Received: 25 July 2019 | Accepted: 24 March 2020

DOI: 10.1111/biom.13265

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Upper bound estimators of the population size based on ordinal models for capture-recapture experiments

Marco Alfò[1] | **Dankmar Böhning**[2] | **Irene Rocchetti**[3]

[1]Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy

[2]Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

[3]Consiglio Superiore della Magistratura, Rome, Italy

Correspondence Marco Alfò, Sapienza Università di Roma, Rome, Italy.
Email: marco.alfo@uniroma1.it

Abstract

Capture-recapture studies have attracted a lot of attention over the past few decades, especially in applied disciplines where a direct estimate for the size of a population of interest is not available. Epidemiology, ecology, public health, and biodiversity are just a few examples. The estimation of the number of *unseen* units has been a challenge for theoretical statisticians, and considerable progress has been made in providing lower bound estimators for the population size. In fact, it is well known that consistent estimators for this cannot be provided in the very general case. Considering a case where capture-recapture studies are summarized by a frequency of frequencies distribution, we derive a simple upper bound of the population size based on the cumulative distribution function. We introduce two estimators of this bound, without any specific parametric assumption on the distribution of the observed frequency counts. The behavior of the proposed estimators is investigated using several benchmark datasets and a large-scale simulation experiment based on the scheme discussed by Pledger.

KEYWORDS

capture-recapture experiments, frequency of frequencies distribution, ordinal data, population size estimation, upper bound

1 | INTRODUCTION

Capture-recapture methods were originally developed in the ecological setting with the aim of estimating the unknown size of an (possibly elusive) animal population; since then, they have been gradually applied to other empirical settings, ranging from epidemiology and public health, see Böhning *et al.* (2004), to biodiversity, see Bunge *et al.* (2012), text analysis, see Efron and Thisted (1976), and software engineering, see Liu *et al.* (2015). An overview on closed population capture-recapture methods in given by Chao *et al.* (2001). For human populations, we usually observe individual records from *multiple systems*; that is, we have information in the form of an indicator variable x_{ij} that is equal to 1 if the ith unit has been recorded by the jth system, $i = 1, \dots, N$, $j = 1, \dots, m$, and equal to 0 otherwise. The goal is to estimate the size N of the population based on several, incomplete, lists of individuals recorded from that population; given the study design, the information is available only for those individuals with $x_i = \sum_{j=1}^{m} x_{ij} > 0$, that is only for individuals that have been registered by at least one source. Being identified by a list (in human/social studies) corresponds to being captured on a sampling occasion (in wildlife studies); the probability of being recorded in a list will be referred to as the capture probability. Often, the number of available lists in human studies is lower than the number of trapping/sampling occasions in wildlife studies.

When the study output is summarized by a frequency of frequencies distribution, say (x, n_x), where n_x denotes the number of units that have been recorded exactly x times, the problem of recovering the population size $N = \sum_{x \geq 0} n_x$ is equivalent to that of recovering the number of unrecorded units, n_0. By using a specific parametric model to fit the observed distribution, it is possible to derive an estimate for n_0 and, therefore, for N. However, it is well known, see Sanathanan (1977), that a consistent estimator for n_0 cannot be

derived in the general family of nonparametric mixture densities. Choosing the *best* model has been proven to be a complex task, with no general solution, as several data-generating processes can provide the same fit to the observed, truncated, distribution, see Link (2003) for a thoughtful example. As a result, lower bound estimators have attracted particular interest, see Chao (1989) and Mao (2007); in fact, it is generally acknowledged that a finite upper bound for the population size does not exist, see Mao and Lindsay (2007).

By using a simple ordinal model to represent the observed frequency distribution, however, we show that the population size may be bounded from above, regardless of the *true* data-generating process. Unfortunately, the bound depends on unknown quantities referring to the complete (untruncated) distribution; therefore, we show how this can be approximated by using the observed frequency counts and give some guidance on conditions for such an approximation to work well.

The paper is structured as follows. In Section 3, we introduce the notation, the elementary treatment of the problem, and some basic relations that can be readily established; in Section 4, an upper bound for N is provided, which can be estimated by using the approximations given in Section 5. Section 6 discusses the application of the proposed estimators to five benchmark datasets. All of these datasets have the essential feature that the population size is known; we may thus compare the *true* value of the upper bound with the proposed estimates, and evaluate the coverage, which is further investigated through a large-scale simulation exercise in Section 8. A nonstandard application is discussed in Section 7, where the distribution of an ordered categorical variable, potentially prone to measurement error, is analyzed. The paper closes with Section 9 providing some concluding remarks. The Supporting Information available at the *Biometrics* website describes the analytical and bootstrap approaches to calculate the variance for the proposed upper bound estimators; it also provides some further results for the real data and simple `R` code to reproduce our results.

2 | MOTIVATION

Before proceeding to introduce the notation and the technical contents of the paper, we would like to motivate our approach and briefly explain why a practitioner could be interested in using it. Our aim is to propose a way to estimate the *maximum* size of a hidden population; the justification for this approach stems from the empirical evidence that, in some specific cases, this count may be of major importance.

For example, let us consider the outbreak of hepatitis A virus (HAV) that occurred in and around a college in northern Taiwan from April to July 1995, see Chao (2001) for further details. To quantify the extent of this outbreak, the number of hepatitis cases was determined using an imperfect screening procedure based on merging three lists (serum test records, hospital records, and epidemiologists' records) of the infected. Twenty-eight students were present in all three lists, 56 in two lists, and 187 in only one list. Therefore, the number of infected students observed was $n = 271$. However, some students with hepatitis were missed by all three sources. In fact, a definitive serological test was later performed on all students at the college and a total of $N = 545$ students were found to have hepatitis. In other words, $n_0 = 274$ infected students were not recorded by any of the sources. The data are presented in the last row of Table 1. In this case, having an estimate of the maximum extent of the outbreak could be important from the perspective of prevention and healthcare. This may help to plan specific health policy actions and to define the size of the corresponding health-providing services. The same question may arise in the social sciences, when we consider deviant (Böhning *et al.*, 2004) or illegal (van der Heijden *et al.*, 2003; King *et al.*, 2014) behavior; in fact, the estimates we provide may give an idea of the social burden of the problem.

TABLE 1 Benchmark data

Dataset	n_0	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n
Golf	88	46	28	21	13	23	14	6	11	162
Sidney	22	8	12	16	21	12	31			100
Cottontail	59	43	16	8	6	0	2	1		76
Taxicab	137	142	81	49	7	3	1			283
Hepatitis	274	187	56	28						271

Note. Complete data distribution.

Even when the problem is that of estimating the richness of a population, the maximum number of classes/species could be of interest in itself or worth using to obtain a more precise estimate, based on both lower and upper bounds for such a quantity; see Eren *et al.* (2012) for an application to archaeology. A similar aim can also be found in software engineering, where the maximum number of defects in a software application is often of interest, see Liu *et al.* (2015) or, just to give an example, the web page Software Testing Help (2020) containing a discussion of the importance of such information in software testing. In all of these cases, having a reliable estimate of the maximum size of the population of interest could be a crucial step toward providing a timely and successful solution to the problem.

