



# Fixed-length interval estimation of population sizes: sequential adaptive Monte Carlo mark–recapture–mark sampling

Ivair R. Silva<sup>1</sup> · Debanjan Bhattacharjee<sup>2</sup> · Yan Zhuang<sup>3</sup>

Received: 26 May 2022 / Revised: 6 April 2023 / Accepted: 25 April 2023 /  
Published online: 15 May 2023

© The Author(s) under exclusive licence to Sociedade Brasileira de Matemática Aplicada e Computacional 2023

## Abstract

Mark–recapture sampling schemes are conventional approaches for population size ( $N$ ) estimation. In this paper, we mainly focus on providing fixed-length confidence interval estimation methodologies for  $N$  under a mark–recapture–mark sampling scheme, where, during the resampling phase, non-marked items are marked before they are released back in the population. Using a Monte Carlo method, the interval estimates for  $N$  are obtained through a purely sequential procedure with an adaptive stopping rule. Such an adaptive decision criterion enables the user to “learn” with the subsequent marked and newly tagged items. The method is then compared with a recently developed accelerated sequential procedure in terms of coverage probability and expected number of captured items during the resampling stage. To illustrate, we explain how the proposed procedure could be applied to estimate the number of infected COVID-19 individuals in a near-closed population. In addition, we present a numeric application inspired on the problem of estimating the population size of endangered monkeys of the Atlantic forest in Brazil.

**Keywords** Markov chain · Population size estimation · Fixed-length confidence intervals · COVID-19

---

Communicated by Clémentine Prieur.

---

The first author has received support from Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq, Grant #302882/2022-7, and from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), under Grant Nos. #PQ-00787-21 and #RED00133-21, Brazil. The authors would like to thank khiem Phung and Lindsay Salvati for their contribution to the literature review of mark–recapture and multiple mark–recapture through Summer Science Research Institute 2020 at Connecticut College.

---

✉ Ivair R. Silva  
ivaiarest@gmail.com

<sup>1</sup> Present Address: Department of Computer Science, Federal University of Ouro Preto, Ouro Preto, MG, Brazil

<sup>2</sup> Utah Valley University, Orem, USA

<sup>3</sup> Connecticut College, New London, USA

**Mathematics Subject Classification** 62L05 · 62L15 · 65C05

## 1 Introduction

When it comes to ecological studies and biodiversity conservation, it is important to know or estimate the population size of the concerned species. However, determining population size is not an easy task. This could be due to the difficulty of catching or the impracticality of enumerating a whole natural population of a species. One is referred to Creel et al. (2003) and Grimm et al. (2014) for more information. Way back in history, Laplace (1786) attempted to estimate the population of France using the registered births for the whole population, the number of parishes of the known total population size, and the number of births in those parishes. Applications in estimating the size of population of American duck and size of population of smolts can be found in Lincoln (1930) and Schwarz and Dempson (1994), respectively. Estimating birth/death rates and extent of registration was discussed in Sekar and Deming (1949). Given a large data set of diabetic persons consisting of a list obtained from a survey and a list obtained from doctors' records, Seber et al. (2000) applied this method to estimate both the probabilities of making list errors and the population size.

This article is on a modification of the the well-known *mark–recapture* tool which also sometimes referred as the *capture–mark–release–recapture* (CMRR) methodology. Typically *mark–recapture* is carried out in three phases: first, one catches a random sample of size  $t$ , without replacement, from the entire population of  $N$  (unknown) individuals; then, these  $t$  individuals are tagged and released back to the population; in the third phase, individuals are recaptured and the number ( $s$ ) of tagged individuals is recorded to get the final estimate of  $N$ . The first use of *mark–recapture* is often attributed to Peterson (1896). In Villella et al (2004), one would find use of mark–recapture techniques to estimate survival and recruitment in a freshwater mussel population at a site on the Cacapon River, a tributary to the Potomac River. Lincoln–Peterson estimates from mark–recapture strategy were used by Moore et al. (2010) for estimating census for tuatara, a burrowing Reptile. It is important to note that the 'mark–recapture' method is under the assumption that the samples experience no change in the population, such as birth, death, or migration. In other words, the population is considered 'closed', according to Seber and Schofield (2019).

