Journal of Statistical Planning and Inference 145 (2014) 165–178

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A regression estimator for mixed binomial capture–recapture data

Irene Rocchetti [a], Marco Alfó [b,*], Dankmar Böhning [c]

[a] ISTAT—Direzione Centrale dei Dati Amministrativi e dei Registri Statistici, Italy
[b] Dipartimento di Scienze Statistiche, "Sapienza" Universitá di Roma, Italy
[c] Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, UK

ARTICLE INFO

Article history:
Received 2 November 2012
Received in revised form
30 July 2013
Accepted 2 August 2013
Available online 11 August 2013

Keywords:
Beta-binomial
Weighted regression
Zero-truncation

ABSTRACT

Mixed binomial models are frequently used to provide estimates for the unknown size of a partially observed population when capture–recapture data are available through a known, finite, number of identification (sampling) sources. In this context, inherently major problems may be the lack of identifiability of the mixing distribution (Link, 2003) and *boundary problems* in ML estimation for mixed binomial models (such as the beta-binomial or finite mixture of binomials), see e.g. Dorazio and Royle (2003, 2005). To solve these problems, we introduce a novel regression estimator based on observed ratios of successive capture frequencies. Both simulations and real data examples show that the proposed estimator frequently leads to under-estimate the *true* population size, but with a smaller bias and a lower variability when compared to other well-known estimators.

1. Introduction

Capture–recapture methods are often used to estimate the unknown size of a *partially observed* population, through samples derived using some identification mechanisms (traps, lists, registers, etc.). These methods have been introduced in the wildlife setting to estimate animal abundance, and extended to epidemiology, public health, quality control, etc., see Chao et al. (2001), Roberts and Brewer (2006), and Böhning and Patilea (2008).

Throughout the paper, we will consider an endogenous mechanism, e.g. a register, which identifies n units from a closed population of unknown size N. The number, m, of sampling occasions (sources) is assumed to be fixed and known; the number of units identified by the mechanism exactly x times is denoted by n_x, $x=1,\ldots,m$, and the number of units identified at least once is $n=n_1+n_2+\cdots+n_m$. Since $N=n_0+n$, cap–recap methods use information on n_x, $x=1,\ldots,m$ to estimate n_0 or, alternatively, N. A common estimation approach is to model the number of times a unit has been identified through a counting distribution. Let p_x $x=1,\ldots,m$ denote the conditional probability of exactly x identifications for a generic unit; we know that the (conditional) maximum likelihood estimator of N is the integer part of the Horvitz–Thompson estimator $\hat{N}=\lfloor n(1+\theta_0)\rfloor$ where $\theta_0=p_0/(1-p_0)$ represents the odds that an individual is unseen. Therefore, to estimate N, we need to estimate p_0 or n_0. According to the hypothesis of a fixed, known, number of sampling occasions (sources), the number of times the ith individual has been identified may be described by a binomial distribution with (possibly subject-specific) probability, π_i, $i=1,\ldots,N$. Heterogeneity may be observed, and summarized by a covariate vector, or unobserved. We discuss the last case where individual-specific variation in identification probabilities is described by a (possibly) known

* Corresponding author. Tel.: +39 0649916072.
E-mail address: marco.alfo@uniroma1.it (M. Alfó).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.08.010

parametric distribution, that is $\pi_i \sim G(\cdot|\xi)$. We start from a simple beta-binomial model and propose a population size estimator based on a regression model for ratios of successive capture frequencies, to avoid boundary issues which often represent a problem in ML approaches, see Dorazio and Royle (2003, 2005) and Mao and You (2009). While the estimator is derived by adopting a beta-binomial model, we show in a simulation study that it may be applied with satisfactory results to general mixed binomial models.

The outline of the paper is as follows: Section 2 reviews recent developments in mixed binomial models for capture–recapture data. In Section 3 the proposed estimator is derived using an Empirical Bayes approach. In Sections 4 and 5 the behavior of the proposed estimator is investigated using the simulation scheme of Pledger (2005), and the analysis of real datasets. Section 6 contains concluding remarks and future research agenda.

2. Modeling unobserved heterogeneity

To estimate the size of a population of interest, we use mixed binomial models, where the mixing distribution models individual-specific heterogeneity in the target population, which could be due to the effect of unobserved covariates. In this context, it could be interesting to work with data representing the *full* capture configuration for a given individual (e.g. 010010 meaning the unit has been captured by two out of six sources); however, in some circumstances, the entire sequence is unknown and the only information we have on the capture history is the total frequency of capture. In these cases, we may not adopt behavioral and/or time response models (see e.g. Otis et al., 1978) and turn to mixed binomial distributions. The estimator we propose is a simple and reliable tool to provide a population size estimate when no time-, individual- or source-specific information has been recorded.

2.1. Mixed models

Discrete (see e.g. Norris and Pollock, 1996; Pledger, 2000) and continuous mixtures (see e.g. Coull and Agresti, 1999; Dorazio and Royle, 2003) have been used to mitigate the potential bias in population size estimates, arising should one not be able to account for individual-specific characteristics. Finite mixture (latent class) models may help reduce the bias by partitioning units into two or more homogeneous groups, see Norris and Pollock (1996) and Mooijaart and van der Heijden (1992); however, finite mixture models may underestimate the population size if additional variation exists within each group (see e.g. Coull and Agresti, 1999). For this reason Morgan and Ridout (2008) propose a two-component mixture with binomial and beta-binomial kernels.

When individual covariates are available, they may be used to account for individual observed heterogeneity, see e.g. Huggins (2002). Bartolucci and Forcina (2001) extend this approach, and propose a Rasch-type model where subjects are homogeneous within a finite set of latent classes and stratified according to a set of discrete covariates. Bartolucci and Pennoni (2007) model observed/latent heterogeneity, through a latent class model where sources are ordered with respect to time. In many empirical cases, covariates are not recorded and mixed binomial distributions, an example of M_h models see e.g. Otis et al. (1978), have to be used to account for individual variation in detection probabilities. In the following, we will focus on these models.

2.2. The choice of a mixing distribution

On the basis of an extensive simulation study, Pledger (2005) claims that neither finite mixture nor beta-binomial models can be proved to outperform the others regardless of the true, but unknown, mixing distribution. As far as the beta-binomial model is concerned, several authors have pointed out its low precision in ML estimates of N; this has been referred to as weak identifiability of model parameters, see e.g. Burnham and Overton (1978, 1979). Failures in the beta-binomial model are also observed when the maximized log likelihood is achieved near the boundaries of the parameter space (Dorazio and Royle, 2003); similar boundary problems occur as well in finite mixture models, see Mao and You (2009). Besides low precision and boundary problems, identifiability of model parameters is of great concern when ML approaches are employed to estimate the population size by mixed binomial models. In this context, estimating the population size reduces to estimating the marginal odds that an individual is unseen (Sanathanan, 1977). This turns out to be completely nonidentifiable in the general class of nonparametric models (Link, 2003); if we fix a specific mixing distribution, it can be identified when the number of capture occasions is not smaller than four (Holzmann et al., 2006). However, as noted by Dorazio and Royle (2003) and Coull and Agresti (1999), several mixed models may fit data reasonably well, and produce substantially different estimates for the population size. It can be proved that two different mixing distributions, say $Q_1 \neq Q_2$, with different *untruncated* marginal distributions may lead to identical *truncated* marginal distributions, i.e. $P_{Q_1}(x) = P_{Q_2}(x)$, $\forall x = 1, \ldots, m$. Since the population size estimates are based on the truncated distributions, different estimates of N are obtained on the basis of identical truncated marginal distributions. Therefore, P_Q is nonidentifiable and the same argument applies to marginal odds an individual is unseen. For these reasons, the choice of the mixing distribution cannot be based on conventional goodness of fit measures; only prior knowledge of potential sources of heterogeneity in individual propensities of capture can be considered when selecting models. Due to identifiability issues, a sensitivity analysis of the population size estimates with respect to different assumptions upon the mixing distribution is mandatory. For example, the ratio plot mentioned by Hoaglin (1980) in supplement of the Poissoness plot, to check for

homogeneity in Poisson-type distributions, and extended by Hoaglin and Tukey (1985) to other discrete distributions, can be used to detect substantial departures from homogeneity.