3 | ORDINAL MODELS FOR CAPTURE-RECAPTURE EXPERIMENTS

Let us start by defining some notation that will be used in the following sections. We start from a target, partially observed,

population, with the aim of estimating its size, denoted by N. For this purpose, we use m identification sources/sampling occasions that register units from the population. We may consider empirical situations where we either use the *same* mechanism repeatedly on m subsequent occasions, or several different registration sources. The mechanism(s) allow us to observe only a portion of the population. More precisely, we consider a binary indicator variable x_{ij} that equals 1 if the ith unit has been identified by the jth source/sampling occasion, and 0 otherwise. Observed units fulfill the condition $x_i = \sum_{j=1}^{m} x_{ij} > 0$, whereas the others, having $x_i = 0$, remain *unobserved*. The number of sampling occasions, m, may be known a priori, or it may be the maximum observed count. Rearranging unit indices, we may denote the global population by $x_1, \dots, x_N$ and the observed sample by $x_1, \dots, x_n$; in this way, $x_{n+1} = \cdots = x_N = 0$, without any loss of generality.

The target population can be described by the probability density function (x, τ_x), where $x = 0, 1, \dots, m$, and τ_x denotes the probability that a generic unit from the population is observed exactly x times. Clearly, the usual constraints $\tau_x \geq 0$ and $\sum_{x=0}^{m} \tau_x = 1$ apply. In the following, we will denote the observed frequency distribution by (x, n_x). The size of the observed sample is $n = \sum_{x>0} n_x$ and the corresponding distribution is truncated at zero, in the sense that units with $x = 0$ are not observed and therefore not in the sample. An obvious estimate for τ_x would be the relative frequency n_x/N that, however, cannot be computed as N is unknown. The empirical relative frequency n_x/n provides an estimate of the zero-truncated probability $\tau_x/(1 - \tau_0)$. Due to such a design, n_0 and $N = \sum_{x=0}^{m} n_x$ are both unknown, and finding an estimate for the population size N on the basis of the observed zero-truncated distribution is a special case of the general capture-recapture problem, see Bunge and Fitzpatrick (1993), Wilson and Collins (1992), Chao (2001).

In this framework, the random variable $X_i = \sum_{j=1}^{m} X_{ij}$, $X_{ij} \in \{0, 1\}$ denotes the number of captures for a given individual; the corresponding values can be considered as the categories of an ordinal random variable. As noted before, the equalities $n = \sum_{i=1}^{N} \mathbb{I}(X_i > 0)$ and $n_0 = N - n = \sum_{i=1}^{N} \mathbb{I}(X_i = 0)$ hold by definition, where $\mathbb{I}(a)$ represents the indicator function for the event a. We write the complete data cumulative distribution function as

$$\Pr(X \leq j) = \sum_{x \leq j} \tau_x = \pi_j = \exp(\theta_j)/[1 + \exp(\theta_j)],$$

where

$$\theta_j = \log\left(\frac{\pi_j}{1 - \pi_j}\right), \quad j = 1, \dots, m.$$

Clearly, we have that $\pi_j \leq \pi_{j'}$ and $\theta_j \leq \theta_{j'}$, $j < j' = 1, \dots, m$. If we consider the zero-truncated distribution, we may write

$$\Pr(X \leq j \mid X > 0) = p_j$$

and the following equality holds:

$$\begin{aligned}\pi_j &= \Pr(X \leq j) = \Pr(X = 0) + \Pr(0 < X \leq j) \\ &= \pi_0 + (1 - \pi_0)p_j.\end{aligned}$$

Solving for π_0, we obtain

$$\pi_0 = \frac{\pi_j - p_j}{1 - p_j},$$

or, rather

$$p_j = \frac{\pi_j - \pi_0}{1 - \pi_0}. \tag{1}$$

Based on the ordinal nature of X, we may write

$$\theta_j = \theta_0 + \psi_j,$$

where

$$\begin{aligned}\psi_j &= \log\left(\frac{\pi_j(1 - \pi_0)}{\pi_0(1 - \pi_j)}\right) = \log\left(\frac{\pi_j}{\pi_0}\frac{1}{(1 - p_j)}\right) \\ &= \log\left(\frac{\pi_j}{\pi_j - p_j}\right),\end{aligned}$$

and obtain the following constraints:

$$\psi_j \geq 0, \quad \psi_j \leq \psi_{j'}, \quad \text{for } j \leq j' = 1, \dots, m.$$

As $\exp(\theta_j) = \exp(\theta_0 + \psi_j) = \exp(\psi_j)\exp(\theta_0)$, the generic term of the truncated probability distribution p_j can be written as

$$\begin{aligned}p_j &= \frac{\exp(\theta_j) - \exp(\theta_j - \psi_j)}{1 + \exp(\theta_j)} = \frac{\exp(\theta_j)}{1 + \exp(\theta_j)}\left(1 - \frac{1}{\exp(\psi_j)}\right) \\ &= \pi_j(1 - \gamma_j),\end{aligned} \tag{2}$$

where the elements $\gamma_j = \exp(-\psi_j) = \frac{\exp(\theta_0)}{\theta_j}$ fulfill the constraints

$$1 \geq \gamma_j \geq \gamma_{j'}, \quad j \leq j' = 1, \dots, m.$$

We can thus write $\pi_j = \frac{p_j}{1 - \gamma_j}$ and obtain the following m equalities for π_0:

$$\pi_0 = \frac{\pi_j - p_j}{1 - p_j} = \frac{\gamma_j p_j}{(1 - \gamma_j)(1 - p_j)}, \quad j = 1, \dots, m. \tag{3}$$

Given that the truncated distribution n_j, $j = 1, \dots, m$, is observed, providing an estimate for π_0 is equivalent to providing an estimate for γ_j, for a given $j = 1, \dots, m$. However,

the only information we can derive for γ_j is that $\gamma_j \leq (1-p_j)$, $j = 1, \ldots, m$, as $\pi_0 \leq 1$ in Equation (3), and this is of limited value. In fact, as $\gamma_j \to (1-p_j)$, $\pi_0 \to 1$, leading to an uninformative infinite upper bound for the population size.

4 | AN UPPER BOUND FOR N

Our purpose is, however, to develop an upper bound for N that can be readily used in empirical applications. For this purpose, let us consider two indices j and h with $j < h = 1, \ldots, m$. We know that the following inequalities hold:

$$\begin{aligned} &\pi_h \geq \pi_j : \\ &p_h \geq p_j; \\ &\gamma_h \leq \gamma_j. \end{aligned} \tag{4}$$

