Computational Statistics and Data Analysis 95 (2016) 122–132

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Confidence intervals for the ratio of two Poisson rates under one-way differential misclassification using double sampling

David J. Kahle [a,*], Phil D. Young [a,b], Brandi A. Greer [a], Dean M. Young [a]

[a] Department of Statistical Science, Baylor University, Waco, TX 76798-7140, United States

[b] Department of Information Systems, Baylor University, Waco, TX, 76798-7140, United States

ARTICLE INFO

Article history:
Received 25 February 2015
Received in revised form 9 September 2015
Accepted 9 September 2015
Available online 8 October 2015

Keywords:
Poisson rates
Differential misclassification
Double-sampling
Under-reporting
R

ABSTRACT

Wald, profile likelihood, and marginal likelihood confidence intervals are derived for the ratio of two Poisson rates in the presence of one-way differentially misclassified data using double sampling. Monte Carlo simulations demonstrate the reliability and relative performance of the intervals, and an example from cancer epidemiology illustrates their application and interpretation in a real-world scenario. All of the methods described are implemented and freely available in the R package **poisDoubleSamp** on the Comprehensive R Archive Network (CRAN).

1. Introduction

When analyzing prevalence rates of a disease across two groups of subjects, differential misclassification refers to the systematic mislabeling of the disease for an alternative at different rates depending on group membership (Gordis, 2014). Such misclassification is an obvious source of information bias with the potential to mask effects of interest relating to the two groups. Consequently, methods that mitigate against such errors are preferable, when available, because they can provide more reliable conclusions. Unfortunately such methods are not as pervasive as one might hope. In this article we consider one-way differential misclassification; that is, we consider the scenario in which the disease of interest is at risk of being misclassified as another ailment, while other ailments are not at risk for being misclassified as the disease of interest.

Double-sampling is often applied to collect count data over a vast observation-opportunity size and is beneficial because it can be utilized to significantly reduce bias in situations of misclassification. Hiridoglou and Sarndal (1998) and Hiridoglou (2001) have provided well-rounded introductions to the theory behind double-sampling, while Chanu and Singh (2014), Choudhury and Singh (2012), and Vishwakarma and Gangele (2014) are among the many who have investigated various aspects of the paradigm within the last few years. A frequently used method for comparing two Poisson rates is to determine and analyze the ratio of the rates. Recently, Gu et al. (2008), Feng et al. (2009), Laurent (2012), Barker and Cadwell (2008), and Sahai and Khurshid (1993) have been among the many who have embraced this method of comparison. In this article we combine the two strategies by employing a double-sampling procedure with count data to determine Wald, profile likelihood, and marginal likelihood confidence intervals (CIs) for the ratio of Poisson rate parameters for two populations. The work described is implemented in the R package **poisDoubleSamp**, which is currently available through the Comprehensive R Archive Network (CRAN), and examples are provided throughout to showcase how the theory can be brought to practice (Kahle et al., 2015; R Core Team, 2014).

* Corresponding author. Tel.: +1 254 710 6102.
E-mail address: david.kahle@gmail.com (D.J. Kahle).

http://dx.doi.org/10.1016/j.csda.2015.09.013
0167-9473/

Table 1
Notation for the Poisson rates used in the running example.

	France	Italy
Cervical cancer	λ_{11}	λ_{21}
Cancer (other)	λ_{12}	λ_{22}

Table 2
The two datasets as they are observed in practice.

(a) The fallible dataset

	$i=1$	$i=2$
$j=1$	z_{11}	z_{21}
$j=2$	z_{12}	z_{22}
Opportunity	N_1	N_2

(b) The infallible dataset

	$i=1$	$i=2$
$j=1$	m_{011}	m_{021}
$j=2$	m_{012}	m_{022}
Under-reported	y_{01}	y_{02}
Opportunity	N_{01}	N_{02}

Example. In our opinion the most efficient and effective way to communicate the topics discussed in this article is with a running example. We describe the study generally in the following example, which we continue throughout the article. Section 5 further explains the quantities referred to in the context of a real-world dataset.

Suppose we wish to compare the mortality rates of cervical cancer in France and Italy. To that end, we represent the numbers of cervical cancer deaths in each country as Poisson variates with rates λ_1 for French women and λ_2 for Italian women. Now, the possibility exists that women who succumbed to cervical cancer are reported as succumbing to a different type of cancer. By aggregating these different types of cancer together as simply "Cancer (Other)", we can reasonably assume that the sum total of other cancer deaths is also Poissonian. The four categories can therefore be represented with a 2×2 contingency table with a Poisson sampling scheme as in Table 1.

Note that the first index, i, corresponds to the population: France is $i = 1$ and Italy is $i = 2$. The second index, j, corresponds to the cause of death: cervical cancer is $j = 1$ and cancer (other) is $j = 2$.

What makes this problem interesting is the misclassification of patients that have succumbed to cervical cancer as having succumbed to another type of cancer, thus creating under-reporting of cervical cancer deaths. We represent this aspect of the problem with under-reporting parameters for each population, denoted θ_i, using a binomial model. Thus, θ_1 and θ_2 denote the under-reporting proportions for French and Italian women, respectively. These rates and under-reporting parameters occur over an observation-opportunity size denoted N_i so that the number of deaths in each case is modeled as $\text{Pois}(N_i\lambda_{ij})$. ‖

We now more precisely and generically describe the data involved in such a study. The double-sampling scheme results in two datasets from each population. The first dataset, called the fallible dataset, consists simply of a contingency table of the observed death counts from both causes in both populations, see Table 2(a); it also contains the observation-opportunity sizes from which the data were gathered. The dataset described in the example above was the fallible dataset. The second dataset, called the infallible dataset, consists of a contingency table as before, but also includes the number of under-reported deaths in each setting learned from a gold standard, see Table 2(b). This dataset is typically far smaller because the resources required to obtain it are typically much larger.

The notation used to describe the data reflects the subtle nature of the problem, so a notational description is essential before we proceed. We begin with components of the fallible dataset. Let m_{ij} denote the actual, but not the observed, count from population i in category j. These counts are latent in the fallible dataset. Let y_i denote the number of counts in group 1 of population i misclassified into group 2, and define z_{ij} to be the error-prone observed count. Using the previously stated model assumptions, one can easily determine

$$[m_{ij}|N_i, \lambda_{ij}] \sim \text{Pois}(N_i\lambda_{ij}), \tag{1}$$

$$[z_{i1}|N_i, \lambda_{i1}, \theta_i] \sim \text{Pois}(N_i\lambda_{i1}\,(1-\theta_i)), , \tag{2}$$

$$[z_{i2}|N_i, \lambda_{i1}, \lambda_{i2}, \theta_i] \sim \text{Pois}(N_i(\lambda_{i2} + \lambda_{i1}\theta_i)), \tag{3}$$

and

$$[y_i|m_{i1}, \theta_i] \sim \text{Bin}(m_{i1}, \theta_i). \tag{4}$$

The notation used for the infallible dataset is the same as that of the fallible dataset, save for a "0" prepended to the index. As a visual reference, a graphical model diagram for the model is included in Fig. 1.

Example (*continued*)**.** Continuing our example, we have that m_{11} and m_{12} reflect the actual number of French women that died from cervical cancer and other types of cancer in the fallible dataset; m_{21} and m_{22} indicate similar counts for Italian women. The under-reported counts, also latent, are denoted y_i, where y_1 and y_2 denote the number of misclassified French and Italian women, respectively, who died because of cervical cancer but were classified as having died from other causes.