Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 3369 – 3385
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Confidence intervals for a binomial parameter based on binary data subject to false-positive misclassification

Doyle H. Boese[a,*], Dean M. Young[b], James D. Stamey[b]

[a] *Infineum USA Inc, Linden, NJ 07036, USA*
[b] *Department of Statistical Sciences, Baylor University, Waco, TX 76798-8005, USA*

Received 7 June 2004; received in revised form 20 August 2005; accepted 23 August 2005
Available online 12 September 2005

Abstract

In this paper we derive five first-order likelihood-based confidence intervals for a population proportion parameter based on binary data subject to false-positive misclassification and obtained using a double sampling plan. We derive confidence intervals based on certain combinations of likelihood, Fisher-information types, and likelihood-based statistics. Using Monte Carlo methods, we compare the coverage properties and average widths of three new confidence intervals for a binomial parameter. We determine that an interval estimator derived from inverting a score-type statistic is superior in terms of coverage probabilities to three competing interval estimators for the parameter configurations examined here. Utilizing the expressions derived, we also determine confidence intervals for a binary parameter using real data subject to false-positive misclassification.

Keywords: Double sampling; Misclassification; False positive; Binary data; Confidence interval; Likelihood statistic

1. Introduction

Inference based on binary data subject to misclassification is discussed in the statistical literature for numerous applications. For instance, Bross (1954) first showed that estimators

* Corresponding author. Tel.: +1 908 4743176; fax: +1 908 4743241.
E-mail address: dboese@comcast.net (D.H. Boese).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2005.08.007

based on data subject to misclassification can be extremely biased when one uses conventional estimators. Cohen (1960) utilized maximum likelihood methods to estimate the population proportion of defects when a defective observation is, on occasion, intentionally recorded as non-defective to allow acceptance of a batch of items that otherwise would have been rejected.

Tenenbein (1970) popularized the double sampling plan for binary data and derived estimators for the population proportion as well as for false-positive and false-negative error rates. Tenenbein (1972) also derived maximum likelihood estimators for multinomial categorical parameters.

Several authors have considered the case where only false-negative counts are obtained. Lie et al. (1994) consider a maximum likelihood approach to the problem where false-negative counts are corrected using multiple fallible classifiers. York et al. (1995) consider this same problem from the Bayesian perspective. Moors et al. (2000) discuss maximum likelihood estimation and one-sided interval estimation for the case where only false-negative misclassification occurs and apply their methods to auditing data. In addition, several authors, including Gaba and Winkler (1992) and Evans et al. (1995), have used Bayesian methods for these applications while assuming both types of misclassification.

To date, two-sided frequentist confidence intervals for a proportion parameter of interest using possibly misclassified data are yet to be derived and investigated. In this paper we derive five likelihood-related large-sample confidence intervals for a binomial parameter derived from two Wald statistics, two score statistics, and a log profile-likelihood difference statistic. The two Wald statistics and the two score statistics are formed using observed Fisher's information and restricted information. Additionally, we derive a log profile-likelihood difference statistic formed with the profile likelihood function of our binomial model estimated with false-positive data. We also consider an asymptotically normal MLE-based confidence interval for a binomial proportion parameter. We refer to this interval estimator as the asymptotic normal (*AN*) interval throughout the remainder of this paper.

We then use a Monte Carlo simulation to compare the coverage and width properties of the four competing confidence intervals.

We have organized the remainder of the paper as follows. We define our false-positive binomial model in Section 2. Likelihood-related interval estimators are briefly reviewed in Section 3, and in Section 4 we derive expressions for the log profile-likelihood difference statistic and four confidence intervals derived from Wald and score statistics with different types of information. In Section 5 we describe a Monte Carlo simulation and report the results of the simulation to compare four confidence intervals—three that we derive in Section 4 as well as an asymptotic normal-based interval. In Section 6 we apply our new confidence intervals to a real-data example and give a brief discussion of our new intervals in Section 7.

2. The false-positive model

In some cases when one estimates a proportion parameter, only one misclassification type may exist naturally or one misclassification type may be minimized by design. The