Based on Equation (1), we may write

$$1 - p_j = 1 - \frac{\pi_j - \pi_0}{1-\pi_0} = \frac{1-\pi_0-\pi_j+\pi_0}{1-\pi_0} = \frac{1-\pi_j}{1-\pi_0} \quad ,$$

and therefore, from Equation (2), we have

$$\begin{aligned} \frac{1-p_j}{1-p_h} &= \frac{1-\pi_j}{1-\pi_h} = \frac{1-\frac{p_j}{1-\gamma_j}}{1-\frac{p_h}{1-\gamma_h}} = \frac{\frac{1-\gamma_j-p_j}{1-\gamma_j}}{\frac{1-\gamma_h-p_h}{1-\gamma_h}} \\ &= \frac{1-\gamma_j-p_j}{1-\gamma_h-p_h}\frac{1-\gamma_h}{1-\gamma_j} = \frac{1-\gamma_j-p_j}{1-\gamma_h-p_h}\frac{p_h}{\pi_h}\frac{\pi_j}{p_j} \\ &= \frac{1-\gamma_j-p_j}{1-\gamma_h-p_h}\frac{p_h}{p_j}\frac{\pi_j}{\pi_h} \leq \frac{1-\gamma_j-p_j}{1-\gamma_h-p_h}\frac{p_h}{p_j} \end{aligned} \tag{5}$$

as $\pi_j \leq \pi_h$. The inequality holds for $j < h = 1, \ldots, (m-1)$ as for $h = m$ we would have $p_h = \pi_h = 1$ and the ratio in Equation (5) would be infinite. Taking first and last terms of Equation (5) into account, we obtain

$$\frac{1-p_j}{1-p_h} \leq \left(\frac{1-\gamma_j-p_j}{1-\gamma_h-p_h}\right)\frac{p_h}{p_j} \tag{6}$$

or, equivalently

$$\left(\frac{1-\gamma_j-p_j}{1-\gamma_h-p_h}\right) \geq \frac{p_j(1-p_j)}{p_h(1-p_h)}. \tag{7}$$

Let us recall the definition for γ_j:

$$\begin{aligned} \gamma_j &= \frac{\exp(\theta_0)}{\exp(\theta_j)} = \frac{\exp(\theta_0)}{\exp(\theta_j)}\frac{\exp(\theta_h)}{\exp(\theta_h)} = \frac{\exp(\theta_0)}{\exp(\theta_h)}\frac{\exp(\theta_h)}{\exp(\theta_j)} \\ &= \gamma_h \frac{\pi_h(1-\pi_j)}{\pi_j(1-\pi_h)} = \gamma_h \frac{\pi_h(1-p_j)}{\pi_j(1-p_h)} = \gamma_h \frac{1}{\alpha_{j,h}}. \end{aligned}$$

This helps us restate the inequality in Equation (7) as follows:

$$\begin{aligned} (1-\gamma_j-p_j) &\geq (1-\gamma_h-p_h)\frac{p_j(1-p_j)}{p_h(1-p_h)} \\ &= (1-\gamma_j\alpha_{j,h}-p_h)\frac{p_j(1-p_j)}{p_h(1-p_h)}. \end{aligned} \tag{8}$$

Simplifying the inequality in Equation (8) gives

$$\begin{aligned} \gamma_j\left[1-\alpha_{j,h}\frac{p_j(1-p_j)}{p_h(1-p_h)}\right] &\leq (1-p_j)-(1-p_h)\frac{p_j(1-p_j)}{p_h(1-p_h)} \\ &= \left(1-p_j\right) - \frac{p_j(1-p_j)}{p_h} = \frac{\left(p_h-p_j\right)\left(1-p_j\right)}{p_h}. \end{aligned}$$

Let us recall that

$$\begin{aligned} \gamma_j\left[1-\alpha_{j,h}\frac{p_j\left(1-p_j\right)}{p_h\left(1-p_h\right)}\right] &= \gamma_j\left[1-\frac{\pi_j\left(1-p_h\right)}{\pi_h\left(1-p_j\right)}\frac{p_j\left(1-p_j\right)}{p_h\left(1-p_h\right)}\right] \\ &= \gamma_j\left[1-\frac{\pi_j}{\pi_h}\frac{p_j}{p_h}\right]. \end{aligned}$$

It is straightforward to show that the simplified term in the square brackets $(1-\frac{\pi_j}{\pi_h}\frac{p_j}{p_h})$ is positive; solving for γ_j, $j < h = 1, \ldots, (m-1)$, we obtain the *upper bound* we have been looking for:

$$\gamma_j \leq \frac{\left(p_h-p_j\right)\left(1-p_j\right)}{p_h\left[1-\frac{\pi_j}{\pi_h}\frac{p_j}{p_h}\right]} = (1-p_j)\left(\frac{p_h-p_j}{p_h-p_j\frac{\pi_j}{\pi_h}}\right) = \gamma_j^u \tag{9}$$

for any j and h such that $j < h = 1, \ldots, (m-1)$. It is worth noting that as $\pi_j \leq \pi_h$ for $j < h = 1, \ldots, (m-1)$, we have that $\gamma_j^u \leq (1-p_j)$, where the term $(1-p_j)$ defines a *trivial* bound for γ_j, as we have discussed previously. Recalling Equation (3), an upper bound is also obtained for π_0:

$$\begin{aligned} \pi_0 = \frac{\pi_j-p_j}{1-p_j} &= \frac{\gamma_j p_j}{(1-\gamma_j)(1-p_j)} \leq \frac{\gamma_j^u p_j}{(1-\gamma_j^u)(1-p_j)} \\ &= \pi_{0j}^u, \quad j = 1, \ldots, m \end{aligned} \tag{10}$$

and for the population size:

$$N_{(j)}^u = \frac{n}{1-\pi_{0j}^u} = \frac{n}{1-\left[\frac{\gamma_j^u}{1-\gamma_j^u}\frac{p_j}{1-p_j}\right]}. \tag{11}$$

5 | AN APPROXIMATION

Unfortunately, γ_j^u cannot be computed as the ratio $\frac{\pi_j}{\pi_h}$ is unknown, as is π_0. For one of the elements in the ratio, say π_h, we may use the fact that if $h = m$, $\pi_m = 1$. But, as we

have seen before, in this case γ_j^u is undefined. To get around this, we can approximate γ_j^u using the data at hand, which is the observed truncated distribution. The first approximation comes from considering that

$$\frac{p_j}{p_h} \simeq \frac{\pi_j}{\pi_h},$$

leading to

$$\gamma_j \simeq \gamma_j^\star = (1-p_j)\left(\frac{p_h - p_j}{p_h - p_j\frac{p_j}{p_h}}\right). \tag{12}$$

To get a good approximation, however, one of two conditions should be met: either π_0 is close to 0 or, if this is not the case, j and h should be large enough to ensure that $\pi_h > \pi_j \gg \pi_0$. In other words, we should look at higher categories (eg, counts). In particular, due to the problems with setting $h = m$ that we have already outlined, we suggest to set $h = m-1$ and $j = h-1 = m-2$. In this case,

$$\gamma_{(m-2)}^\star = \frac{p_{m-1}}{(p_{m-1}+p_{m-2})}(1-p_{m-2}). \tag{13}$$