Starting from the idea of the ratio plot, and extending the proposal by Rocchetti et al. (2011), we introduce a regression estimator which is based on an alternative parameterization of the odds which does not suffer of any of the weak identifiability/boundary issues discussed above. The estimator we propose extends the proposal by Rocchetti et al. (2011) in several aspects. Here, we aim at estimating the total amount of an unknown population when dealing with a few, fixed, number of capture occasions, while Rocchetti et al. (2011) discuss the case of a large, potentially unknown, number of sampling occasions. The situation we discuss in the present paper is equally common in practice because of the availability of a few number of sources (due to different purposes of each other and/or privacy issues which make difficult the correct identification of recaptured individuals) which can be integrated in order to produce more precise prevalence estimates. In the case of mixed binomial distributions, as pointed out by Link (2003), different mixing distributions, each with support bounded away from zero, may produce identical sampling distributions for the observed data, but lead to inconsistent inferences about n_0. This makes it impossible to define the *best* mixing distribution and, thus, to establish the *best* estimate for the probability of missing a unit (p_0). We introduce a suitable parameterization, in terms of the posterior odds parameter, that allows us to provide a non-parametric estimate of the quantity of interest, overcoming weak identifiability and boundary issues in beta-binomial models. While the approach in Rocchetti et al. (2011) applies to the distributions in the Katz family, the approach we propose is based on the extension of the Chao (1989) inequality to m sources, that is on the monotonicity of posterior odds, and therefore applies to the whole family of mixed binomial distributions.

3. An empirical Bayes estimator

Let us start by assuming that the observed counts are drawn from a truncated beta-binomial distribution, since the zero count is not observed. To estimate n_0 (or equivalently the population size given the equality $N = n + n_0$), we define a linear regression for (adjusted) ratios of successive frequency counts; the estimator for n_0 is obtained by projecting the regression function back to zero. The starting point comes from a simple extension of the Chao (1989) lower bound for binomial mixtures when $m > 2$ sampling occasions are considered; monotonicity of the posterior odds parameter is proved in Appendix A. Let us consider a binomial distribution with a number of trials equal to m, and probability of success π. The number of successes (the number of times an individual has been registered) is denoted by $x = 0, 1, 2, \ldots, m$. The binomial distribution can be written as

$$P(X = x|\pi, m) = \binom{m}{x}\pi^x(1-\pi)^{m-x} = \alpha_x\theta^x\mu(\theta)$$

where $\theta = \pi/(1-\pi)$, $\mu(\theta) = (1+\theta)^{-m}$, $\alpha_x = \binom{m}{x}$. If we model unobserved individual-specific variation in π through a mixing distribution $g(\cdot)$, the marginal pdf is

$$P(X = x|m) = P_x = \int \alpha_x\theta^x\mu(\theta)g(\theta)\,d\theta$$

and the posterior mean for θ is

$$\theta_{x,B} = \frac{\int \alpha_x\theta^{x+1}\mu(\theta)g(\theta)\,d\theta}{\int \theta^x\alpha_x\mu(\theta)g(\theta)\,d\theta} = \frac{P_{x+1}\alpha_x}{P_x\alpha_{x+1}} = \left(\frac{x+1}{m-x}\right)\frac{P(x+1|m)}{P(x|m)}$$

Note that *Px* relates to the marginal probability whereas *Px* defined in the introduction is a general notation for the probability of observing exactly x recaptures. The posterior odds conditional to x captures is proportional, for fixed m, to the ratio between the marginal distribution evaluated at $x+1$ and x. In the case of a homogeneous binomial model, we have $\theta_{x,B} = \pi/(1-\pi)$ and, therefore, the plot for the couples $(x, \theta_{x,B})$ would be constant and parallel to the x-axis. The plot can be easily linked to *Poissoness* plot introduced by Hoaglin (1980) and extended by Hoaglin and Tukey (1985) to binomial and other discrete distributions. In the case of mixed binomial models, the ratio-plot can be proven to be monotone non-decreasing with x, see Appendix A for a simple proof. Therefore, we have that

$$\frac{P_x}{\alpha_x}\Big/\frac{P_{x-1}}{\alpha_{x-1}} \le \frac{P_{x+1}}{\alpha_{x+1}}\Big/\frac{P_x}{\alpha_x}. \tag{1}$$

When $x = 1$, the previous inequality leads to

$$P_0 \ge \frac{P_1^2(m-1)}{2mP_2} \tag{2}$$

defining a lower bound for P_0 (the marginal probability of missing a unit), see Chao (1989). Replacing the unknown probabilities by the observed frequencies n_x/N, the Chao lower bound estimator is achieved:

$$\hat{n}_0^C = \frac{n_1^2(m-1)}{2n_2m} \Rightarrow \hat{N}_C = n + \hat{n}_0^C = n + \frac{n_1^2(m-1)}{2n_2m}.$$

We will use the monotonicity result to motivate the proposed estimator. Let us write

$$\theta_{x,B} = h(x;\phi)$$

where $h(\cdot)$ is an appropriate *response* function and ϕ denotes the corresponding parameter vector. Let us replace $\theta_{x,B}$ by its empirical counterpart

$$\hat{\theta}_{x,EB} = \frac{(x+1)n_{x+1}}{(m-x)\ n_x}$$

We may estimate ϕ by least squares and solve for $x=0$, to get

$$\hat{\theta}_{0,EB} = \frac{1\, n_1}{m n_0} = h\left(0;\hat{\phi}\right) \Rightarrow \hat{n}_0 = \frac{n_1}{m h(0;\hat{\phi})} \tag{3}$$

This approach is quite general and applies to the whole class of mixed binomial distributions; it can be specialized when $\pi \sim \text{Beta}(\alpha,\beta)$; in this case, the marginal distribution is

$$P_x = \alpha_x \frac{B(x+\alpha, m-x+\beta)}{B(\alpha,\beta)} \quad \text{where } B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

and $\Gamma(x)$ is the ordinary Gamma function. The posterior mean is given by

$$\theta_{x,B} = \frac{P_{x+1}\alpha_x}{P_x\alpha_{x+1}} = \frac{(x+\alpha)}{(m-x-1+\beta)} \frac{B(x+\alpha, m+\beta-x)}{B(x+\alpha, m+\beta-x)} = \frac{x+\alpha}{(m-x-1+\beta)}$$

that is, a non-linear function in x. Let us consider the following monotone transform of $\theta_{x,B}$:

$$\left(\frac{\theta_{x,B}}{1+\theta_{x,B}}\right) = \frac{\dfrac{P_{x+1}\alpha_x}{P_x\alpha_{x+1}}}{1+\dfrac{P_{x+1}\alpha_x}{P_x\alpha_{x+1}}} = \frac{\dfrac{x+\alpha}{(m-x-1+\beta)}}{1+\dfrac{x+\alpha}{(m-x-1+\beta)}} = \frac{x+\alpha}{(m+\alpha+\beta-1)}. \tag{4}$$

Rewriting

$$\frac{\alpha}{(m+\alpha+\beta-1)} = \gamma \quad \text{and} \quad \frac{1}{(m+\alpha+\beta-1)} = \delta$$

Eq. (4) leads to the linear regression model

$$\left(\frac{\theta_{x,B}}{1+\theta_{x,B}}\right) = \gamma + \delta x. \tag{5}$$

Estimates for γ and δ can be obtained by plugging in observed frequencies on the left-hand side of (5)

$$\frac{\alpha_x n_{x+1}}{\alpha_{x+1} n_x + \alpha_x n_{x+1}} = \gamma + \delta x$$

where $\delta \in (0,1)$ and $\gamma \in (0,1)$. We may use ordinary least squares to provide estimates $(\hat{\gamma},\hat{\delta})$, and invert the relation at $x=0$ to obtain the estimate for n_0

$$\frac{n_1\alpha_0}{n_1\alpha_0 + n_0\alpha_1} = \hat{\gamma} \Rightarrow \hat{n}_0 = \frac{n_1(1-\hat{\gamma})}{m\hat{\gamma}}. \tag{6}$$