The marking process is not just limited to two samples but can be repeated furthermore to multi-record systems. For instance, when a third sample is constructed, the unmarked subjects will be marked and the already marked could be marked again using a different marker or number if numbered tags are applied. As a result, a capture history is created for the population. This method is called multiple capture–recapture (MCR) method by Seber and Schofield (2019) or the K-sample method. Schnabel (1938) was recognized as the first to apply it to a closed population in a lake to estimate the total number of fish. The theory of this model was further developed by Chapman (1952) and Darroch (1958) developed a sequential Bayes algorithm in population estimation from multiple mark–recapture experiments, which was discussed to work for all cases regardless of the sample size, number of sequences, and sampling procedures. Smith (1988) presented Bayesian methods for obtaining point and interval estimates from data gathered from capture–recapture surveys. Smout et al. (2010) implemented an integrated approach where they simultaneously analyzed data obtained using three different marking techniques (permanent brands, tags, and natural pelage markings), assuming that animals can be cross-classified across the different mark types. They also provided estimates of mark-loss probabilities for tag and pelage individuals by essentially

comparing the capture histories with those for the branded seals. Bonner and Holmberg (2013) provided a method of modeling a population's dynamics using data from multiple, non-invasive marks by constructing an explicit model of the observation process that allows for multiple marks and applying Bayesian methods of inference via Markov chain Monte Carlo sampling. They modeled the aggregation of whale sharks in Ningaloo Marine Park, off the west coast of Australia. Epopa et al. (2017) estimated the seasonal variation in population size and the survival and dispersal of male *Anopheles coluzzii* mosquitoes in a small sub-Saharan Africa village via repeated mark–recapture experiments.

However, there is not much literature focusing on estimating a population size while providing an adequate stopping rule using the mark–recapture method. As for real-life problems, it is often the case that people need to balance between of cost of sampling and the accuracy of estimation. Thus, it is of great importance to make the best estimation with the minimum required sample. In the framework of a mark–recapture strategy, Mukhopadhyay and Bhattacharjee (2018) proposed sequential methodologies to come up with bounded risk point estimators of an optimal choice of recaptured number of tagged individuals, which led to an appropriate sequential estimator of  $N$ . Zhuang and Bhattacharjee (2021) further developed purely sequential and accelerated sequential sampling schemes to obtain the total number of tagged individuals to observe in the recapturing phase, incorporating the cost of each observation into a weighted squared-error loss function and with the goal of achieving minimum risk for estimating the size of a population.

Many researchers apply sequential analysis to solve a wide variety of problems in areas, such as reliability and quality control, clinical trials, and life tests, with the fact that sequential analysis is more efficient than a customary fixed-sample-size method and it often requires fewer sampling inspections with accurate outcome. Moreover, for some problems, sequential analysis is essential, because no fixed-sample-size procedure may work, according to Dantzig (1940). A comprehensive review of sequential methodologies can be found in Mukhopadhyay and de Silva (2009). In the spirit of Anscombe (1952) and Chow and Robbins (1965), fixed-width confidence interval for a parameter  $\theta$  is defined as  $(\theta - d, \theta + d)$  with  $2d$  to be the prefixed width of the constructed interval. Then, due to the fact that for a positive parameter  $\theta$ , the lower bound of a fixed-width confidence interval,  $\theta - d$  may possibly be a negative number, in the spirit of Banerjee and Mukhopadhyay (2016), it may be more appropriate to construct a fixed-accuracy confidence interval typically given by  $(\theta/\delta, \theta\delta)$  with  $\delta$  to be the prefixed accuracy measurement. For most recent work in sequential analysis and confidence interval estimation, one is referred to Chattopadhyay and Mukhopadhyay (2013), Mukhopadhyay and Zhuang (2017), Zhuang et al. (2020), Khalifeh et al. (2020), and Hu et al. (2021).

In this paper, we will provide solutions for constructing confidence interval estimations of the unknown population size,  $N$ , considering both fixed-length and traditional intervals, under a sequential  $s$ -adaptive multiple mark–recapture sampling scheme. We use a with-replacement procedure where a recaptured non-marked item will be marked before being put back into the population, named as a *mark–recapture–mark* scheme, which substantially decreases the recapturing effort by the end of the analysis. We would like to point out that Leslie and Chitty (1951) had proposed a similar mark–recapture sampling scheme to estimate population parameters where the authors are marking all the items after recapturing. They proposed repeated marking on items that are already marked. In our process, marked items are not marked again. This way, when the marking process is difficult and re-marking is potentially harmful for an organism, our method becomes more suitable. Schnabel (1938), Jackson (1939), and Bailey (1951) have used similar ideas, but they are doing multiple recaptures in groups to observe the number of tagged items without changing the number

of tagged items in the population by marking the unmarked before releasing back to the population.

The rest of the paper develops in the following sequence. For the non-fixed-length confidence interval, the exact solution is offered in the first part of Sect. 2. In Sect. 2.1, the solution for the fixed-length interval is derived using both exact and Monte Carlo approaches, where the latter is recommended in practice for computational viability. Section 3 contains the results of a simulation study comparing the proposed method with the *Practical Accelerated Sequential Procedure* introduced by Zhuang and Bhattacharjee (2021). In Sect. 4, the proposed mark–recapture–mark procedure is illustrated in two situations using simulated data. The first one is inspired by the problem of estimating the number of persons infected by the COVID-19 virus, and the second one is based on the experiment by Screnci-Ribeiro (2016) to estimate the population size of the endangered golden lion tamarin in Brazil. The paper is finished in Sect. 5 with the concluding remarks.