Obviously, we have that

$$\frac{\pi_{(m-2)}}{\pi_{(m-1)}} = \frac{p_{(m-2)}}{1-\gamma_{(m-2)}}\frac{1-\gamma_{(m-1)}}{p_{(m-1)}} = \frac{p_{(m-2)}}{p_{(m-1)}}\frac{1-\gamma_{(m-1)}}{1-\gamma_{(m-2)}} \geq \frac{p_{(m-2)}}{p_{(m-1)}}$$

as $\gamma_{(m-2)} \geq \gamma_{(m-1)}$. The equality holds either when $\pi_0 = 0$ or $p_{(m-2)} = p_{(m-1)}$, that is $\pi_{(m-2)} = \pi_{(m-1)}$, which, however, leads to the *trivial* bound $\gamma_{(m-2)}^\star = (1-p_{(m-2)})$. According to the previous inequality, in the general case, it holds that

$$\gamma_{(m-2)}^u > \gamma_{(m-2)}^\star.$$

Using $\gamma_{(m-2)}^\star$ instead of $\gamma_{(m-2)}^u$, we obtain the following approximation for the upper bound in Equation (11):

$$\begin{aligned} N_{(m-2)}^u = N^u &= \frac{n}{1-\pi_0(m-2)} \simeq N_\star^u \\ &= \frac{n}{1-\left[\frac{\gamma_{(m-2)}^\star}{1-\gamma_{(m-2)}^\star}\frac{p_{(m-2)}}{1-p_{(m-2)}}\right]} = \frac{n}{1-\left(\frac{p_{(m-1)}}{1+p_{(m-1)}}\right)}. \end{aligned} \tag{14}$$

This estimator has a remarkably simple structure and is easy to obtain; if we look at the observed frequency distribution (x, n_x), $x = 1, \ldots, m$ and use $\widehat{p}_j = \frac{\sum_{l\leq j} n_l}{n}$, $j = 1, \ldots, m$, we get the estimate:

$$\begin{aligned} \widehat{N}_\star^u &= \frac{n}{1-\left(\frac{\widehat{p}_{(m-1)}}{1+\widehat{p}_{(m-1)}}\right)} = n(1-\widehat{p}_{(m-1)}) \\ &= n + \sum_{j=1}^{(m-1)} n_j = 2n - n_m. \end{aligned} \tag{15}$$

A sufficient condition for $N_\star^u$ to be an upper bound for N is that $\gamma_{(m-2)}^\star \geq \gamma_{(m-2)}$, that is

$$\begin{aligned} \gamma_{(m-2)}^\star &= \frac{p_{(m-1)}}{p_{(m-1)}+p_{(m-2)}}(1-p_{(m-2)}) \\ &\geq 1 - \frac{p_{(m-2)}}{\pi_{(m-2)}} = \gamma_{(m-2)}. \end{aligned} \tag{16}$$

If we solve for $\pi_{(m-2)}$, we obtain the following condition for $N_\star^u$ to be an upper bound of the true population size, N:

$$\pi_{(m-2)} \leq \frac{p_{(m-1)}+p_{(m-2)}}{p_{(m-1)}+1}. \tag{17}$$

By exploiting the dependence of $\pi_{(m-2)}$ on π_0 and $p_{(m-2)}$, we may also derive a condition on π_0 for $N_\star^u$ to be an upper bound of N:

$$\pi_0 \leq \frac{p_{(m-1)}}{1+p_{(m-1)}} \simeq 0.5. \tag{18}$$

However, this sufficient condition cannot be fulfilled in every situation, as the probability of zero counts could be higher than 0.5, and/or the number of capture occasions could be very small (eg, three), leading to a substantial difficulty in the use of $p_{(m-2)}/p_{(m-1)}$ to approximate the ratio $\pi_{(m-2)}/\pi_{(m-1)}$. More importantly, no method can be used to check (empirically) that the condition in Equation (18) holds. Therefore, we introduce a further upper bound approximate estimator, which we will refer to as N_c^u.

This is simply based on estimating π_0 from the observed distribution, plugging in the estimate to obtain an estimate for π_j, $j = 1, \ldots, m-1$, and deriving the upper bound based on this *augmented* empirical distribution.

For this purpose, we use the Chao's (1984) lower bound estimator to approximate the number of missing units n_0:

$$\widehat{n}_0^c = \frac{n_1^2}{2n_2}$$

or, equivalently

$$\widehat{\pi}_0^c = \frac{\widehat{n}_0^c}{\widehat{n}_0^c + n}.$$

Based on such an estimator, we derive a *completed* distribution $(\widehat{n}_0^c, n_1, \ldots, n_m)$ and calculate the *upper bound* for N on such a completed distribution. Clearly, a different estimator for n_0 can be used, such as those described in Lanumteang and Böhning (2011) or Rocchetti *et al.* (2014), as the procedure does not depend on the specific form of the estimator. The approach we propose leads to the following values for the cumulative probabilities of the *completed* data:

$$\widehat{\pi}_j^c = \widehat{\pi}_0^c + \left(1-\widehat{\pi}_0^c\right)p_j, \quad j = 1, \ldots, (m-1).$$

The approximation for $\gamma^u_{(m-2)}$ obviously follows:

$$\gamma^c_{(m-2)} = (1 - p_{(m-2)})\left(\frac{p_{(m-1)} - p_{(m-2)}}{p_{(m-1)} - p_{(m-2)}\frac{\widehat{\pi}^c_{(m-2)}}{\widehat{\pi}^c_{(m-1)}}}\right)$$

and the approximation to the upper bound follows from Equation (11):

$$N^u_c = N^u_{(m-2),c} = \frac{n}{1 - \left[\frac{\gamma^c_{(m-2)}}{1-\gamma^c_{(m-2)}}\frac{p_{(m-2)}}{1-p_{(m-2)}}\right]}. \tag{19}$$

As we have mentioned previously, we know that $N \leq N^u_c$ whenever $\gamma_{(m-2)} \leq \gamma^c_{(m-2)}$. Starting from this inequality, we obtain the following condition for N^u_c to be an upper bound for the true population size, N:

$$\pi_{(m-2)} \leq p_{(m-2)}\left[1 - \left(1 - p_{(m-2)}\right)\left(\frac{p_{(m-1)} - p_{(m-2)}}{p_{(m-1)} - p_{(m-1)}\frac{\widehat{\pi}^c_{(m-2)}}{\widehat{\pi}^c_{(m-1)}}}\right)\right]^{-1} \tag{20}$$

or, by doing a little algebra, the equivalent condition

$$\pi_0 \leq \frac{(p_{(m-1)} - p_{(m-2)})}{\left[\left(1 - \frac{\widehat{\pi}^c_{(m-2)}}{\widehat{\pi}^c_{(m-1)}}\right) + \left(p_{(m-1)} - p_{(m-2)}\right)\right]}. \tag{21}$$

That is, with an increasing number of capture occasions, m, and/or an increasing value of the ratio $\frac{\widehat{\pi}^c_{(m-2)}}{\widehat{\pi}^c_{(m-1)}}$, the proposed estimator gives a reliable upper bound for N.

6 | REAL DATA EXAMPLES

In this section, we provide a re-analysis of five well-known benchmark datasets. For each of these, the global population size N, as well as n_0, are known; this is the reason for selecting them, as we aim to compare the approximate upper bound estimate and the *true* population size. For each of these studies, we calculate the *exact* value for N^u, the upper bound estimate in (11), and compare it with the estimates $\widehat{N}^u_\star$ and $\widehat{N}^u_c$, also looking at the difference between $p_{(m-1)}/p_{(m-2)}$ and $\pi_{(m-1)}/\pi_{(m-2)}$. We also provide the analytical and bootstrap estimates for the standard deviation of $\widehat{n}^u_{0\star} = \widehat{N}^u_\star - n$ and $\widehat{n}^u_{0c} = \widehat{N}^u_c - n$. These are calculated using the procedure described in the Supporting Information, available at the *Biometrics* website.