Comparing expression (6) with expression (3), we may notice that, in this case, $h(0;\hat{\phi}) = \hat{\gamma}/(1-\hat{\gamma})$. By using the previous regression estimator, we are implicitly assuming that the monotone non-decreasing trend in $\theta_{x,B}$ holds, at least approximately, also when the marginal probabilities are estimated through relative frequencies n_x/N, $x=1,\ldots,m-1$, even if sample (random) variation in the n_x could break this monotonicity. In practice, it is preferable to fit the response on a logarithmic scale, which is approximately linear near the origin and avoids negative fitted values, which can occur when $\hat{\gamma} < 0$. Negative estimates could also be avoided by defining appropriate non-linear transforms of γ and δ, and employing nonlinear least squares (NLS, see e.g. Lawson and Hanson, 1974). For sake of simplicity, we will not pursue this approach further. By using a log-transform, we get

$$\begin{aligned}\log\left(\frac{\theta_{x,B}}{1+\theta_{x,B}}\right) &= \log\left(\frac{x}{m-1+\beta+\alpha} + \frac{\alpha}{m-1+\beta+\alpha}\right) \\ &= -\log(m-1+\beta+\alpha) + \log(x+\alpha).\end{aligned} \tag{7}$$

Applying McLaurin expansion around α, we obtain

$$\log\left(\frac{\theta_{x,B}}{1+\theta_{x,B}}\right) \simeq \log\left(\frac{\alpha}{m-1+\beta+\alpha}\right) + \frac{x}{\alpha} = \gamma_1 + \delta_1 x, \quad \gamma_1 < 1, \delta_1 \in \mathbb{R}_+$$

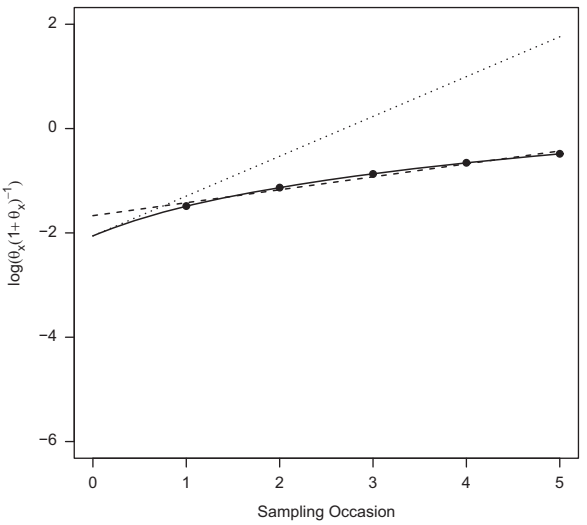

Fig. 1. Plot of ratios $r_x = \log(\theta_{x,EB}/(1+\theta_{x,EB}))$ versus x. Data simulated from a mixed binomial distribution. Mixing distribution Beta(1.76, 9.99). $N=100$, $m=6$. *True* curve (solid line), regression estimator N_r (dashed line), and tangent line at $x=0$ (dotted line).

By plugging-in observed frequencies, we obtain the following regression model:

$$r_x = \log\left(\frac{\hat{\theta}_{x,EB}}{1+\hat{\theta}_{x,EB}}\right) = \log\left[\frac{(x+1)n_{x+1}}{(m-x)n_x + (x+1)n_{x+1}}\right] \simeq \gamma_1 + \delta_1 x \tag{8}$$

Note that estimates $\hat{\gamma}_1$ and δ_1 can be obtained by weighted least squares, as detailed below. Solving for n_0, at $x=0$, we obtain the following estimate:

$$\hat{n}_0^r = \frac{n_1\left[1-\exp(\hat{\gamma_1})\right]}{\exp(\hat{\gamma_1})m} \quad \text{and} \quad \hat{N}_r = \hat{n}_0^r + n \tag{9}$$

where the suffix r stands for *regression*. In this case, $h(0;\hat{\phi}) = \exp(\hat{\gamma}_1)/(1-\exp(\hat{\gamma}_1))$.

While we have used the beta-binomial model as motivation for the proposed estimator, we must notice that all we really need is that $\log[\theta_{x,B}(1+\theta_{x,B})]$ follows, at least approximately, a linear pattern. We should stress that the regression model in Eq. (8) includes, as a special case, the binomial distribution as a reference term, that can be obtained for $\delta=0$. In this sense, our aim is not that of testing for monotonicity of the observed ratios or comparing different mixing distributions, which is an ill-posed question. Rather, we aim at defining an estimator which is valid under mild conditions for the whole class of binomial models, regardless of the specific mixing distribution, since the proposed approach does not suffer from weak identifiability or boundary issues which are often encountered when beta or discrete mixing distribution are considered. For a more formal use of the ratio plot, the interested reader is referred to Böhning et al. (2013). The linear function in (8) is the tangent line to the curve at $x=0$; given that $\log(\cdot)$ is a concave function of x (and $\theta_{x,B}$ is non-decreasing in x), the linear function will be above the curve, and the approximation would be satisfactory should the corresponding equation be estimated on x-points sufficiently near to 0. As much as we move away from zero, $\hat{\gamma}_1$ tends to move away from the *true* (if any) γ_1, taking higher values. Therefore, the proposed approach leads to underestimate the *true* n_0; Fig. 1 displays a ratio plot for data generated according to case B1 of Pledger (2005). As it can be easily observed, the intercept term of the tangent curve is lower than the estimate derived by the regression model, which (slightly) underestimates the true n_0. As far as the estimator bias is concerned, we must admit that a general expression for the bias cannot be derived; what we can do is to study the bias issue when a beta-binomial model holds. As it can be evinced from Eq. (7) in the paper, when a beta-binomial model holds, we use the following approximation:

$$\log\left(\frac{\theta_{x,B}}{1+\theta_{x,B}}\right) = -\log(m-1+\beta+\alpha) + \log(x+\alpha) \simeq \log\left(\frac{\alpha}{m-1+\beta+\alpha}\right) + \frac{x}{\alpha} = \gamma_1 + \delta_1 x$$

Apart from the intercept term, the bias corresponds, loosely speaking, to the accuracy of the term $\delta_1 x$ in reproducing $\log(x+\alpha)$ when $x \to 0$. The bias is increasing with $\alpha \to 0$, as it can be noticed by looking at Fig. 2, where lines correspond to linear models fitted on ratios corresponding to counts $x=1,\ldots,6$ (that is to the truncated distributions), with $\alpha \in [0.01, 3]$ and fixed $\beta > 1$. For $\alpha < 1$, the bias may reduce when also $\beta < 1$, since these values imply a higher mass at the right tail of the distribution of the binomial parameter, producing a lower regression estimate for γ. Fig. 3 below reports case 13 of the simulation study (beta mixing with $\alpha = 0.49$, $\beta = 2.76$, plot a), and a modification ($\alpha = 0.49$, $\beta = 0.76$, plot b).

To account for potential heteroscedasticity, we used a weighted OLS estimator. The weight matrix **W** is the inverse of the response covariance matrix; we assume that, conditional on fixed and observed n, counts $n_1, \ldots, n_m$ follow a multinomial distribution with cell probabilities $\tau = (\tau_1, \ldots, \tau_m)'$

$$n_1, \ldots, n_m | n \sim \text{Multinomial}(\tau)$$

where $\mathrm{E}(n_x) = n\tau_x$, $Cov(n_1, \ldots, n_m) = n[\Lambda(\tau) - \tau\tau'])$, and $\Lambda(\tau) = \mathrm{diag}(\tau)$. The proposed estimator is based on a transformation of observed frequencies, say $t(n_x)$, $x = 1, \ldots, m$; therefore, we adopt the multivariate delta method to approximate the covariance matrix for $t(n_x)$, $x = 1, \ldots, m$. When we face small population sizes, as in the simulation study, the distribution of the random vector $(n_1, \ldots, n_m)/n$ may not be close to a multivariate normal distribution and, thus, one may wonder whether the multivariate delta method lead to reliable results. Our empirical findings, based on the simulation study in Section 4, are that weighting mainly influences variability/dispersion rather than point estimates.

4. Simulation study

Pledger (2005) describes a simulation study where mixed binomial distributions built from several mixing distributions for individual-specific detection probabilities are used. The aim is to give a comprehensive assessment of bias and precision for a class of population size estimators. This scheme has been widely adopted in the literature; for this reason we used it as well to provide a detailed comparison with other, well-known, estimators. The study is based on 1000 samples drawn from different mixed binomial distributions with population size $N = 100$ and $m = 6$ sampling occasions. Comparison entails estimators from homogenous (binomial) and two heterogeneous models, the beta-binomial and two components finite mixture model, referred to as BB and 2PM, respectively.