## 2 Mark–recapture–mark

Let  $t$  denote the number of marked items in a target population of unknown size, say  $N$ . Items are sampled, with replacement, one-by-one from the population. If a sampled item is already tagged, then it is just put back in the population. Otherwise, the item is marked before being returned to the population. This way, let  $X_n$  denote the number of marked items appearing up to the  $n$ th resampled item. Because  $X_n$  is a Markovian process, for a given population with size being  $N$ , its probability function for  $n \geq 2$  can be expressed through as a recursion using the law of total probability, as follows:

$$\begin{aligned} Pr[X_n = x; N] &= Pr[X_n = x|X_{n-1} = x - 1; N]Pr[X_{n-1} = x - 1; N] \\ &\quad + Pr[X_n = x|X_{n-1} = x; N]Pr[X_{n-1} = x; N] \\ &= Pr[X_{n-1} = x - 1; N] \frac{(t + n - x)}{N} + Pr[X_{n-1} = x; N] \frac{(N - t - n + x + 1)}{N}, \end{aligned} \tag{1}$$

where  $x = \max(0, n - N + t), \dots, n$ .

In addition, let  $Y_s$  denote the number of resampled items until  $X_{Y_s} = s$ , where  $s(\geq 1)$  is fixed in advance.

The probability of  $\{Y_s = y\}$  is given by

$$Pr[Y_s = y; N] = Pr[X_{y-1} = s - 1; N] \frac{(t + y - s)}{N}, \quad y = s, \dots, N - t + s. \tag{2}$$

For fixed  $N$ , the probability in (2) can be calculated by running the Markov chain related to (1). Using the same routine, if an actual value  $Y_s = y_0$  is observed, one can evaluate the likelihood,  $L(N|y_0)$ , through (2), that is

$$L(N|y_0) = Pr[Y_s = y_0; N]. \tag{3}$$

To find the maximum-likelihood estimator (MLE) for  $N$ , here denoted by  $\hat{N}_s(Y_s)$ , we use a trisection method, which is a modification of the classical bisection algorithm by Burden and Faires (1985) to get the MLE of  $N$ . The intuition behind this method is the following. Suppose that lower and upper bounds for the MLE are available, say  $N_1$  and  $N_4$ , respectively. Note that even the naive approach based on scanning each integer in the set  $N_1, N_1 + 1, \dots, N_4$

will find the maximum of  $L(N|y_0)$ . Naturally, to save execution time in comparison to such an exhaustive procedure, it is more efficient to divide the range  $[N_1, N_4]$  into sub-intervals to be scanned in parallel. As demonstrated by Van Der Burg (1995), the bisection method is the optimal choice in comparison to dividing the ranges in a larger number of sub-intervals at each iteration of the algorithm. And most recently, from Sathasivam and Alzaeemi (2022), the trisection method was discussed as a modification of the classical bisection method, and they stated that dividing an interval into two sections leads to slow convergence. For brevity, we will not get into too many details of the discussion on finding the optimal number of sub-intervals, because it is clear that one will find the optimal value either way and it is the difference between slower or faster convergence. The bisection algorithm is typically applied to find the roots for the gradient function in problems of optimization. Here, we are looking for a point to maximize a function where derivatives do not apply. For our problem specifically, the traditional bisection method does not apply, and hence, we consider a modification of the process. We decide to divide the range  $[N_1, N_4]$  in three parts,  $[N_1, N_2]$ ,  $[N_2, N_3]$ , and  $[N_3, N_4]$  and state the steps in details in the following.

Because the maximum point of  $L(N|y_0)$  is in one of the sub-intervals  $[N_1, N_2]$ ,  $[N_2, N_3]$ , or  $[N_3, N_4]$ , the range of integers at which the solution is confined is updated by comparing the likelihood for each of these  $N_i$  candidates. This way, the algorithm is as follows. Take the initial auxiliary tuning parameters  $N_4 = \max(y_0, y_0 - s + t, N_{naive})$  and  $a = L(N_4|y_0)$ , and update  $N_4 := N_4 + N_{naive}$ , where

$$N_{naive} = t \frac{Y_s}{s},$$

would be the MLE of  $N$  under a without-replacement sampling scheme, which is not the case in the present context. Because  $N_{naive}$  is the maximum-likelihood estimator and an unbiased estimator of  $N$  under the with-replacement mark-recapture scheme (Mukhopadhyay and Bhattacharjee 2018), and we are seeking for a  $N_4$  value greater than the maximum-likelihood estimator of  $N$ , it is intuitive to start the search for  $N_4$  among values greater than  $N_{naive}$ .

In the initial auxiliary tuning parameters described above, we do three calculations: (1)  $N_4 = \max(y_0, y_0 - s + t, N_{naive})$ , then (2)  $a = L(N_4|y_0)$ , and then (3)  $N_4 := N_4 + N_{naive}$ . Note that  $a$  is based on the old  $N_4$ . The new  $N_4$ , after calculation (2), will provide the first iteration of Step (i) below.