6.1 | Golf Tees data

The 1999 statistics honors class at the University of St. Andrews (Scotland) participated in the following experiment, see Borchers *et al.* (2004). A set of 760 golf tees of two different colors were arranged into 250 groups of various sizes, which were placed in a survey region of 1680 m^2, either exposed above the surrounding grass, or partly hidden by it. A total of 162 groups of tees were found by the participants while $n_0 = 88$ were not. The truncated empirical distribution (see the first row of Table 1) refers to the number of times each group of tees was found, with $m = 8$ sources corresponding to eight independent observers. The approximate estimates of the upper bound for the population size are $\widehat{N}^u_\star = 313$ and $\widehat{N}^u_c = 350$; these are both reasonable values for $N = 250$, as well as good approximations for $N^u = 401$. The bootstrap estimate for the standard deviation of $\widehat{n}^u_{0\star}$ is $b.sd(\widehat{n}^u_{0\star}) = 8.9$, whereas the analytical estimate is higher, $a.sd(\widehat{n}^u_{0\star}) = 10.9$. The bootstrap estimate for the standard deviation of $\widehat{n}^u_{0c}$, $b.sd(\widehat{n}^u_{0c}) = 23.08$, is considerably higher than the former as it likely suffers from the additional variability in Chao's estimator $\widehat{n}^c_0$. We also note that the truncated and untruncated probability ratios for counts $(m-2)$ and $(m-1)$ are very close to each other (0.96 and 0.975, respectively).

6.2 | Sidney data

The data concern a screening test for bowel cancer, the fecal occult blood test: this can be used to detect the presence of a small amount of blood in the bowel motion as an indicator for cancer prior to the manifestation of clear symptoms. From 1984 onward, about 50 000 subjects were screened for bowel cancer at St Vincent's Hospital in Sydney (Australia), see Lloyd and Frommer (2004a, 2004b, 2008). Given that a single application of the test is not sufficiently accurate, the screening procedure was based on a sequence of $m = 6$ binary diagnostic tests performed on consecutive days; for each test, the presence ($x = 1$) or the absence ($x = 0$) of blood in feces was recorded. People with negative results in all six tests did not undergo further assessment and their *true* disease status remained unknown, as this can only be ascertained by carrying out a definitive diagnostic test. People with at least one positive test result had their true disease status verified by physical examination, sigmoidoscopy, and colonoscopy. A sample of $N = 122$ individuals with confirmed bowel cancer have been screened again using the same procedure and the same number of testing occasions ($m = 6$). The corresponding distribution is reported in the second row of Table 1. The approximate estimates of the upper bound for the population size, $\widehat{N}^u_\star = 169$ and $\widehat{N}^u_c = 172$, are quite close to each other and to the *true* value $N^u = 191$, see the second row

TABLE 2 Benchmark data

Dataset	N	$\widehat{N}^u_\star$	$\widehat{N}^u_c$	N^u	$b.sd(\widehat{n}^u_{0\star})$	$a.sd(\widehat{n}^u_{0\star})$	$b.sd(\widehat{n}^u_{0c})$	$\frac{p_{(m-2)}}{p_{(m-1)}}$	$\frac{\pi_{(m-2)}}{\pi_{(m-1)}}$
Golf	250	313	350	401	8.92	10.91	23.08	0.96	0.975
Sidney	122	169	172	191	6.41	7.26	14.94	0.826	0.868
Cottontail	135	151	212	210	6.03	6.99	16.41	0.973	0.985
Taxicab	420	565	691	702	12.02	13.21	34.42	0.989	0.993
Hepatitis	545	543	827	817	11.77	12.45	23.16	0.897	0.949

Note. Upper bound estimates for population size N.

of Table 2. The bootstrap estimate for the standard deviation is $b.sd(\widehat{n}^u_{0\star}) = 6.41$, whereas the analytical estimate, $a.sd(\widehat{n}^u_{0\star}) = 7.26$, is higher. As in Section 6.1, the bootstrap estimate for $\widehat{n}^u_{0c}$ is higher than the other two, $b.sd(\widehat{n}^u_{0c}) = 14.94$. The ratios of the truncated and untruncated probabilities for the counts $(m-2)$ and $(m-1)$ are equal to 0.83 and 0.87, respectively.

6.3 | Cottontail data

A live trapping study of a cottontail rabbit population with known size $N = 135$, confined within a 4-acre zone, was conducted for $m = 18$ consecutive nights by Edwards and Eberhardt (1967). This dataset has been studied extensively, starting from the seminal paper by Chao (1987). Here, the number of trapping occasions is $m = 18$ but the maximum count is $\widetilde{m} = 7$, see the third row of Table 1.

The approximate estimates of the upper bound for the population size are $\widehat{N}^u_\star = 151$ and $\widehat{N}^u_c = 212$; the latter is essentially equal to the *true* value $N^u = 210$. The bootstrap estimate of the standard deviation for $\widehat{n}^u_{0\star}$ is equal to $b.sd(\widehat{n}^u_{0\star}) = 6.03$, the analytical estimate is slightly higher, $a.sd(\widehat{n}^u_{0\star}) = 6.99$, and $b.sd(\widehat{n}^u_{0c}) = 16.41$ is, as we have seen before, much higher than both the previous estimates. The ratios of the truncated and untruncated probabilities for the counts $(m-2)$ and $(m-1)$ are equal to 0.97 and 0.98, respectively.

6.4 | Taxicab data

Carothers (1973) conducted an experiment on the taxicab population of Edinburgh, Scotland. In this example, the population has a known size $N = 420$; the experiment roughly consisted of a capture when a taxicab was sighted. The taxicab population was observed on $m = 10$ consecutive days, with observation points and times changing with each day; however, no taxis were observed more than $\widetilde{m} = 6$ times, see the fourth row of Table 1.

In this case, $N^u = 702$ is much higher than the proposed approximation $\widehat{N}^u_\star = 565$, but it is well approximated by $\widehat{N}^u_c = 691$. Both approximate estimates are, however, quite successful given that $N = 420$. The ratios of the truncated and the untruncated probabilities for the counts $(m-2)$ and $(m-1)$ are equal to two decimal places with a value of 0.99.

6.5 | Hepatitis data

Now we return to the hepatitis data introduced in Section 2. As we mentioned previously, the data refer to an outbreak of the HAV recorded in and around a college in northern Taiwan from April to July 1995, see Chao (2001). Three sources were merged and the corresponding zero-truncated distribution is reported in the last row of Table 1. To estimate the extent of the outbreak, a screening serological check for all students was conducted, so that the true population size $N = 545$ is known. The proposed approximate estimates are $\widehat{N}^u_\star = 543$ and $\widehat{N}^u_c$=827. Although the former fails in providing an upper bound for N, likely due to the small number of capture occasions considered and the condition $n_0 < n$ not being fulfilled, the latter satisfactorily approximates the true value of the bound $N^u = 817$. The ratios of the truncated and untruncated probabilities, provided in Table 1, are equal to 0.9 and 0.95, respectively.

