Statistical Methodology 8 (2011) 204–212

Contents lists available at ScienceDirect

Statistical Methodology

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

Likelihood-based confidence intervals for the risk difference of two-sample binary data with a fallible classifier and a gold standard

Dewi Rahardja *, Yan D. Zhao

Department of Clinical Sciences and Simmons Cancer Center, UT Southwestern Medical Center, Dallas, TX 75390-8822, USA

ARTICLE INFO

Article history:
Received 23 July 2010
Received in revised form
3 September 2010
Accepted 18 September 2010

Keywords:
Binary data
Confidence interval
Misclassification
Risk difference
Validation substudy

ABSTRACT

We develop likelihood-based confidence intervals for risk difference in two-sample misclassified binary data. Such data consist of two studies. The first study is the main study where individuals are classified by an inexpensive fallible classifier which may misclassify. The second study is a validation substudy where individuals are classified by using both the fallible classifier and an expensive gold standard which classifies perfectly. We propose and examine three likelihood-based confidence interval methods and conclude that the modified Wald method applied to small-number adjusted new data performs well and has nominal coverage probabilities.

1. Introduction

We consider two-sample data where the response variable is obtained using a binary classifier which dichotomizes each response into one of two distinctive categories. Interests of such data are to quantify the binomial proportion within each sample and to compare the two binomial proportions. The grouping variable which categorizes individuals to one of the two samples is perfect. However, sometimes, the response classifier is fallible and misclassification may occur. When the extent of misclassification is substantial, it is well known that classical estimators can be severely biased. For example, [2,4] evaluated the bias of classical estimators of risk difference.

If an infallible classifier (gold standard) exists for response classification, then a double-sampling scheme [11] can be used for making valid inference. The double-sampling scheme can be described as follows. We note that the gold standard may be too expensive or time-consuming to be applied to each individual in the original data. Therefore, it is more practical to first randomly select a subset

* Corresponding author.
E-mail addresses: rahardja@gmail.com (D. Rahardja), yandzhao@gmail.com (Y.D. Zhao).

1572-3127/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2010.09.005

Table 1
Data for sample i.

Study	Gold standard	Fallible method		
		0	1	Total
Validation	0	n_{i00}	n_{i01}	$n_{i0\cdot}$
	1	n_{i10}	n_{i11}	$n_{i1\cdot}$
	Total	$n_{i\cdot 0}$	$n_{i\cdot 1}$	n_i
Main	NA	y_i	x_i	m_i

NA: not available.

of the original data or to get another set of individuals beyond the original data, and then obtain the infallible response for this subset (or set) of individuals using the gold standard. If this is a new set of individuals beyond the original data, also obtain the response classification using the fallible classifier. In other words, for this subset (or set) of individuals, both response classifications are obtained. Such a subset (or set) is known as the validation substudy. The complement of this substudy is known as the main study. For the remainder of this article, we are interested in analyzing such data consisting of a main study and a validation substudy.

In epidemiology it is common to analyze binary data obtained using a double-sampling scheme. In such retrospective studies, an appropriate measure to compare the two samples is the odds ratio. For odds ratio, [3] developed maximum likelihood (ML) and pseudo-likelihood (PL) methods for constructing confidence intervals (CI) of the odds ratio. In addition, [5,9] provided matrix method and inverse matrix method for inference on odds ratio, respectively. Furthermore, [12,10] proposed Bayesian methods.

We intend to provide frequentist methods for analyzing prospective binary data obtained using the double-sampling scheme. Such data are commonly collected in the medical field, sociology, and economics. In such data, risk difference is more commonly used than odds ratio to compare two samples. The aforementioned frequentist methods on odds ratio cannot be easily modified to analyze risk difference. To date, no frequentist methods for inference on risk difference have been developed for two-sample misclassified binary data obtained using the double-sampling scheme. In this article, we develop likelihood-based method to tackle this problem. In Section 2 we describe the data. In Section 3 we derived maximum likelihood estimator and three likelihood-based CIs. When constructing CIs, we perform a reparameterization of the model parameters and a transformation of the original data to improve the performance of our CIs. In Section 4 we illustrate three CI methods using a real data. The performance of three methods is examined and compared in Section 5 and a discussion can be found in Section 6.

2. Data

In our data two classifiers are used to classify individuals into two distinctive response categories (0 or 1). The gold standard is used only in the validation substudy, while the fallible classifier is used in both the main study and the validation substudy. There are two samples in both the main study and the validation study. For Sample i (1 or 2), let m_i and n_i be the number of individuals in the main study and the substudy, respectively. In addition, we define $N_i = m_i + n_i$ as the sample size for Sample i.

For the jth individual in the ith sample, where $i = 1, 2$ and $j = 1, \ldots, N_i$, let F_{ij} and T_{ij} be the classifications by the fallible classifier and the gold standard, respectively. We denote $F_{ij} = 1$ if the result is positive by the fallible classifier and $F_{ij} = 0$ otherwise. Similarly, we denote $T_{ij} = 1$ if the result is truly positive by the gold standard and $T_{ij} = 0$ otherwise. Note that F_{ij} is observed for all individuals in both the main study and the validation study, while T_{ij} is observed for individuals in the validation study but not in the main study. Clearly, misclassification occurs when $F_{ij} \neq T_{ij}$.

In the validation study, we use n_{ijk} to denote the number of individuals in Sample i classified as j and k by the infallible classifier and the gold standard, respectively. In the main study, let x_i and y_i be the number of positive and negative classifications in Sample i by the fallible classifier, respectively. The summary counts in both the main study and the validation study for sample i are displayed in Table 1.