Computational Statistics and Data Analysis 55 (2011) 813–823

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Likelihood-based confidence intervals for the risk ratio using double sampling with over-reported binary data

Dewi Rahardja *, Dean M. Young

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

ARTICLE INFO

Article history:
Received 24 November 2009
Received in revised form 7 June 2010
Accepted 5 July 2010
Available online 30 July 2010

Keywords:
Binary data
Double sampling
Misclassification
Relative risk
Risk ratio

ABSTRACT

In this article we derive likelihood-based confidence intervals for the risk ratio using over-reported two-sample binary data obtained using a double-sampling scheme. The risk ratio is defined as the ratio of two proportion parameters. By maximizing the full likelihood function, we obtain closed-form maximum likelihood estimators for all model parameters. In addition, we derive four confidence intervals: a naive Wald interval, a modified Wald interval, a Fieller-type interval, and an Agresti–Coull interval. All four confidence intervals are illustrated using cervical cancer data. Finally, we conduct simulation studies to assess and compare the coverage probabilities and average lengths of the four interval estimators. We conclude that the modified Wald interval, unlike the other three intervals, produces close-to-nominal confidence intervals under various simulation scenarios examined here and, therefore, is preferred in practice.

1. Introduction

Binary data are sometimes obtained when experimental units are classified into two mutually exclusive categories. Generally, a classifier is not perfect and, therefore, misclassified binary data can occur. Usually two types of misclassification are present in misclassified data: false-positive and false-negative. For example, visual inspection by a midwife or obstetrician may erroneously classify a normal child as having Down's syndrome (false-positive), or it may classify a child with Down's syndrome as being healthy (false-negative). In other cases, only one type of misclassification may exist. For instance, Perry et al. (2000) have displayed blood testing data that had only false-positive or over-reported errors, and Moors et al. (2000) have presented auditing data indicating only false-negative or under-reported errors to be present.

Many researchers, including Bross (1954), have demonstrated that classical estimators that ignore misclassification are biased when applied to misclassified binary data. Therefore, additional external information or additional data is needed to correct the bias. Several methods in the literature are dedicated to this purpose. In the Bayesian paradigm, when an infallible procedure is unavailable or prohibitively expensive, one can use informative priors to obtain model identifiability. Another information-producing method is to use multiple fallible classifiers. The method we focus on in this article is using additional training data via a double-sampling scheme first proposed by Tenenbein (1970).

Tenenbein's double-sampling method is used when infallible and fallible classification procedures are available. Usually, the infallible procedure is very expensive; the fallible procedure is usually cheap but is generally prone to error. Therefore, the use of an infallible procedure on only a small portion of the data and a fallible procedure on all the data is an economically feasible means of promoting model identifiability.

A rich literature of research is available on binary data subject to misclassification that provides point and interval estimation methods on various functions of the proportion parameters of interest. For one-sample problems, several researchers have considered the case in which only one type of error is present. Lie et al. (1994) have used a maximum

* Corresponding author.
E-mail address: rahardja@gmail.com (D. Rahardja).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.07.004

Table 1
Data from the fallible method for sample i, $i = 1, 2$.

Classification	0	1	Total
Count	Y_i	X_i	M_i

likelihood approach, where false-negative errors were corrected using multiple fallible classifiers. York et al. (1995) have considered this same problem from a Bayesian perspective. When data are obtained using a double-sampling scheme, Moors et al. (2000) have discussed the method of moment and maximum likelihood estimation, in addition to one-sided interval estimation. Boese et al. (2006) have derived several likelihood-based confidence intervals (CIs) for a single proportion parameter, while Lee and Byun (2008) have provided Bayesian credible intervals using noninformative priors for the same problem.

Moreover, several researchers have also studied one-sample problems with both types of misclassification errors. In conjunction with double sampling, Tenenbein (1970) has proposed a maximum likelihood estimator for a proportion parameter and has derived an expression for the asymptotic variance. For the case when training data is unavailable in one-sample problems, Gaba and Winkler (1992) and Viana et al. (1993) have developed Bayesian approaches using sufficiently informative priors.

For two-sample problems with both types of misclassification errors, Bayesian inference methods using sufficiently informative priors have also been developed when training data is unavailable. For example, see Evans et al. (1996) for risk difference estimation, that is, the difference of two proportion parameters, and Gustafson et al. (2001) for odds ratios. When training data is obtained through a double-sampling scheme, Boese (2003) has derived several likelihood-based CIs for the risk difference.

Recently, Rahardja and Young (2010) developed a Bayesian approach for point and interval estimation of the risk ratio for two-sample misclassified binary data with only false-positive error. The risk ratio (RR) is defined as the ratio of two proportion parameters and is also known as the relative risk. In this article we propose frequentist approaches to the same problem. The remainder of this paper is organized as follows. In Section 2 we describe the data, and in Section 3 we propose four likelihood-based methods for interval estimation of a risk ratio using double sampling with over-reported data. In Section 4 we illustrate the newly derived interval estimation methods using real cervical cancer data. We examine the performance of four proposed interval estimation methods in Section 5 using Monte Carlo simulation, and we provide a brief discussion in Section 6.

2. The data

In this section we describe two-sample misclassified binary data with one type of misclassification error. We assume that the data is obtained using a fallible classification procedure that yields false-positive but not false-negative counts.

We next introduce notation useful for describing the data. Let F_{ij} be the observed classification by the fallible classification for the jth individual in the ith sample, where $i = 1, 2, j = 1, \ldots, M_i$, and

$$F_{ij} = \begin{cases} 1, & \text{if the result is positive by the fallible classifier} \\ 0, & \text{otherwise.} \end{cases}$$

Let $X_i = \sum_j F_{ij}$ and $Y_i = M_i - X_i$ be the observed number of positive and negative classifications, respectively. The data obtained by the fallible classification for sample i, $i = 1, 2$, is displayed in Table 1.

Similarly, we define the unobserved true classification of the jth individual in the ith sample as

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

We remark that misclassification occurs when $T_{ij} \neq F_{ij}$.

Also, we let

$$p_i \equiv \Pr\left(T_{ij} = 1\right),$$

$$\pi_i \equiv \Pr\left(F_{ij} = 1\right),$$

and

$$\phi_i \equiv \Pr\left(F_{ij} = 1 | T_{ij} = 0\right).$$

Here, p_i is the actual proportion parameter of interest, π_i is the proportion parameter of the fallible procedure, and ϕ_i is the false-positive rate for the fallible procedure. We allow the false-positive rates to be different between the two samples, i.e., $\phi_1 \neq \phi_2$. Note that π_1 and π_2 are not additional unique parameters because

$$\begin{aligned} \pi_i &= \Pr(T_i = 1) \Pr(F_i = 1 \mid T_i = 1) + \Pr(T_i = 0) \Pr(F_i = 1 \mid T_i = 0) \\ &= p_i + q_i\phi_i, \end{aligned} \tag{1}$$

where $q_i = 1 - p_i$, $i = 1, 2$.