- Step (i) While  $L(N_4|y_0) \geq a$ , update  $a = L(N_4|y_0)$ , and at this point  $N_4 := N_4 + N_{naive}$ .
- Step (ii) Define  $N_1 = \max(y_0, y_0 - s + t)$ ,  $N_2 = N_1 + \lceil (N_4 - N_1)/3 \rceil$ , and  $N_3 = N_2 + \lfloor (N_4 - N_1)/3 \rfloor$ .
- Step (iii) For  $i = 1, 2, 3, 4$ , if  $\max_i L(N_i|y_0) = L(N_1|y_0)$ , then update  $N_4$  as  $N_4 := N_2$ . If  $\max_i L(N_i|y_0) = L(N_2|y_0)$ , then update  $N_4 := N_3$ . If  $\max_i L(N_i|y_0) = L(N_3|y_0)$ , then update  $N_1$  as  $N_1 := N_2$ . However, if  $\max_i L(N_i|y_0) = L(N_4|y_0)$ , then update  $N_1 := N_3$ .
- Step (iv) Update  $N_2 := N_1 + \lceil (N_4 - N_1)/3 \rceil$ , and  $N_3 := N_2 + \lfloor (N_4 - N_1)/3 \rfloor$ .
- Step (v) If  $N_4 - N_1 \leq 6$ , then stop the iterations and go to Step (vi) below. Otherwise, run Step (iii) again.
- Step (vi) Take  $\hat{N}_s(Y_s) = \operatorname{argmax}_{N^* \in \{N_1, N_1+1, \dots, N_4\}} L(N^*|y_0)$ .

To find the lower bound,  $N_1$ , to initiate the search, the trivial choice is the maximum between  $y_0$  and  $y_0 - s + t$ . Thus, to ensure that the algorithm converges, it is sufficient to show that step (i) always run in a finite number of iterations to deliver the upper bound,  $N_4$ . Note from (2) and (3), for fixed  $y_0$ , the likelihood converges to zero with  $N$ , and it is because  $L(N|y_0) = Pr[Y_s = y_0; N] \leq \frac{(t+y_0-s)}{N} \rightarrow 0$  as  $N \rightarrow +\infty$ . That is, there always exists

**Table 1** Expectation and standard deviation of  $\hat{N}_s(Y_s)$  calculated for  $N = 100, 200, \dots, 1000$  with  $t = s = 50, 100, 200$

$N$	$t = 50$		$t = 100$		$t = 200$	
	$E[\hat{N}_s(Y_s)]$	$sd[\hat{N}_s(Y_s)]$	$E[\hat{N}_s(Y_s)]$	$sd[\hat{N}_s(Y_s)]$	$E[\hat{N}_s(Y_s)]$	$sd[\hat{N}_s(Y_s)]$
100	100.63	8.30	100.00	0	na	na
200	200.66	20.26	200.41	11.04	200.00	0
300	300.16	33.44	300.80	20.12	300.19	9.68
400	400.23	46.31	400.22	29.17	400.03	15.92
500	499.95	59.32	500.05	38.18	500.67	22.33
600	599.87	72.68	600.47	47.50	600.46	28.69
700	699.85	85.79	700.56	56.45	700.26	35.01
800	799.69	99.04	800.54	65.60	800.23	41.24
900	899.58	112.47	900.47	74.84	900.77	48.05
1000	999.53	125.90	100.40	84.06	1000.80	53.99

an integer  $N_0$ , such that for  $N_4 > N_0$ , we have  $L(N_4|y_0) < L(N_0|y_0)$ . Therefore, step (i) is always executable in a finite number of iterations.

In step (ii), we setup the first division of the feasible interval in three sub-intervals.

Step (iii) is meant to update  $N_1$  and  $N_4$  to obtain a shorter feasible interval according to the maximum likelihood among the four initial  $N_i$  choices, namely interim maximum. Note that, if  $N_1$  promotes the interim maximum, it means that the global maximum is not in  $[N_2, N_4]$ , and thus, we have a new upper bound given by  $N_4 := N_2$ . However, if the interim maximum was obtained with  $N_2$ , it means that the global maximum is not in  $[N_3, N_4]$ , and hence, we can update  $N_4 := N_3$ . If  $N_3$  promoted the interim maximum, then the global maximum is not in  $[N_1, N_2]$ , and then, we can update  $N_1 := N_2$ . Finally, if interim maximum occurs with  $N_4$ , then the global cannot be in  $[N_1, N_3]$ , and thus, we can updated  $N_1 := N_3$ .

Step (iv) performs the division of the updated feasible interval in three new sub-intervals for each iteration of step (iii).