6.6 | Summarizing the empirical findings

Based on the results we have just discussed, the approximate estimate $\widehat{N}^u_\star$ may fail in two cases: when dealing with a very low number of capture occasions and/or when the number of unseen units is greater than 50% of the population size, see the condition in Equation (18). We face this issue empirically in the hepatitis example, see Section 6.5, where only three capture occasions and a high proportion of unseen units were available, and in the simulation study, see Section 8, which includes different types of "true" data-generating processes. However, even in these extreme cases, we are still able to rely on $\widehat{N}^u_c$, with the (negligible) further effort of estimating $\widehat{n}_{0c}$ and completing the observed data.

Table 2 summarizes the results obtained by calculating the proposed upper bound for each of the five benchmark datasets we considered. We report the following quantities: $\widehat{N}^u_\star$, the

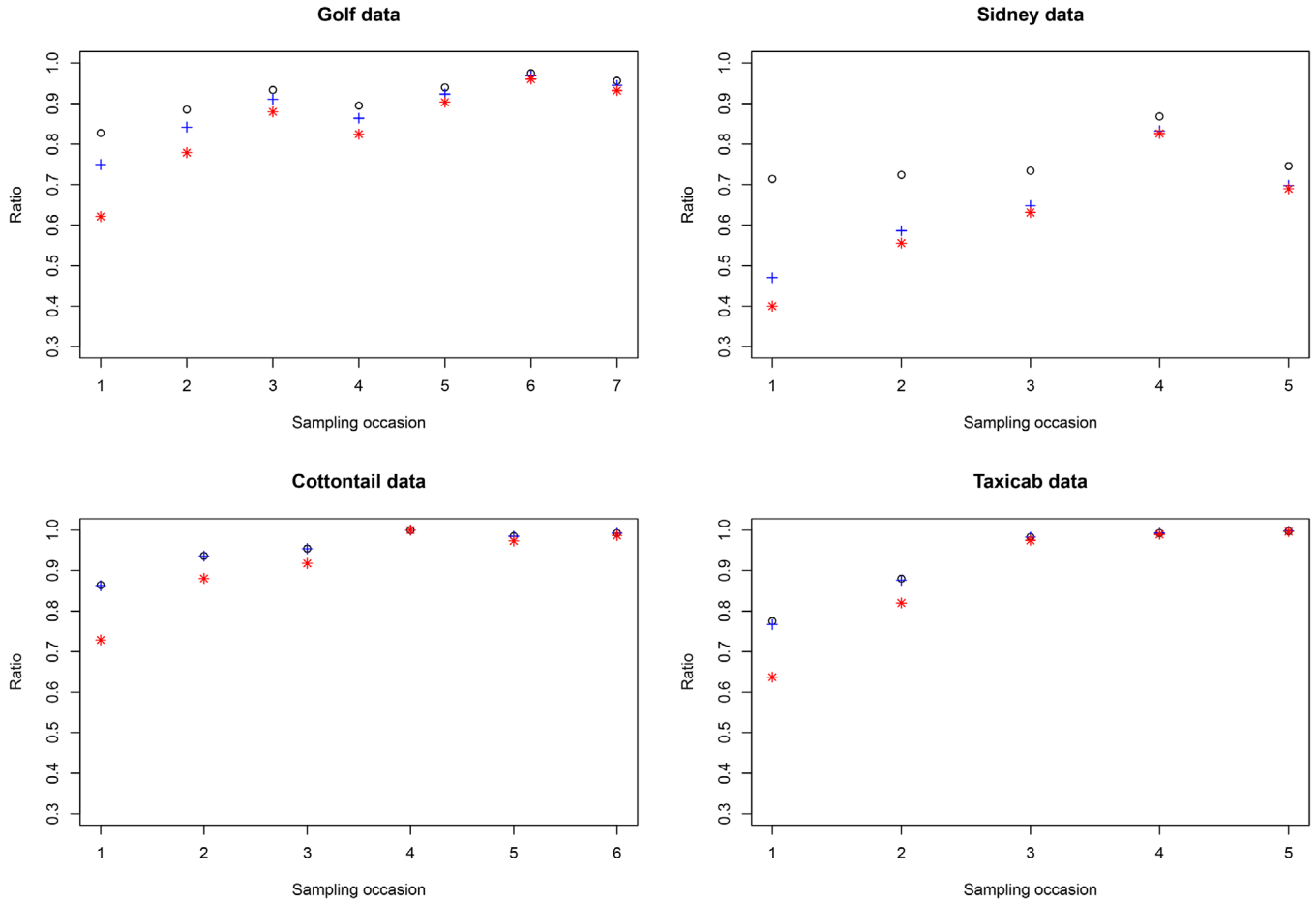

FIGURE 1 Benchmark data examples. Successive frequency ratios for the truncated (red stars), *completed* (blue pluses), and untruncated (black dots) distributions. This figure appears in color in the electronic version of the article, and any mention of color refers to that version

TABLE 3 Oman accident data

Gender	No injury	Mild	Moderate	Severe	Fatal	Number of crashes
Total	6790	13 346	9758	2226	3665	35 785
Male (%)	19.9	35.8	27.0	6.5	10.9	31 763
Female (%)	11.9	49.3	29.6	4.0	5.1	4022

Note. Complete data distribution.

approximate estimate of the upper bound for N given in Equation 11; $\widehat{N}_c^u$, the approximate estimate of the upper bound calculated by completing the observed distribution with $\widehat{n}_0^c$, the Chao estimate for n_0; N^u, the *true* value of the upper bound for N; and the ratios $p_{(m-2)}/p_{(m-1)}$ and $\pi_{(m-2)}/\pi_{(m-1)}$ to give an idea of the approximation order. We also report the estimates of the standard deviation for $\widehat{n}_{0\star}^u$ and $\widehat{n}_{0c}^u$. The former has been calculated analytically (a.) and via nonparametric bootstrap (b.), while the bootstrap approach is the only option for the latter, as detailed in Section 1 of the Supporting Information available at the *Biometrics* website. In Figure 1, we report, excluding the hepatitis data where $m = 3$, the ratios $p_{(j)}/p_{(j+1)}$ and $\pi_{(j)}/\pi_{(j+1)}$ for $j = 1, \ldots, (m-2)$, to further explore the differences we may observe, in empirical cases, between the truncated and untruncated distributions.

7 | A FURTHER, NONSTANDARD, EXAMPLE

Up to this point, we have presented the proposed upper bound in the context of a partially observed count distribution; however, it can also be fruitfully applied to truncated distributions involving ordinal categorical variables. The following example may suggest an alternative use of the proposed approach. According to Al Aamri (2018), a total of 35 785 road traffic

TABLE 4 Oman accident data

Dataset	$\widetilde{N}$	$\widehat{N}^u_\star$	$\widehat{N}^u_c$	$\widetilde{N}^u$	$b.sd(\widehat{n}^u_{0\star})$	$a.sd(\widehat{n}^u_{0\star})$	$b.sd(\widehat{n}^u_{0c})$	$\frac{p_{(m-2)}}{p_{(m-1)}}$	$\frac{\pi_{(m-2)}}{\pi_{(m-1)}}$
Crash	35 785	54 325	63 450	61 115	118.49	120.88	267.47	0.912	0.931
Crash(male)	31 763	47 440	54 969	53 753	110.27	113.07	244.35	0.906	0.927
Crash(female)	4022	6884	8535	7360	41.28	42.57	94.24	0.952	0.958

Note. Upper bound estimates for population size N.

