positive misclassification. Our algorithm is much easier to implement than existing algorithms. In addition, simulations showed that our algorithm produced credible intervals with the nominal coverage probabilities and have narrower intervals than the existing approximate Bayesian credible intervals, especially when the sample sizes were small. This was not surprising because Lee and Byun's method was based on asymptotic theory while our algorithm sampled from the exact posterior distribution.

Because both $\lambda$ and $\pi$ have marginal beta posteriors, exact Bayesian inference without Monte Carlo simulation can be made on them. However, the parameter $p$ is a product of $\lambda$ and $\pi$ and does not have a recognizable marginal posterior functional form, therefore, Monte Carlo simulation is needed to make Bayesian inference on $p$.

There are several advantages for our closed-form algorithm which draws samples from the exact posterior distribution. First, because we sample directly from the posterior distribution, we do not need to specify initial values and do not have burn-in period or convergence issue. Second, our algorithm can handle zero counts. Lastly, we do not rely on asymptotic theory and therefore the algorithm works well for data with small sample sizes.

We need to be cautious when $p$ is close to either 0 or 1. In these scenarios, the overall sample size and the proportion for the training data need to be large for valid statistical inference, regardless of the statistical methodology used.

## Acknowledgements

The authors thank the referees and the editor for their constructive comments which helped improve the presentation of this paper.

## References

Boese, D. H., Young, D. M. and Stamey, J. D. (2006). Confidence intervals for a binomial parameter based on binary data subject to false-positive misclassification. *Computational Statistics & Data Analysis* **50**, 3369-3385.

Bross, I. (1954). Misclassification in 2×2 tables. *Biometrics* **10**, 478-486.

Lee, S. C. and Byun, J. S. (2008). A bayesian approach to obtain confidence intervals for binomial proportion in a double sampling scheme subject to

false-positive misclassification. *Journal of the Korean Statistical Society* **37**, 393-403.

Lie, R. T., Heuch, I. and Irgens, L. M. (1994). Maximum likelihood estimation of the proportion of congenital malformations using double registration systems. *Biometrics* **50**, 433-444.

Moors, J. J. A., van der Genugten, B. B. and Strijbosch, L. W. G. (2000). Repeated audit controls. *Statistica Neerlandica* **54**, 3-13.

Perry, M., Vakil, N. and Cutler, A. (2000). Admixture with whole blood does not explain false-negative urease tests. *Journal of Clinical Gastroenterology* **30**, 64-65.

Raats, V. M. and Moors, J. J. A. (2003). Double-checking auditors: a Bayesian approach. *The Statistician* **52**, 351-365.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of American Statistical Association* **65**, 1350-1361.

York, J., Madigan, D., Heuch, I. and Lie, R. T. (1995). Birth defects registered by double sampling: a bayesian approach incorporating covariates and model uncertainty. *Applied Statistics* **44**, 227-242.

Received April 19, 2011; accepted August 31, 2011.

Dewi Rahardja
Department of Clinical Sciences and Simmons Cancer Center
University of Texas Southwestern Medical Center
Dallas, TX 75390, USA
rahardja@gmail.com

Yan D. Zhao
Department of Clinical Sciences and Simmons Cancer Center
University of Texas Southwestern Medical Center
Dallas, TX 75390, USA
yandzhao@gmail.com

Hongmei Zhang
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC 29208, USA
hzhang@mailbox.sc.edu