Table 1 shows the mixing distributions divided into three groups (A, B and C) with corresponding mean, variance and skewness values. In group A, both mean and heterogeneity are low and the probability of missing a unit is slightly higher than 0.4; group B has similar generating distributions, with higher mean and heterogeneity, and lower masses at zero, while group C distributions have low mean, high heterogeneity and sometimes huge masses at zero. Thus, three different scenarios are considered to analyze the behavior of the proposed estimator under different settings.

Table 2 shows the median of the proposed estimator values $\hat{N}_r(b)$, over samples $b = 1, \ldots, 1000$, and the corresponding median absolute deviations $\mathrm{MAD} = \mathrm{Med}[\hat{N}_r(b) - \mathrm{Med}(\hat{N}_r(b))]$. We also report, from Table 4 of Mao and You (2009), the median and Mad estimates for the BB model (N_{h_β}), the 2PM model (N_{h_2}), the lower bound estimators of Mao (2007b), N_ϕ, and Chao (1989), N_ψ.

When a beta mixing is used (choices A1, B1 and C1), $\hat{N}_r$ tends to slightly (significantly in C1) underestimate the population size. When observed counts come from a finite mixture model, the proposed estimator may be sensitive to larger counts, and tend to slightly overestimate the population size. This is not true for choice C2, where $\hat{N}_r$ satisfactory models the mass near to zero, leading to a moderately negative bias. When the generating distribution is clear of zero but has high skewness (choices A2, A4, B2 and B4), $\hat{N}_r$ behaves similarly to the 2PM model, and both are preferable to the BB model.

While Pledger (2005) pointed out serious underestimation for all the estimators therein considered when the beta-binomial model is used in the presence of low true skewness (choices A3, A5, B3, B5, C5 and C6), the proposed estimator

Table 1
Mixing distributions for individual capture probabilities. Mean (μ), variance (σ^2) and skewness coefficient of the mixing distributions, π=component weight, θ=component specific parameter, from Pledger (2005).

Mixing distribution	Details	μ	σ^2	Skew	P_0
Group A					
A1. Beta	$B(1.76, 9.99)$	0.15	0.010	1.02	0.448
A2. Two-point	$\pi = (0.942, 0.058)$, $\theta = (0.125, 0.552)$	0.15	0.010	3.78	0.422
A3. Two-point	$\pi = (0.5, 0.5)$, $\theta = (0.05, 0.25)$	0.15	0.010	0.00	0.457
A4. Two-point	$\pi = (0.964, 0.036)$, $\theta = (0.131, 0.669)$	0.15	0.010	5.00	0.415
A5. Four-point	$\pi = (0.4, 0.1, 0.1, 0.4)$, $\theta = (0.05, 0.1, 0.2, 0.25)$	0.15	0.009	0.00	0.445
A6. Uniform	a=0, b=0.3 on [0,b]	0.15	0.010	0.00	0.435
Group B					
B1. Beta	$B(1.31, 3.94)$	0.25	0.030	0.80	0.306
B2. Two-point	$\pi = (0.866, 0.134)$, $\theta = (0.182, 0.690)$	0.25	0.030	2.14	0.259
B3. Two-point	$\pi = (0.5, 0.5)$, $\theta = (0.077, 0.423)$	0.25	0.030	0.00	0.329
B4. Two-point	$\pi = (0.916, 0.084)$, $\theta = (0.198, 0.822)$	0.25	0.030	3.00	0.242
B5. Four-point	$\pi = (0.4, 0.1, 0.1, 0.4)$, $\theta = (0.06, 0.2, 0.3, 0.44)$	0.25	0.029	0.00	0.324
B6. Quadratic	$f_x = 85.7(x-0.4)^2$ on $(0.1, 0.6)$	0.26	0.030	1.04	0.265
Group C					
C1. Beta	$B(0.49, 2.76)$	0.15	0.030	1.54	0.550
C2. Two-point	$\pi = (0.935, 0.065)$, $\theta = (0.104, 0.807)$	0.15	0.030	3.53	0.441
C3. Exponential	$\lambda = 6$, truncated to $(0, 1]$	0.16	0.030	1.68	0.479
C4. Log	$f_x = -\log(x)$ on $(0, 1]$	0.25	0.050	0.89	0.371
C5. Beta mix	$\pi = (0.5, 0.5)$, $B(0.43, 8.08)$ and $B(9.13, 27.38)$	0.15	0.015	0.27	0.492
C6. Beta mix	$\pi = (0.5, 0.5)$, $B(0.81, 4.57)$ and $B(3.63, 6.74)$	0.25	0.030	0.48	0.315

Table 2
Simulation results. Median (Mad) values for the proposed estimator, $\hat{N}_r$, for the beta-binomial $\hat{N}_{h_\beta}$ and the 2PM $\hat{N}_2$ model, Mao (2007a,b) $\hat{N}_\phi$, and Chao (1989) $\hat{N}_\psi$ lower bounds. $N=100$, $m=6$ sources, $B=1000$ samples.

Choice	$\hat{N}_r$	$\hat{N}_{h_\beta}$	$\hat{N}_{h_2}$	$\hat{N}_\psi$	$\hat{N}_\phi$
A1	98(11.9)	95(30.0)	95(34.7)	101(43.8)	81(12.9)
A2	110(11.6)	273(275.0)	108(30.3)	114(38.9)	95(16.7)
A3	91(10.2)	81(15.9)	82(20.6)	88(30.1)	74(10.4)
A4	111(15.7)	451(541.3)	106(21.0)	109(25.5)	96(15.9)
A5	96(10.6)	86(19.2)	88(24.9)	94(33.2)	78(11.7)
A6	97(10.7)	89(19.7)	91(27.5)	96(33.5)	80(13.0)
B1	92(6.0)	98(20.1)	89(13.8)	97(26.5)	84(8.5)
B2	105(8.8)	248(193.1)	102(12.3)	109(21.5)	98(11.2)
B3	86(4.9)	83(10.3)	99(39.1)	129(84.5)	79(7.7)
B4	112(9.3)	1463(2011.6)	102(10.5)	108(15.7)	99(10.5)
B5	88(5.1)	82(10.4)	98(37.5)	130(86.1)	80(7.9)
B6	97(6.3)	114(28.0)	100(16.6)	107(26.0)	92(9.4)
C1	70(6.9)	93(51.0)	66(18.6)	80(40.0)	60(10.3)
C2	99(16.5)	∞(–)	102(22.0)	111(32.7)	94(20.6)
C3	84(8.7)	120(70.9)	78(19.3)	93(40.5)	72(11.3)
C4	82(5.8)	102(31.1)	78(11.2)	91(28.7)	76(8.0)
C5	79(7.9)	71(14.3)	74(21.3)	77(26.3)	66(9.4)
C6	85(5.4)	90(14.1)	88(15.7)	97(30.9)	82(7.6)

seems to produce more precise estimates. It is commonly acknowledged, see e.g. Dorazio and Royle (2003), that boundary problems may occur for both the BB and the 2PM models; just to give an example, as stressed also by Mao and You (2009), the former leads to infinite estimates in the 97.2% of cases when the choice C2 is entailed. The proposed estimator does not suffer from such problems; in fact, regardless of the *true* mixing distribution, it is always well defined.

While neither estimators can be considered to clearly outperform all the others, we may observe that the proposed estimator N_r always (but in a very few cases) outperforms the BB model, and, often, the 2PM model; in all analyzed conditions, $\hat{N}_r$ has lower variability when compared to 2PM or BB estimators, as can be stated by looking at corresponding MADs. Compared to Chao and Mao lower bounds, N_r seems to produce more reliable and/or less dispersed estimates, and could thus represent a potential alternative to lower bound estimators. Indeed, for the simulation reported in Table 2, $\hat{N}_r$ is a lower bound in 14 out of 18 cases, and in the remaining cases, the positive bias is not substantial, as it can be evinced by corresponding MADs. Further simulation results, discussing the performance of the proposed variance estimator, and the analysis of the bias of the regression estimator according to Mao and You (2009) results, are given in Appendix C.