TABLE 5 Mixing distributions for individual capture probabilities

Mixing distribution					
Group A	Details	μ	σ^2	skew	π_0
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

Note. Mean (μ), variance (σ^2) and skewness coefficient of the mixing distributions, π = component weight, θ = component-specific parameter, from Pledger (2005).

crashes were reported in Oman in 2015; 71% of them resulted in injuries, whereas 10% were fatal. Table 3 details the frequencies and percentages, stratified by gender.

As the burden of mortality and disability has considerable economic, social, and healthcare implications, it could be of interest to estimate the number of crashes resulting in *no injuries*, as the corresponding count could be severely downward biased. In fact, not all of the crashes have been reported, as those with minor consequences are likely to be underreported. That is, the corresponding frequency (n_0 in a capture-recapture setting) may be considered as observed with error; the aim here is to recover this quantity by looking at the zero-truncated (ie, truncated at the *no injury* category) distribution. We report in Table 4 the approximate estimates $\widehat{N}^u_\star$ and $\widehat{N}^u_c$, the latter obtained by completing the observed distribution with the Chao estimate $\widehat{n}^c_0$ and N^u. Further, we give the naive, likely underestimated, population size $\widetilde{N}$ and the corresponding value of the upper bound $\widetilde{N}^u$ that is likely underestimated as well. The ratios $\widehat{p}_{(m-2)}/\widehat{p}_{(m-1)}$ and $\widehat{\pi}_{(m-2)}/\widehat{\pi}_{(m-1)}$ are also reported to give an idea of the approximation order. We further provide the standard deviation estimates for $\widehat{n}^u_{0\star}$ and $\widehat{n}^u_{0c}$, the approximate estimates of the upper bound for n_0. As before, the former has been calculated both analytically (a.) and via a nonparametric bootstrap (b.) approach, whereas the latter has been calculated by nonparametric bootstrap only.

8 | SIMULATION STUDY

Pledger (2005) describes a simulation experiment concerning mixed binomial distributions, where several prior distributions for individual-specific detection probabilities are considered. The aim is to give a comprehensive assessment of bias and precision for a class of population size estimators.

TABLE 6 Simulation study

Pledger's choice	N^u	$\widehat{N}^u_\star$	$\min(\widehat{N}^u_\star)$	$\widehat{N}^u_c$	$\min(\widehat{N}^u_c)$	$sd(\widehat{N}^u_\star)$	$a.se(\widehat{N}^u_\star)$	$sd(\widehat{N}^u_c)$	f_1	f_2
A3	1545	1092	991	1354	1172	31.395	18.069	53.915	0.998	1
B1	1687	1383	1283	1568	1404	29.044	21.567	43.012	1	1
B3	1671	1346	1246	1501	1385	30.246	21.426	40.867	1	1
B5	1671	1348	1234	1492	1381	29.266	21.389	39.257	1	1
C1	1573	1300	1223	1389	1288	26.679	21.093	33.992	1	1
C4	1613	1247	1148	1414	1289	30.251	20.804	41.825	1	1
C5	1510	1022	917	1212	1051	30.973	17.737	47.371	0.761	1
C6	1678	1363	1278	1520	1399	29.261	21.445	38.847	1	1
Bin(1000, 0.3)	1880	1764	1693	1903	1812	20.268	23.601	27.935	1	1
Bin(1000, 0.7)	1885	1879	1845	1880	1846	10.216	24.009	10.246	1	1

Note. $sd(\cdot)$ and $a.se(\cdot)$ denote the standard deviation of the estimator across simulations, and the mean of the analytical standard error, respectively. All quantities are averages over simulation samples, $\min(\cdot)$ refers to the minimum over simulation samples; $f_1 = f(\widehat{N}^u_\star > N)$ and $f_2 = f(\widehat{N}^u_c > N)$ are the relative frequencies for the approximate upper bound being greater than the true population size. In all cases $N = 1000$.

The scheme is reported in Table 5. Rocchetti *et al.* (2014) have presented a point estimator for the size of the population, using this simulation scheme as a benchmark. We propose to use again the same simulation scheme and, in particular, those scenarios where the population size was more difficult to recover, and a substantial underestimation of the true population size was recorded. The study is based on $B = 1000$ replications (samples) drawn from different mixed binomial distributions with population size $N = 1000$ and $m = 6$ sampling occasions.

Table 5 shows the mixing distributions in the Pledger scheme divided into three groups: the first (A) is characterized by a low mean and heterogeneity and a probability of having a zero-count (π_0) slightly higher than 0.4. The second group (B) refers to a situation with a higher mean and heterogeneity, but a smaller mass at zero; group C is an intermediate situation between A and B, where the mixing distributions have a low mean but a high heterogeneity and a large mass at zero.

We consider one choice from the first group (A3), three choices from the second group (B1, B3, B5), and four from the third (C1, C4, C5, C6); to get an idea of the behavior of the proposed estimator under complete homogeneity conditions, we have also drawn counts from a binomial distribution with common probability of observing a generic unit equal to 0.3 and 0.7, respectively. Table 6 shows the upper bound for N that we would get if the complete distribution were known (N^u), the mean (across simulations) of the proposed approximate estimates ($\widehat{N}^u_\star$ and $\widehat{N}^u_c$), the corresponding minimum values obtained in the 1000 replicates ($\min(\widehat{N}^u_\star)$ and $\min(\widehat{N}^u_c)$), the standard deviation of the estimates across simulations ($sd(\widehat{N}^u_\star)$ and $sd(\widehat{N}^u_c)$), the mean of the analytical standard error ($se.a(\widehat{N}^u_\star)$), the true n_0, and the mean estimates $\widehat{n}^u_\star$ and $\widehat{n}^u_c$. Finally, we report the observed *coverage frequency* across simulations for $\widehat{N}^u_\star$ and $\widehat{N}^u_c$, computed as the proportion of samples for which the approximate estimates of the upper bound are greater than the true population size N.

Looking at the results in Table 6, we notice that the proposed approximate estimator of the upper bound $\widehat{N}^u_\star$ always performs well in both the heterogeneous and homogeneous settings: its minimum value across replications is always greater than the true N except for choices A3 and C5, where the *coverage frequency* is slightly less than 1 (0.998 and 0.761, respectively). As a matter of fact, these choices (A3 and C5) correspond to a very high percentage of unseen units, leading to moderate underestimation of the *true* upper bound. In all cases, the estimator $\widehat{N}^u_c$ gives a reliable approximation to N^u and its coverage is always equal to 1; consequently, it may be used as a reliable estimator for the upper bound, regardless of the *true* data generating distribution.

