Statistical Methodology 22 (2015) 17–22

Contents lists available at ScienceDirect

Statistical Methodology

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

On the Chao and Zelterman estimators in a binomial mixture model

Chang Xuan Mao *, Nan Yang, Jinhua Zhong

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China

ARTICLE INFO

Article history:
Received 27 June 2012
Received in revised form
14 June 2014
Accepted 14 June 2014
Available online 23 June 2014

Keywords:
Capture–recapture
Population size

ABSTRACT

Data from a surveillance system can be used to estimate the size of a disease population. For certain surveillance systems, a binomial mixture model arises as a natural choice. The Chao estimator estimates a lower bound of the population size. The Zelterman estimator estimates a parameter that is neither a lower bound nor an upper bound. By comparing the Chao estimator and the Zelterman estimator both theoretically and numerically, we conclude that the Chao estimator is better.

1. Introduction

Estimating population sizes has been thoroughly investigated [5]. There are various mechanisms from which data are generated. Particularly, we will focus on applications in which individuals are subject to detection via surveillance systems. For instance, a household can serve as a useful unit of disease surveillance, and a binomial mixture model can arise by assuming that the number of disease cases in a household is binomial and that the probability that one person is infected is allowed to vary over households [1,16,14,15,8–10]. There are various epidemiological applications of the binomial mixture model (e.g., [6,18,19]).

We will use the nonparametric binomial mixture model. Some parametric mixture models were considered in the literature and their disadvantages were discussed in [13]. In the nonparametric model, each of the estimators in [2,20] developed in the Poisson setting and favored by epidemiologists, admits a counterpart in the present binomial setting [4,17]. If the size parameter for binomial

* Corresponding author. Tel.: +86 21 56901035.
E-mail addresses: mao.changxuan@mail.shufe.edu.cn (C.X. Mao), yangnan@mail.shufe.edu.cn (N. Yang).

http://dx.doi.org/10.1016/j.stamet.2014.06.002
1572-3127/

distribution is larger than two, then the Zelterman estimator consistently estimates a parameter that is neither a lower bound nor an upper bound of the population size, while the Chao estimator consistently estimates a lower bound. The Zelterman estimator may yield confidence intervals for the population size with poor coverage probabilities. Although the Zelterman estimator was declared to be robust, its performance is worse than the Chao estimator. The Chao estimator deserves our recommendation.

The results are presented in Section 2. A simulation experiment and a real example are reported in Section 3. The proofs are provided in the Appendix.

2. Results

Suppose that there are s individuals subject to t times of detection, $t \geqslant 2$. Let X_i be the number of times when individual i is detected, which is binomial with size t and probability π_i. The X_i are independent given the π_i that arise as a random sample from a mixing distribution G. When G is discrete, it can be written as $G = \sum_{k=1}^{K} \psi_k \delta(\varpi_k)$, where $\psi_k \geqslant 0$, $\sum_{k=1}^{K} \psi_k = 1$, and $\delta(\pi)$ is a distribution degenerate at π. Marginally, the X_i arise as a random sample from a mixture density

$$h_G(x) = \int \binom{t}{x} \pi^x (1-\pi)^{t-x} \, dG(\pi) = \sum_{k=1}^{K} \psi_k \binom{t}{x} \varpi_k^x (1-\varpi_k)^{t-x}, \quad x = 0, 1, \ldots, t.$$

Let $n_x = \sum_{i=1}^{s} I(X_i = x)$. The number of observed individuals is $n_+ = \sum_{x=1}^{t} n_x$. Given n_+, $(n_1, n_2, \ldots, n_t)^\top$ is multinomial with index n_+ and cell probabilities $h_G(x)/\{1 - h_G(0)\}$. To reformulate these cell probabilities, we use a mixing distribution $Q = \sum_{k=1}^{K} \eta_k \delta(\varpi_k)$, where

$$\eta_k = \frac{\{1 - (1 - \varpi_k)^t\}\psi_k}{\sum_{\kappa=1}^{K} \{1 - (1 - \varpi_\kappa)^t\}\psi_\kappa}.$$

It is easily shown that $h_G(x)/\{1 - h_G(0)\}$ can be written as a mixture density

$$f_Q(x) = \int \binom{t}{x} \frac{\pi^x (1-\pi)^{t-x}}{1 - (1-\pi)^t} \, dQ(\pi) = \sum_{k=1}^{K} \eta_k \binom{t}{x} \frac{\varpi_k^x (1-\varpi_k)^{t-x}}{1 - (1-\varpi_k)^t}.$$

These facts yield a conditional likelihood

$$L(Q) = \frac{n_+!}{\prod_{x=1}^{t} n_x!} \prod_{x=1}^{t} \left\{ f_Q(x) \right\}^{n_x}.$$

The nonparametric maximum likelihood estimator (NPMLE) $\widehat{Q}$ satisfies $L(Q) \leqslant L(\widehat{Q})$, $\forall Q$, which is discrete with finitely many support points. To ensure its existence, $\widehat{Q}$ is allowed to put some mass on zero. Moreover, $\widehat{Q}$ is not necessarily unique. The reason is that the model of mixtures of zero-truncated binomial densities is non-identifiable in the sense that there are mixing distributions Q and M with $Q \neq M$ and $f_Q = f_M$.

To estimate the population size s, we observe that

$$s = \frac{E(n_+)}{1 - h_G(0)} = E(n_+) + E(n_+) \cdot \frac{h_G(0)}{1 - h_G(0)}.$$

The problem becomes estimating the odds $h_G(0)/\{1 - h_G(0)\}$ that can be re-written as

$$\theta(Q) = \int \frac{(1-\pi)^t}{1 - (1-\pi)^t} \, dQ(\pi) = \sum_{k=1}^{K} \eta_k \frac{(1-\varpi_k)^t}{1 - (1-\varpi_k)^t}.$$

The odds $\theta(Q)$ is non-identifiable in the sense that there are mixing distributions Q and M such that $\theta(Q) \neq \theta(M)$ and $f_Q = f_M$ [9,10]. This invites one to ask what has been estimated by an existing estimator for the population size s. The fact is that, given an estimator for s, it admits a *de facto* estimand that may differ from s.