5. Examples

In this section, we discuss real data examples: the golf tees data (Borchers et al., 2002) with $m=8$, the meadow voles data (Pollock et al., 1990) with $m=5$ and the Hong Kong bird data, with $m=20$ sampling occasions. Nonparametric bootstrap CI have been calculated using $S=1000$ resamples. For all the considered examples, we used the function `closedN` of the `R` library `secr`, to provide competing estimators, namely the 2-point mixture, the beta-binomial, the jackknife of Otis et al. (1978) estimators. Obtained results have been compared with those previously discussed in the literature, to give also a look at potential differences in the point or interval estimates. All confidence intervals are built considering a 95% nominal level.

5.1. Golf tees data

In a field experiment, 250 groups (760 individuals) of golf tees were placed in groups of different sizes in a survey region of 1680 m^2, either exposed above the surrounding grass, or, at least partly, hidden by it. They were surveyed by the 1999 statistics honors class at the University of St Andrews (Scotland), see Borchers et al. (2002). A total of $n=162$ groups of tees were seen, but a (potentially unknown) number is missed and needs to be estimated. Table 3 shows the corresponding frequency distribution. Fig. 4 provides a plot of the observed ratios and the estimated regression line. This is a well-known example where, due to weak identifiability, the beta-binomial model provides an unreliable estimate of the confidence interval of the population size, as we will discuss in the following. The regression estimate of the number of missed units is $\hat{n}_0^r = 54$ leading to a population size $\hat{N}_r = 216$; the confidence intervals are $[193, 238]$ and $[188, 247]$, depending on whether we use the asymptotic approximation for the variance of N_r or a nonparametric bootstrap approach. The bootstrap estimate of the standard deviation is $\hat{sd}(\hat{N}_r) = 14.17$, while the asymptotic standard error estimate is $\hat{sd}(\hat{N}_r) = 11.26$. The maximum likelihood estimate based on the beta-binomial model is $N_{h_\beta} = 302$, with confidence interval $[219, 501]$, $(\hat{sd}(\hat{N}_{h_\beta}) = 66.56)$; this is still much wider than the one for the regression estimator, but substantially narrower than the one reported in King et al. (2010), that is $[209, 7705]$. As it can be evinced by the simulation section for those cases with a substantial mass at zero and a high variability, the 2PM model clearly underestimates the true population size, with a point estimate $\hat{N}_{h_2} = 184$, and

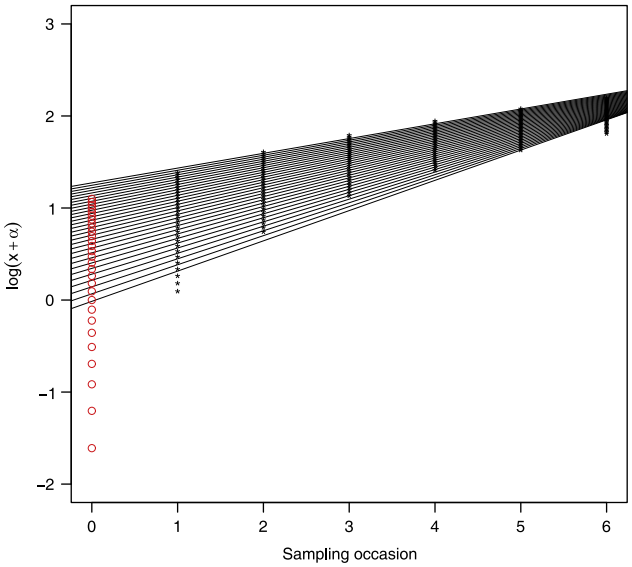

Fig. 2. Beta-binomial model. Bias of the linear estimator for $\alpha \in (0,3]$. $\log(x+\alpha)$ ('*'), estimated regression models (solid lines), and values at $x=0$ (red '○'). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

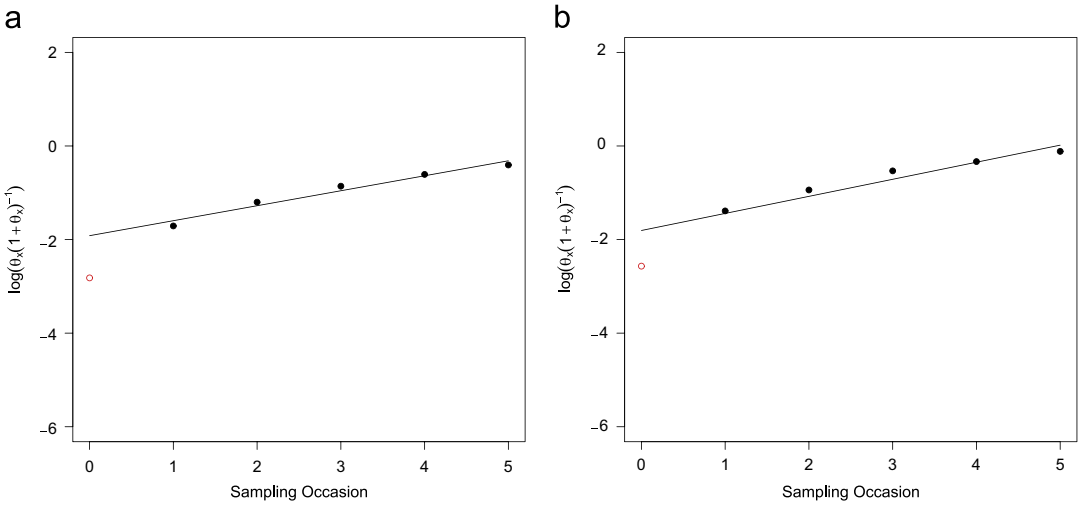

Fig. 3. Beta-binomial model. Bias of the linear estimator for $\alpha = 0.49$; (a) $\beta = 2.76$, (b) $\beta = 0.76$.

Table 3
Frequency distribution of golf tees groups detected by eight observers.

n_0	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8
–	46	28	21	13	23	14	6	11

a very narrow confidence interval [179, 191], $\hat{sd}(\hat{N}_{h_2}) = 2.97$, for further comparison see also Morgan and Ridout (2008). The jackknife estimator of Otis et al. (1978) provides a point estimate $\hat{N}_{jackk} = 213$, which is slightly more biased than the estimate obtained through the proposed regression model, with a confidence interval [190, 254] which covers the *true* population size, even if it is slightly wider than those we have shown before for the regression estimator, $\hat{sd}(\hat{N}_{jackk}) = 15.77$. Chao's lower bound estimate is given by $\hat{N}_{\psi} = 200$ with a 95% confidence interval [180, 240], $\hat{sd}(\hat{N}_{\psi}) = 14.59$, showing a slightly higher variability when compared to the log–linear estimator. As it has been mentioned before, the true population size is $N = 250$; also in this case, the regression estimator tends to produce a more reliable and efficient estimate than the ML estimates obtained by using the beta-binomial or the 2-component finite mixture models.

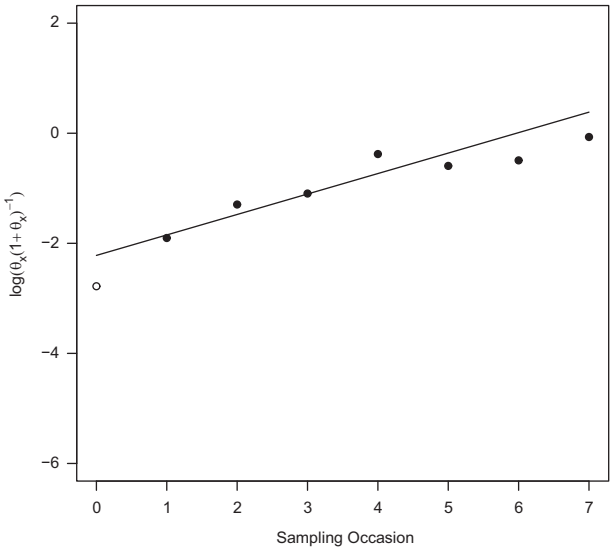

Fig. 4. Golf tees data. Plot of ratios $r_x = \log(\theta_{x,EB}/(1+\theta_{x,EB}))$ versus x. Solid line: regression estimator. Observed (filled circles) and true (empty circle) ratios.