9 | CONCLUDING REMARKS

Capture-recapture studies have been frequently used to estimate the size of a partially observed population; several approaches based either on semiparametric, for example, finite mixtures, see Pledger (2005), or parametric, for example, negative or beta binomial, see Rocchetti *et al.* (2011), count distribution models have been used in such a context. Starting by considering a frequency of frequencies distribution and simply exploiting the ordinal nature of the observed counts, we develop a potentially useful upper bound for the population size, which seems to be a novel and undeveloped area in the capture-recapture literature as the obvious upper bound is usually infinite. As the upper bound depends on unobservable terms, we propose two approximate estimators; we study their behavior by using a series of benchmark datasets and a simulation study based on the scheme by Pledger (2005), see Section 8. Across the variety of settings we have discussed, the performance of the proposed

estimators of the upper bound is reasonably successful, as at least one of the two gives a reliable upper bound for the unknown population size. From the results of the simulation experiment, we observe that $\hat{N}^u_\star$ is always very close to or larger than the true population size, and in the majority of the analyzed samples it produces a bound that is larger than the population size. When this is not the case we may rely on $\hat{N}^u_c$, which always closely resembles the *true* upper bound for the population at hand. There is a further argument as to why the proposed ordinal approach is appropriate in a capture-recapture setting. In fact, conventionally, only the lower counts in the observed truncated distribution are used to get an estimate for the unknown population size. This is motivated by the higher counts being more sensitive to outliers and containing less information. In the proposed ordinal approach, however, we use the cumulative distribution function and, therefore, make use of the higher counts in a more robust framework. An additional reason to consider the proposed approximate upper bound is that it can be used in a Bayesian setting to define a more detailed prior for N, see, for example, Farcomeni and Tardella (2010), King *et al.* (2014), Alunni Fegatelli *et al.* (2017). One possible route to develop such an approach is based on following the advice given by Garthwaite *et al.* (2005), who discuss the *elicitation* of prior distributions based on expert knowledge.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/y8vtk.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available in the Supporting Information of this article.

ORCID

Marco Alfò https://orcid.org/0000-0001-7651-6052
Dankmar Böhning
https://orcid.org/0000-0003-0638-7106

REFERENCES

Al Aamri, A.K. (2018) *Quantifying the patterns of road traffic crashes in the sultanate of Oman: Statistical evaluation of aggregate data from police records*. PhD Thesis, University of Southampton.

Alunni Fegatelli, D., Farcomeni, A. and Tardella, L. (2017) Bayesian population size estimation with censored counts. In: Böhning, D., van der Hejiden, P.G.M. and Bunge, J. (Eds.) *Capture-Recapture Methods for the Social and Medical Sciences*. New York, NY: Chapman and Hall/CRC. Chapter 25, 371–385.

Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W. and Viwatwongkasem, C. (2004) Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *European Journal of Epidemiology*, 19, 1075–1083.

Borchers, D.L., Buckland, S.T. and Zucchini, W. (2004) *Estimating Animal Abundance: Closed Populations*. London: Springer.

Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: a review. *Journal of the American Statistical Association*, 88, 364–373.

Bunge, J., Woodard, L., Böhning, D., Foster, J.A., Connolly, S. and Allen, H.K. (2012) Estimating population diversity with CatchAll. *Bioinformatics*, 28, 1045–1047.

Carothers, A.D. (1973) Capture–recapture methods applied to a population with unknown parameters. *Journal of Animal Ecology*, 42, 125–146.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11, 265–270.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43, 783–791.

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

Chao, A. (2001) An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6, 158–175.

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

Edwards, W.R. and Eberhardt, L.L. (1967) Estimating cottontail abundance from live trapping data. *Journal of Wildlife Management*, 31, 87–96.

Efron, B. and Thisted, R. (1976) Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63, 435–447.

Eren, M.I., Chao, A., Hwang, W.-H. and Colwell, R.K. (2012) Estimating the richness of a population when the maximum number of classes is fixed: a nonparametric solution to an archaeological problem. *PLoS ONE*, 7, e34179.

Farcomeni, A. and Tardella, L. (2010) Reference Bayesian methods for recapture models with heterogeneity. *Test*, 19, 187–208.

Garthwaite, P.H., Kadane, J.B. and O'Hagan, A. (2005) Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680–701.

van der Heijden, P.G.M., Cruyff, M.J.L.F., van Howelingen, H.C. (2003) Estimating the size of a criminal population from police records using the truncated Poisson model. *Statistica Neerlandica*, 57, 1–16.

King, R., Bird, S.M., Overstall, A.M., Hay, G. and Hutchinson, S.J. (2014) Estimating prevalence of injecting drug users and associated heroin-related death rates in England by using regional data and incorporating prior information. *Journal of the Royal Statistical Society, Series A*, 177, 209–236.

Lanumteang, K. and Böhning, D. (2011) An extension of Chao's estimator of population size based on the first three capture frequency counts. *Computational Statistics and Data Analysis*, 55, 2302–2311.

Link, W.A. (2003) Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. *Biometrics*, 59, 1123–1130.

Liu, G., Rong, G., Zhang, H. and Shan, Q. (2015) The adoption of capture-recapture in software engineering: a systematic literature review. *EASE '15 Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, pp. 1–13.

Lloyd, C.J. and Frommer, D. (2004a) Estimating the false negative fraction for a multiple screening test for bowel cancer when negatives are

not verified. *Australian and New Zealand Journal of Statistics*, 46, 531–542.

Lloyd, C.J. and Frommer, D. (2004b) Regression based estimation of the false negative fraction when multiple negatives are unverified. *Journal of the Royal Statistical Society, Series C*, 53, 619–631.

Lloyd, C.J. and Frommer, D. (2008) An application of multinomial logistic regression to estimating performance of a multiple-screening test with incomplete verification. *Journal of the Royal Statistical Society, Series C*, 57, 89–102.

Mao, C.X. (2007) Estimating population sizes for capture–recapture sampling with binomial mixtures. *Computational Statistics and Data Analysis*, 51, 5211–5219.

Mao, C.X. and Lindsay, B.G. (2007) Estimating the number of classes. *Annals of Statistics*, 35, 917–930.

Pledger, S.A. (2005) The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics*, 61, 868–876.

Rocchetti, I., Bunge, J. and Böhning, D. (2011) Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics*, 5, 1512–1533.

Rocchetti, I., Alfò, M. and Böhning, D. (2014) A regression estimator for mixed binomial capture-recapture data. *Journal of Statistical Planning and Inference*, 145, 165–178.

Sanathanan, L. (1977) Estimating the size of a truncated sample. *Journal of the American Statistical Association*, 72, 669–672.

Software Testing Help. Available at: https://www.softwaretestinghelp.com/tips-to-find-valid-defects-in-any-application/ [Accessed 27 January 2020].

Wilson, R.M. and Collins, M.F. (1992) Capture-recapture estimation with samples of size one using frequency data. *Biometrika*, 79, 543–553.

SUPPORTING INFORMATION

Web appendices and Tables referenced in Sections 1 and 6 are available with this paper at the Biometrics website on Wiley Online Library. A simple R code and the analyzed data are given to perform the estimation and the standard error calculation for the approximate estimators of the upper bound

How to cite this article: Alfò M, Böhning D, Rocchetti I. Upper bound estimators of the population size based on ordinal models for capture-recapture experiments. *Biometrics*. 2020;1–12. https://doi.org/10.1111/biom.13265