Regarding step (v), by evaluating the likelihood on each new sequence of  $N_i$ s, and following the updates in step (iii), these intervals are made narrower and narrower by following the maximum likelihood among these four interim candidates. Because of the ceiling and floor truncation operators for getting the integer candidate for  $N$  in steps (iii) and (iv), the method can end up with  $N_4 - N_1 = 3$ , or there may occur a difference of two units between  $N_i - N_{i-1}$  for  $i = 2, 3, 4$ , case where  $N_4 - N_1 = 6$ . If  $N_4 - N_1 > 6$ , then the algorithm will still have a different sequence of  $N_i$ 's to pass through step (iii).

Finally, in step (vi), the MLE is the argument of maximum in the sequence of remaining candidates  $N_1, N_1 + 1, \dots, N_4$ .

Let  $F_{\hat{N}_s}(n; N)$  denote the probability distribution of  $\hat{N}_s(Y_s)$ , which can be computed through (2) for any point  $N^*$  in the parameter space. Analogously, one can calculate the  $r$ th moment of  $\hat{N}_s(Y_s)$  in the following way:

$$E[\hat{N}_s^r(Y_s)] = \sum_{y=s}^{N-t+1} \hat{N}_s^r(y) Pr[Y_s = y; N].$$

For example, for  $t = s = 50, 100, 200$ , the values of  $E[\hat{N}_s(Y_s)]$  and of  $sd[\hat{N}_s(Y_s)] = \sqrt{Var[\hat{N}_s(Y_s)]}$  are shown in Table 1 for  $N = 100, 200, \dots, 1000$ .

Although small disturbs are found in the decimals due to the numeric precision of the probabilities calculated while running the Markov chain, the MLE of  $N$  is apparently unbiased. It is also evident that the variance of  $\hat{N}_s(Y_s)$  decreases with  $t$ .

Note that  $F_{\hat{N}_s}(n; N)$  is decreasing with  $N$ , since  $\hat{N}_s(Y_s)$  tends to increase with  $N$ . This way, for a given observation  $Y_s = y_0$  and an arbitrary  $\alpha \in (0, 1)$ , the  $100(1 - \alpha)\%$  confidence interval for  $N$  is given by

$$[L_\alpha^s(y_0), U_\alpha^s(y_0)], \tag{4}$$

where  $L_\alpha^s(y_0)$  is the largest  $N^*$ , such that  $1 - F_{\hat{N}_s}(\hat{N}_s(y_0); N^*) \leq \alpha/2$ , and  $U_\alpha^s(y_0)$  is the smallest  $N^*$ , such that  $F_{\hat{N}_s}(\hat{N}_s(y_0); N^*) \leq \alpha/2$ .

### 2.1 Fixed-length confidence intervals

The length of the confidence interval in (4) is random, but one may desire an interval estimator with length somehow fixed in advance. One option is the fixed-width interval of the form  $[\hat{N}_s(Y_s) - d, \hat{N}_s(Y_s) + d]$ , where  $d(> 0)$  is an user-defined precision of interest. Another possibility is the fixed-accuracy interval of the form  $[\delta^{-1}\hat{N}_s(Y_s), \delta\hat{N}_s(Y_s)]$ , where  $\delta(> 1)$  is interpreted as an user-defined interval accuracy. Hence, the sequential procedure may be extended as follows.

Denote the fixed-width (or fixed-accuracy) interval simply by  $[L^s(y_0), U^s(y_0)]$ , which shall be calculated with the information  $Y_s = y_0$ . The resampling is stopped and the interval  $[L^s(y_0), U^s(y_0)]$  is a  $(1 - \alpha)100\%$  confidence interval for  $N$  if

$$\beta_0 = \Pr[Y_s > y_0; N = L^s(y_0)] + \Pr[Y_s \leq y_0; N = U^s(y_0)] \leq \alpha. \tag{5}$$

Otherwise, the resampling shall continue until  $(s + 1)$ st marked item is observed. Let  $Y_{s+1}$  represent the total number of resampled items needed to collect  $(s + 1)$  marked ones. Define  $\Omega_0$ , the set of points in the sample space of  $Y_s$  for which (5) does not hold. Then, for an observation  $Y_{s+1} = y_1$ ,  $[L^{s+1}(y_1), U^{s+1}(y_1)]$  is a  $(1 - \alpha)100\%$  confidence interval for  $N$  if

$$\begin{aligned} \beta_1 = \Pr[Y_{s+1} > y_1, Y_0 \in \Omega_0; N = L^{s+1}(y_1)] + \Pr[Y_{s+1} \leq y_1, \\ Y_0 \in \Omega_0; N = U^{s+1}(y_1)] \leq \alpha. \end{aligned} \tag{6}$$

Otherwise, the resampling is continued until the next marked items is observed, and the stopping rule is the same as before, that is, having the set  $\Omega_1$  of points from the sample space of  $Y_{s+1}$  for which (6) does not hold, one needs to evaluate the inequality  $\beta_2 \leq \alpha$ .