Table 4
Frequency distribution of captured voles over $m=5$ occasions.

n_0	n_1	n_2	n_3	n_4	n_5
–	29	15	15	16	27

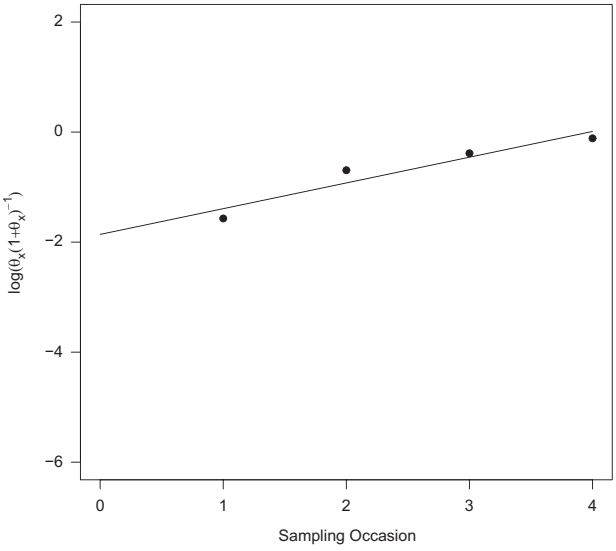

Fig. 5. Meadow voles data. Plot of ratios $r_x = \log(\theta_{x,EB}/(1+\theta_{x,EB}))$ versus x. Solid line: regression estimator.

5.2. Meadow voles data

The dataset of meadow voles has been previously analyzed by Pollock et al. (1990) and Lee and Chao (1994). There were five consecutive trapping days ($m=5$) and totally 102 distinct voles were captured. The frequency distribution of captured voles is shown in Table 4 and Fig. 5 provides a plot of the observed ratios r_x, $x=1,\ldots,4$, together with the regression line corresponding to N_r (solid line). The regression model provides an estimate $\hat{n}_0^r = 31$ and a total population size $\hat{N}_r = 133$ with 95% non-parametric bootstrap confidence interval [116, 175], and an approximated Wald-type confidence interval [113, 153]. To give a comparison term, we may notice that the value $\hat{N}_r$ is similar to the results obtained by employing the lower bound

Chao (1989) estimate, leading to an estimate $\hat{N}_{\psi} = 130$, with a confidence interval [113, 172], $\hat{sd}(\hat{N}_{\psi}) = 13.74$, which is comparable in size with the one for the regression estimator (at least in its bootstrap version). The MLe for the beta-binomial model reported by Mao and You (2009) is $\hat{N}_{h_{\beta}} = 659$, with $\hat{\alpha} = 0.045$ and $\hat{\beta} = 0.446$ obtained near the boundaries of the parameter space. By using the `closedN` function, we obtain a point estimate $\hat{N}_{h_{\beta}} = 295$, with a completely unreliable confidence interval [102, 411893], which is clearly indicating that the standard error is growing to infinity. The estimate obtained through the 2PM model is $\hat{N}_{h_2} = 117$, with (still quite narrow) confidence interval [110,129]; the jackknife estimator gives a point estimate $\hat{N}_{jackk} = 139$, with associated confidence interval [122, 171] and standard error estimate $\hat{sd}(\hat{N}_{jackk}) = 12.18$. Also in this case, the proposed estimator seems to provide a more reliable estimate when compared to the beta-binomial and the 2PM models, a reliable competitor for lower bound estimators of the population size when the general class of mixed binomial distributions is entailed, and leads to an estimated population size which is quite similar to the estimate produced by the jackknife procedure of Otis et al. (1978).

5.3. Hong Kong bird data

This dataset comes from the Hong Kong Big Bird Race (BBR), an annual competition among teams of birdwatchers. The challenge is to record as many wild bird species in the Hong Kong territory as possible during a fixed interval of time. Twenty teams ($m = 20$) competed in the Year 2000 BBR, each team had four members who went around the city to record the number of distinct bird species they observed. The frequency counts of species recorded by the 20 teams are displayed in Table 5, while Fig. 6 provides a plot of the observed ratios r_x, $x = 1, \ldots, 19$, together with the estimated regression line. In this example, the regression model leads to an estimate $\hat{n}_0^r = 9$ with $\hat{N}_r = 229$, and a non-parametric bootstrap confidence interval [223,240], with standard error $\hat{sd}(\hat{N}_r) = 4.18$, while if we use the asymptotic variance estimator the confidence interval is [221,238], with $\hat{sd}(\hat{N}_r) = 3.89$. The jackknife estimate (Otis et al., 1978) is $\hat{N}_{jackk} = 238$ with a confidence interval [228, 256], and a standard error estimate $\hat{sd}(\hat{N}_{jackk}) = 6.67$. As far as the beta-binomial model is concerned (see also Lloyd and Yip, 1991; Lloyd, 1992), the maximum likelihood estimate is 368 (s.e. 177.9); by using the `closedN` function we obtained the same point estimate, but asmaller standard error $\hat{sd}(\hat{N}_{h_{\beta}}) = 73.41$, leading to a confidence interval [279, 591]. As before, also in this case the MLe for the beta-binomial model is unusually high when compared to other estimates; the same can be argued for the corresponding standard error and the confidence interval. The 2PM model leads to a null estimate and standard error for n_0 and to an undefined confidence interval for the population size; here the model failed to converge as it

Table 5
Frequency distribution of bird species observed by 20 teams in Hong Kong, year 2000.

n_0	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}
–	21	16	13	10	4	13	6	4	11	1
	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	n_{17}	n_{18}	n_{19}	n_{20}
	6	5	8	3	4	6	11	15	8	55

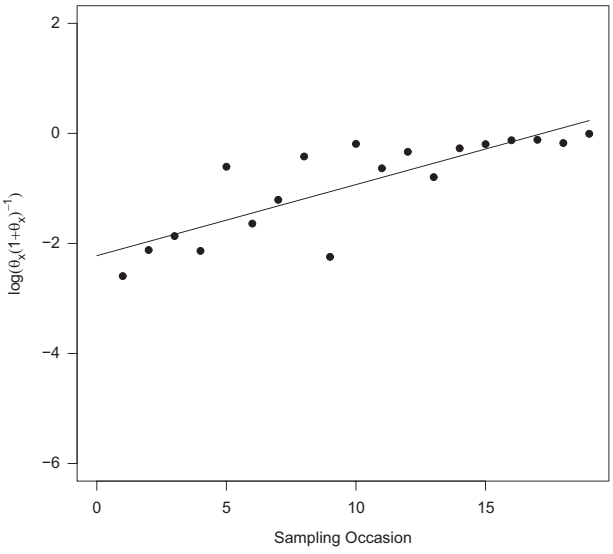

Fig. 6. Hong Kong bird data. Plot of ratios $r_x = \log(\theta_{x,EB}/(1 + \theta_{x,EB}))$ versus x. Solid line: regression estimator.

could happen and, currently, happened in a number of the simulation cases discussed by Mao and You (2009). The Chao (1989) lower bound estimate is $\hat{N}_{\psi} = 234$, with a standard error $\hat{sd}(\hat{N}_{\psi}) = 7.86$ and a confidence interval [225, 259], which is wider than the one obtained through the proposed regression estimator. In this perspective, the proposed regression estimator seems to provide a potential and reliable alternative to ML when a beta or a discrete mixing is assumed, and a reliable lower bound with a lower variability than the Chao estimator.

6. Concluding remarks

In this paper, by using an appropriate reparameterization, we extend the approach developed by Rocchetti et al. (2011) to situations where the number of sampling occasions is known and fixed. We propose an estimator based on a novel weighted regression model for (log) ratios of successive capture frequencies for the whole class of mixed binomial distributions. The proposed estimator is introduced also to prevent identifiability and boundary problems which are quite standard in ML estimation for mixed binomial models, especially when a beta mixing is considered. The observed data are used to provide estimates for the number of unregistered individuals, and, therefore, of the unknown population size, through very simple regression approach.

The beta-binomial and the proposed regression models have the same number of parameters (the regression model is, in a sense, a reparameterization of the former), but the proposed estimator does not suffer from any boundary problem, since it is neither designed to provide estimates of the beta distribution parameters, say (α, β), nor it needs to employ an Horvitz–Thompson approach to estimate N. Since observed ratios are always well-defined through continuity correction, the intercept and slope estimates are always finite, and the same is true for the estimates of n_0 and N. The proposed estimator is quite simple to be implemented in standard worksheets and do not need any complex software to be used. An asymptotic formula for the corresponding variance is given in Appendix B; the performed simulation study shows that this approximation works quite well in practice, as shown in Section 4 and Appendix C.

When we move across choices in the Pledger (2005) scheme, we have very often that $\hat{N}_r \leq N$. With increasing sample size, the relative bias is slightly decreasing (while the variability decreases more substantially), while the estimator tends to converge when the number of sampling occasions increases. When lower bound estimators are considered, we may observe that the proposed regression estimator $\hat{N}_r$ gives a satisfactory approximation to it, with good properties in terms of bias and, in particular, in terms of dispersion, which is substantially lower than those of the lower bounds of Mao (2007b) and Chao (1989). The proposed estimator has been applied to some real data, producing reliable estimates when compared to results in the literature. In general, the regression estimator could be considered a valid lower bound estimator: it produces results very close to the Chao estimator, and is characterized by a substantially lower variability. Future research should be focused on analyzing the influence, on the population size estimate, of the particular specification of the regression function we use to model the ratio plot distribution, in a sensitivity analysis perspective.

Appendix A. A monotonicity result for ratios in *m*-captures mixed binomials

Let us denote the number of successes (here, the number of times an individual is registered) with $x = 0, 1, 2, \ldots, m$. The binomial distribution can be written as

$$P(X = x|\pi, m) = \binom{m}{x}(\pi)^x(1-\pi)^{m-x} = \binom{m}{x}\left(\frac{\pi}{1-\pi}\right)^x(1-\pi)^m = \alpha_x\theta^x(1+\theta)^{-m} = \alpha_x\theta^x\mu(\theta)$$

where

$$\theta = \frac{\pi}{1-\pi}, \quad \mu(\theta) = (1+\theta)^{-m}, \quad \alpha_x = \binom{m}{x}.$$

If we consider a distribution on θ representing unobserved heterogeneity in the individual-specific probability π, the marginal pdf can be written as follows:

$$P_x = P(X = x|m) = \int \alpha_x\theta^x\mu(\theta)G(\theta)\, d\theta$$

where $G(\theta)$ represents the mixing distribution. Writing $V = \sqrt{[\mu(\theta)\theta^{x-1}]}$ and $W = \sqrt{[\mu(\theta)\theta^{x+1}]}$, and using the Cauchy–Schwarz inequality, we may write

$$[E(VW)]^2 = \{E[\mu(\theta)\theta^x]\}^2 = \left(\frac{P_x}{\alpha_x}\right)^2 \leq E\left(V^2\right)E\left(W^2\right)$$

$$= E\left(\sqrt{[\mu(\theta)\theta^{x-1}]}\right)^2 E\left(\sqrt{[\mu(\theta)\theta^{x+1}]}\right)^2 = \frac{P_{x-1}P_{x+1}}{\alpha_{x-1}\alpha_{x+1}}.$$

Table 6
Simulation results. Asymptotic and Monte Carlo (Asy.se, MC.se) estimates for the standard error of the proposed estimator, the beta-binomial $\hat{N}_{h_\beta}$ and the 2PM $\hat{N}_2$ model. Empirical coverage rates for the asymptotic and the nonparametric bootstrap confidence interval, Cov(Asy.CI) and Cov(Emp.CI). $N=100$, $m=6$ sources, $B=1000$ samples, $N_b=1000$ bootstrap resamples.

Choice	Cov(Asy.CI)	Cov(Emp.CI)	$A\overline{sy}.se(N_r)$	$MC.se(N_r)$	$\overline{se}(\hat{N}_{h_\beta})$	$\overline{se}(\hat{N}_{h_2})$
A1	0.931	0.932	16.4	17.0	59	99
A2	0.943	0.944	21.5	23.9	170	78
A3	0.875	0.888	14.8	15.1	30	79
A4	0.939	0.936	23.4	25.5	202	57
A5	0.925	0.910	16.7	17.5	22	55
A6	0.896	0.905	18.4	19.2	27	94
B1	0.817	0.807	8.5	9.1	33	29
B2	0.939	0.939	13.6	13.3	255	17
B3	0.622	0.631	7.0	7.7	13	72
B4	0.917	0.895	13.9	14.1	371	10
B5	0.708	0.721	6.9	7.9	12	56
B6	0.930	0.924	9.3	9.7	57	41
C1	0.586	0.582	10.4	11.4	101	58
C2	0.952	0.944	26.7	26.1	436	28
C3	0.759	0.763	11.5	11.9	147	57
C4	0.752	0.741	7.8	8.5	55	22
C5	0.666	0.671	11.5	12.6	22	89
C6	0.653	0.689	8.5	8.6	20	38

Thus

$$\frac{\frac{P_x}{\alpha_x}}{\frac{P_{x-1}}{\alpha_{x-1}}} \leq \frac{\frac{P_{x+1}}{\alpha_{x+1}}}{\frac{P_x}{\alpha_x}}$$

which implies the posterior mean of the odds is non-decreasing in x.

Appendix B. The formula for the asymptotic variance

Following Böhning (2008), we can use conditioning in combination with the delta-method to give an approximate expression for the variance of the proposed regression estimator. In this case, we have

$$Var(\hat{n}_0^r) \simeq Var(n_1)\left[\frac{\partial \hat{n}_0^r}{\partial n_1}\right]^2 + Var(\hat{\gamma}_1)\left[\frac{\partial \hat{n}_0^r}{\partial \hat{\gamma}_1}\right]^2 \tag{10}$$

$$Var(\hat{n}_0^r) \simeq n_1\left(1-\frac{n_1}{\hat{N}}\right)\left[\frac{1-\exp(\hat{\gamma}_1)}{\exp(\hat{\gamma}_1)m}\right]^2 + Var(\hat{\gamma}_1)\left[\frac{-n_1}{\exp(\hat{\gamma}_1)m}\right]^2 \tag{11}$$

where $Np_1(1-p_1)$ is the variance of n_1 assuming a multinomial distribution with parameter p_1 (probability of being caught once), which can be estimated by $\hat{N}(n_1/\hat{N})(1-(n_1/\hat{N})) = n_1(1-(n_1/\hat{N}))$, while $Var(\hat{\gamma}_1)$ is the variance of the intercept estimate in the regression model.

Appendix C. Further simulation results

Table 6 reports the simulation study results for standard error estimates and empirical coverage rates. In both cases, we provide values averaged over the $B=1000$ simulated samples for the asymptotic variance approximation detailed in Appendix B and Wald-type confidence intervals. To check for empirical behavior of these estimators, we also provide Monte Carlo estimates for the standard error and the empirical coverage rates obtained through nonparametric bootstrap ($\mathcal{S}=1000$ resamples). Some points are worth of a discussion; first, as far as the standard error estimates are concerned, the average asymptotic standard error approximation resembles quite accurately the Monte Carlo estimates. This means that, even with moderate population sizes, the proposed standard error estimator works accurately. When looking at empirical coverage rates, the behavior of the Wald-type CI is very similar to the one based on nonparametric bootstrap; therefore, we may use the former avoiding unnecessary computational effort. The observed coverage rates are quite good and often near to the nominal 0.95 level, but for cases B5, C1 and C3–C6, where the observed level is substantially lower than the nominal one. This is probably due to a quite pronounced curvature at $x=0$ which could not be recovered by the adopted linear approximation. In these cases, however, only a negligible portion of asymptotic CIs overestimate the *true N* value, 0.012 at maximum in all analyzed cases, and this still points out that the proposed estimator can be used as an efficient lower bound estimator.

Table 7
Simulation results. Marginal odds θ_0, sharpest lower bound $\phi(f_Q)$, the intrinsic bias (i-bias) $\phi(f_Q)-\theta(Q)$, approximation bias (a-bias) $E(\theta)-\phi(f_Q)$, and estimation bias (e-bias) $\hat{\theta}-E(\theta)$. $N=100$, $m=6$ sources, $B=1000$ samples.