Generalizing, define  $\Omega_{j-1}$ , the set formed by the points of the sample space of  $Y_{s+j-1}$  for which  $\beta_{j-1} > \alpha$ . The sequential procedure is stopped at the  $(s + j)$ th marked observation only if, for  $j = 1, 2, 3..$

$$\begin{aligned} \beta_j = Pr \left[ Y_{s+j} > y_j, \cap_{i=1}^{j-1} \{Y_i \in \Omega_i\}; N = L_d^{s+j}(y_j) \right] + \\ + Pr \left[ Y_{s+j} \leq y_j, \cap_{i=1}^{j-1} \{Y_i \in \Omega_i\}; N = U_d^{s+j}(y_j) \right] \leq \alpha. \end{aligned} \tag{7}$$

The confidence interval resulting from this sequential procedure is exact, because it is based on the comprehensive method of inverting a family of  $(1 - \alpha)$ -level two-tailed testing as described in page 181, Section 8.5, by Jennison and Turnbull (1999). The measure  $\beta_j$  is the sum of the lower tail and the upper tail distributions of  $Y_s$ , which can be solved for  $N$

to find the upper and the lower confidence limits. This way, the  $100(1 - \alpha)\%$  confidence interval for  $N$  is exact by construction.

The MLE of  $N$ , needed for calculating  $L^{s+j}(y_j)$  and  $U^{s+j}(y_j)$ , can be calculated with the steps (i)–(vi) already described. For this, the likelihood function for the adaptive procedure is given by

$$L(N|y_0, \dots, y_j) = Pr [Y_s = y_0|N] \prod_{i=1}^j q(y_i) \frac{t + y_i - s - i}{N},$$

where

$$q(y_i) = \begin{cases} \prod_{k=1}^{y_i - y_{i-1} - 1} \frac{N - (t + y_{i-1} - s + k - i)}{N} & \text{for } (y_i - y_{i-1}) > 1, \\ 1 & \text{otherwise.} \end{cases}$$

### 2.2 Monte Carlo solution

The exact solution for the fixed-length confidence interval with the mark–recapture–mark procedure is computationally intensive. Seeking a simpler approach, now, we introduce an alternative adaptive sequential stopping rule for the mark–recapture–mark scheme based on a Monte Carlo bound, say  $\hat{\beta}_j$ , for  $\beta_j$ .

Note from (7) that

$$\beta_j \leq Pr [Y_{s+j} > y_j; N = L_d^{s+j}(y_j)] + Pr [Y_{s+j} \leq y_j; N = U_d^{s+j}(y_j)]. \tag{8}$$

Therefore, if the right-hand side of the inequality in (8) is smaller than  $\alpha$ , so is  $\beta_j$ .

Let  $Y_1^{(l)}(j), Y_2^{(l)}(j), \dots, Y_m^{(l)}(j)$  denote a  $m$ -dimensional Monte Carlo sample of  $Y_{s+j}$  generated under  $N = L_d^{s+j}(y_j)$ . Likewise,  $Y_1^{(u)}(j), Y_2^{(u)}(j), \dots, Y_m^{(u)}(j)$  denote a Monte Carlo sample of  $Y_{s+j}$  generated under  $N = U_d^{s+j}(y_j)$ . This way, by the law of large numbers

$$\hat{\beta}_j = \frac{\# \{Y_k^{(l)}(j) > y_j\}}{m + 1} + \frac{1 + \# \{Y_k^{(u)}(j) \leq y_j\}}{m + 1} \xrightarrow{ae} Pr [Y_{s+j} > y_j; N = L_d^{s+j}(y_j)] + Pr [Y_{s+j} \leq y_j; N = U_d^{s+j}(y_j)], \tag{9}$$

as  $m \rightarrow \infty$ , where ‘*ae*’ means almost everywhere convergence.

Therefore, the adaptive sequential Monte Carlo procedure for a fixed-length confidence interval is stopped at the  $(s + j)$ th marked observation only if

$$\hat{\beta}_j \leq \alpha.$$

This method shall be referred simply by MRM. In MRM procedure, a total of  $Y_{s+j-1}$  items should be resampled to observe  $s + j$  marked items, where  $j = 1, 2, 3, \dots$

### 3 Comparison of MRM with the accelerated sequential method

The problem of determining the size of a finite population can also be addressed by standard sequential procedures and references of those were already provided in the Introduction. One of the main goal of the MRM strategy is to expedite the estimation process. Likewise, accelerated sequential procedure also aims at achieving a reasonable estimate quicker in



time than what would be required in a regular sequential procedure. In this section, we will consider a **mark–recapture** sampling and we will adopt the *Practical Accelerated Sequential Procedure* proposed by Zhuang and Bhattacharjee (2021) to compare with the MRM procedure.

Let us define  $p = t/N$ , where  $N$  and  $t$  are same as defined in the MRM from Sect. 2. Suppose that at the *recapturing* phase, we wish to gather a random sample of appropriate size that will afford us with exactly  $s$  observed tagged elements where  $s$  is fixed for now.