Choice	θ_0	$\phi(f_Q)$	i-bias	$E(\theta_l)$	$\overline{\hat{\theta}}_l$	a-bias(θ_l)	e-bias(θ_l)
A1	0.81	0.67	−0.14	0.66	0.78	−0.01	0.12
A2	0.73	0.73	0.00	0.81	0.88	0.08	0.07
A3	0.84	0.84	0.00	0.77	0.67	−0.07	−0.10
A4	0.71	0.71	0.00	0.82	0.84	0.11	0.02
A5	0.80	0.79	−0.01	0.78	0.73	−0.01	−0.05
A6	0.77	0.67	−0.11	0.68	0.71	0.01	0.03
B1	0.44	0.31	−0.13	0.34	0.32	0.03	−0.02
B2	0.35	0.35	0.00	0.38	0.42	0.03	0.04
B3	0.49	0.49	0.00	0.38	0.31	−0.11	−0.07
B4	0.32	0.32	0.00	0.43	0.45	0.11	0.02
B5	0.48	0.47	−0.02	0.42	0.33	−0.05	−0.09
B6	0.36	0.35	−0.01	0.32	0.32	−0.03	0.00
C1	1.22	0.52	−0.70	0.53	0.51	0.01	−0.02
C2	0.79	0.94	0.15	0.85	0.91	−0.09	0.06
C3	0.92	0.54	−0.38	0.56	0.53	0.02	−0.04
C4	0.59	0.31	−0.28	0.31	0.30	0.00	−0.01
C5	0.97	0.49	−0.47	0.44	0.49	−0.05	0.05
C6	0.46	0.29	−0.18	0.26	0.28	−0.03	0.02

Since the marginal odds that an individual is unseen can not be consistently estimated, see Sanathanan (1977), we may follow Mao and You (2009) and question whether proposed estimators should be compared with the *true* population size or, rather, with what we could consistently estimate. In this case, we may distinguish between *intrinsic*, *approximation* and *estimation* bias. The first term measures the departure of the true marginal odds from what we could consistently estimate, that is the Mao (2007b) lower bound; the second term represents the ability of the proposed estimator to produce unbiased estimates of the lower bound, and is measured as the deviation between the lower bound and the expected value of the proposed estimator. The last term measures the deviation between the sample mean of the proposed estimator and its expected value and roughly measures bias due to sample variability. Table 7 summarizes the results, where columns referring to θ_0, $\phi(f_Q)$, and i-bias are drawn from Table 3 of Mao and You (2009). Quantities $E(\theta_0^r)$ for the proposed estimator have been derived by taking the expected value of the regression model predictions, while $\overline{\hat{\theta}_0^r}$ is the mean value of the marginal odds, estimated by averaging empirical ratios $\hat{\theta}_0^r$, over $B=1000$ simulation samples. As it can be easily observed, estimator $\hat{N}_r$ seems to provide a good approximation to the sharpest lower bound, with usually moderate downward bias, but for cases A2, A4, B4 where a consistent skewness is present, and some slight upward bias is observed.

To evaluate the behavior of the proposed estimator in other settings, we extended the simulation study by varying the population size, $N=500, 1000$ and the number of sampling occasions $m=9, 12$. We do not report corresponding tables here for sake of brevity. However, the main findings are that the *relative* bias is slightly decreasing with increasing population size, for a given number of sampling occasions, while the observed variability is substantially decreasing. This points out, again, the problem of identifiability stated by Link (2003). Rather, a quite clear path arises when, for a given population size, the number of sampling occasions is increased; in this case, the bias substantially reduces.

References

Bartolucci, F., Forcina, A., 2001. Analysis of capture–recapture data with a Rasch-type model allowing for conditional dependence and multidimensionality. Biometrics 57, 714–719.

Bartolucci, F., Pennoni, F., 2007. A class of latent Markov models for capture–recapture data allowing for time, heterogeneity, and behavior effects. Biometrics 63, 568–578.

Böhning, D., 2008. A simple variance formula for population size estimators by conditioning. Statistical Methodology 5, 410–423.

Böhning, D., Patilea, V., 2008. A capture–recapture approach for screening using two diagnostic tests with availability of disease status for the test-positives only. Journal of the American Statistical Association 103, 212–221.

Böhning, D., Baksh, M.F., Lerdsuwnasri, R., Gallagher, J., 2013. Use of the ratio plot in capture–recapture estimation. Journal of Computational and Graphical Statistics 22, 135–155.

Borchers, D.L., Buckland, S.T., Zucchini, W., 2002. Estimating Animal Abundance, Closed Populations. Springer-Verlag, London.

Burnham, K.P., Overton, W.S., 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrics 65, 625–633.

Burnham, K.P., Overton, W.S., 1979. Robust estimation of population size when capture probabilities vary among animals. Ecology 60, 927–936.

Chao, A., 1989. Estimating population size for sparse data in capture–recapture experiments. Biometrics 45, 427–438.

Chao, A., Tsay, P.K., Lin, S.-H., Shau, W.-Y., Chao, D.-Y., 2001. Tutorial in biostatistics. The applications of capture–recapture models to epidemiological data. Statistics in Medicine 20, 3123–3157.

Coull, B.A., Agresti, A., 1999. The use of mixed logit models to reflect heterogeneity in capture–recapture studies. Biometrics 55, 294–301.

Dorazio, R.M., Royle, J.A., 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics 59, 351–364.

Dorazio, R.M., Royle, J.A., 2005. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics 61, 874–876.

Holzmann, H., Munk, A., Zucchini, W., 2006. On identifiability in capture–recapture models. Biometrics 62, 934–939.
Hoaglin, D.C., 1980. A Poissonness plot. The American Statistician 34, 146–149.
Hoaglin, D.C., Tukey, J.W., 1985. Checking the shape of discrete distributions. In: Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), Exploring Data Tables, Trends and Shapes. Wiley, New York.
Huggins, R.M., 2002. A parametric empirical Bayes approach to the analysis of capture–recapture experiments. Australian and New Zealand Journal of Statistics 44, 55–62.
King, R., Morgan, B.J.T., Gimenez, O., Brooks, S.P., 2010. Bayesian Analysis for Population Ecology. Chapman & Hall, CRC, Boca Raton.
Lawson, C.L., Hanson, R.J., 1974. Solving Least Squares Problems. Prentice-Hall.
Lee, S.M., Chao, A., 1994. Estimating population size via sample coverage for closed capture–recapture models. Biometrics 50, 88–97.
Link, W.A., 2003. Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. Biometrics 59, 1123–1130.
Lloyd, C.J., Yip, P.S.F., 1991. A unification of inference from capture–recapture studies through martingale estimating functions. In: Godambe, V.P. (Ed.), Estimating Functions. Clarendon Press, Oxford, pp. 65–88.
Lloyd, C.J., 1992. Modified martingale estimation for recapture experiments with heterogeneous capture probabilities. Biometrika 79, 833–836.
Mao, C.X., 2007a. Estimating population sizes by catch-effort methods. Statistical Methodology 4, 111–119.
Mao, C.X., 2007b. Estimating population sizes for capture–recapture sampling with binomial mixtures. Computational Statistics and Data Analysis 51, 5211–5219.
Mao, C.X., You, N., 2009. On comparison of mixture models for closed population capture–recapture studies. Biometrics 65, 547–553.
Morgan, B.J.T., Ridout, M.S., 2008. A new mixture model for capture heterogeneity. Journal of the Royal Statistical Society C 57, 433–446.
Mooijaart, Ab., van der Heijden, Peter G.M., 1992. The EM algorithm for latent class analysis with equality constraints. Psychometrika 57, 261–269.
Norris, J.L.I., Pollock, K.H., 1996. Nonparametric MLE under two closed capture–recapture models with heterogeneity. Biometrics 52, 639–649.
Otis, D.L., Burnham, K.P., White, G.C., Anderson, D.R., 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62.
Pledger, S.A., 2000. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. Biometrics 56, 434–442.
Pledger, S.A., 2005. The performance of mixture models in heterogeneous closed population capture–recapture. Biometrics 61, 868–876.
Pollock, K.H., Nichols, J.D., Brownie, C., Hines, J.E., 1990. Statistical inference for capture–recapture experiments. Wildlife Monographs 107, 1–97.
Roberts, J.M., Brewer, D.D., 2006. Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture–recapture method. Journal of the Royal Statistical Society A 169, 745–756.
Rocchetti, I., Bunge, J., Böhning, D., 2011. Population size estimation based upon ratios of recapture probabilities. Annals of Applied Statistics 5, 1512–1533.
Sanathanan, L., 1977. Estimating the size of a truncated sample. Journal of the American Statistical Association 72, 669–672.