Let  $X_i$  stand for the number of *independent and identically distributed* (i.i.d.) trials required to observe the  $i$ th tagged item,  $i = 1, \dots, s$ . Then, the  $X_i$ 's are i.i.d. having the common geometric distribution, referred to as  $\text{geometric}(p)$ , with the following *probability mass function* (p.m.f.):

$$f(x; p) = q^{x-1}p, \quad x = 1, 2, \dots, \quad \text{and} \quad q = 1 - p, \quad 0 < p < 1, \quad \text{with} \\ \text{mean } \mu = 1/p \text{ and variance } \sigma^2 = q/p^2. \tag{10}$$

Consider the total number of items observed, namely

$$Y \equiv X_1 + \dots + X_s,$$

at the recapture phase. Clearly, in view of (10), we have  $E_p[Y] \equiv s\mu = s/p$  and  $\text{Var}_p[Y] \equiv s\sigma^2 = sq/p^2$ . Now, an unbiased estimator of  $1/p$  being  $\bar{X}_s = s^{-1}Y$ , the population size will be estimated unbiasedly by

$$\hat{N}_s \equiv \widehat{tp^{-1}} = t\bar{X}_s \text{ with an estimated variance } t^2s^{-1}(\bar{X}_s^2 - \bar{X}_s) \text{ for a fixed } s (\geq 1). \tag{11}$$

As  $s \rightarrow \infty$ , by *Central Limit Theorem* and (11), we have

$$\sqrt{s} \left( \hat{N}_s - N \right) \xrightarrow{d} N(0, \sigma^2), \text{ where } \sigma^2 = t^2(p^{-2} - p^{-1}). \tag{12}$$

Slutsky's theorem and (12) together will imply

$$\sqrt{s} \left( \log(\hat{N}_s) - \log(N) \right) \xrightarrow{d} N(0, \tau^2), \text{ as } s \rightarrow \infty, \text{ where } \tau^2 \equiv 1 - p. \tag{13}$$

### 3.1 Fixed-width confidence interval

Again, for a  $(1 - \alpha)100\%$  confidence interval, with fixed pre-assigned width  $2d (> 0)$ , we want to construct a confidence interval for  $N$  of the form  $J_s = [\hat{N}_s - d, \hat{N}_s + d]$ . For large  $s$ , in view of (12), we will have  $Pr_N [N \in J_s]$  approximately greater than or equal to  $(1 - \alpha)$  if

$$s \text{ is the smallest integer } \geq a^2\sigma^2/d^2 \equiv s_d^*,$$

where  $a = z_{\alpha/2}$  is the upper  $100(\alpha/2)\%$  point of  $N(0, 1)$ . Since  $s_d^*$  remains unknown, from Zhuang and Bhattacharjee (2021), we propose the following accelerated sequential stopping rule.

For a fixed number  $\rho \in (0, 1)$ , we take one observation at-a-time as needed according to the stopping rule

$$T_0 \equiv T_0(\rho) = \inf \left\{ s \geq 1 : s \geq \rho a^2 d^{-2} t^2 \left( \bar{X}_s^2 - \bar{X}_s + s^{-\gamma} \right) \right\}; \tag{14}$$

with arbitrary, but fixed  $\gamma (> \frac{1}{2})$ . Upon termination of the stopping rule in (14), we define the final terminated variable  $T$  as

$$T \equiv T(\rho) = \lfloor \rho^{-1} T_0 \rfloor + 1, \tag{15}$$

where  $\lfloor x \rfloor$  is the greatest integer less than or equal to  $x$ . After determining  $T$ , we adopt the *Practical Accelerated Sequential Procedure* proposed by Zhuang and Bhattacharjee (2021), to gather the remaining number of tagged items. According to that procedure, one has to collect  $B_T$  items in a batch and count the number of tagged items in that batch where

$$B_T = \left\lceil \frac{(T - T_0)(t\bar{X}_{T_0} + 1)}{t} \right\rceil, \tag{16}$$

with  $\lfloor x \rfloor$  being the greatest integer less than or equal to  $x$ . The final estimator of the population size will be given by

$$\hat{N}_T = t \left( \frac{B_T + \sum_{i=1}^{T_0} X_i}{t_{B_T} + T_0} \right), \tag{17}$$

where  $t_{B_T}$  is the number of tagged items observed in a batch of  $B_T$ . Observe that  $T$  is a positive integer-valued observable random variable. Based on finally gathered data,  $\{T, X_1, \dots, X_T\}$ , upon termination, we propose the fixed-width approximate  $(1 - \alpha)100\%$  confidence interval for  $N$

$$J_{T_d} = \left[ \hat{N}_T - d, \hat{N}_T + d \right]. \tag{18}$$

One should note that, for the accelerate sequential procedure given by, (14)–(18), the factor  $\rho$  accelerates the termination process. Moreover, for the second batch of sampling, one is required to take samples in a batch, instead of one-at-a-time, without checking the stopping rule (Tables 2, 3).

### 3.2 Fixed-accuracy confidence interval

The lower confidence bound  $\hat{N}_T - d$  from (18) may be negative with a positive probability. Hence, for arbitrary  $\delta (> 1)$ , we may require a fixed-accuracy confidence interval for  $N$

$$K_s = \left[ \delta^{-1} \hat{N}_s, \delta \hat{N}_s \right]. \tag{19}$$

For ‘large’  $s$ , in view of (13) and (19), we will have, approximately,  $P_N [N \in K_s] \geq 1 - \alpha$ , with arbitrary  $\alpha \in (0, 1)$ , if

$$s \text{ is the smallest integer } \geq a^2 \tau^2 / (\log \delta)^2 \equiv s_\delta^*,$$

where recall that  $a = z_{\alpha/2}$  is the upper  $100(\alpha/2)\%$  point of  $N(0, 1)$ . Since  $s_\delta^*$  remains unknown, we propose the following purely sequential stopping rule.

For  $0 < \rho < 1$ , let

$$R_0 \equiv R_0(\rho) = \inf \left\{ s \geq 1 : s \geq \rho a^2 (\log \delta)^{-2} (\hat{\tau}_s^2 + s^{-\gamma}) \right\}, \text{ where } \hat{\tau}_s^2 = 1 - \bar{X}_s^{-1}, \tag{20}$$

with arbitrary, but fixed  $\gamma (> \frac{1}{2})$ . Upon termination of the stopping rule in (20), we define the final terminated variable  $R$  as

$$R \equiv R(\rho) = \lfloor \rho^{-1} R_0 \rfloor + 1. \tag{21}$$

Again, following the *Practical Accelerated Sequential Procedure*, one should collect  $B_R$  items in a batch leading to the final estimator:

$$\widehat{N}_R = t \left( \frac{B_R + \sum_{i=1}^{R_0} X_i}{t_{B_R} + T_0} \right),$$

where

$$B_R = \left\lceil \frac{(R - R_0)(t\bar{X}_{R_0} + 1)}{t} \right\rceil,$$

and  $t_{B_R}$  is the number of tagged items observed in a batch of  $B_R$ . Observe that  $R$  is a positive integer-valued observable random variable. Based on finally gathered data,  $\{R, X_1, \dots, X_R\}$ , upon termination, we propose the fixed-accuracy approximate  $(1 - \alpha)100\%$  confidence interval for  $N$

$$J_{R\delta} = \left[ \delta^{-1}\widehat{N}_R, \delta\widehat{N}_R \right]. \tag{22}$$

### 3.3 Discussions on the comparisons

An important criterion to compare sampling strategies for confidence interval estimations of population sizes is the coverage probability. Moreover, the expected total recaptures under each procedure are also critical to understand the efficiency of the procedures. In the following discussions, we will use MRM to denote the mark–recapture–mark procedure that is developed in Sect. 2 and AS to denoted the accelerated sequential procedures discussed in Sect. 3.

We compare the simulation results for both MRM and AS with the unknown population sizes being  $N = 500, 1000$ , and initial tagged items to be  $t = 50, 100, 200$ . For the fixed-width confidence interval, we choose the fixed-width measurement,  $d$ , to be 50, 100, 200; for the fixed-accuracy confidence interval, we choose the fixed-accuracy measurement,  $\delta$ , to be 1.1, 1.2, 1.3. Moreover, for AS, we further fixed  $\rho = 0.9$  and  $\gamma = 0.7$ ; and for MRM, we further fixed  $s = 50, 100, 200$ . The coverage probabilities for Tables 4 and 5 are estimates for  $1 - \alpha$ . It is calculated as the ratio of the number of intervals which includes the true population size and the number of total intervals that are created. Here, for brevity, we only include the simulation results with  $\alpha = 0.1$ . For AS, the total number of recaptures are given by the following expressions:

$$\text{Fixed-width procedure: } Y_T = B_T + \sum_{i=1}^{T_0} X_i;$$

$$\text{Fixed-accuracy procedure: } Y_R = B_R + \sum_{i=1}^{R_0} X_i.$$

Corresponding to each replication in the simulation, one would get a value of  $Y_T$  and  $Y_R$  for fixed-width and fixed-accuracy confidence intervals, respectively. The average values of  $Y_T$  and  $Y_R$  obtained from all the simulations are estimates of the expected total recaptures.

Now, we are in a position to summarize the comparisons of MRM and AS according to the simulation studies. First of all, the MRM requires much less recaptures than the AS for most scenarios. MRM increases the number of tagged items in the population by tagging recaptured and non-tagged items before putting them back to the population, while the AS

**Table 2** Expected sample size at the recapture stage of MRM and AS for  $N = 500, 1000$  with  $\alpha = 0.1, t = 50, 100, 200,$  and  $d = 50, 100, 200$

$d$	$t = 50$		$t = 100$		$t = 200$	
	MRM	AS	MRM	AS	MRM	AS
$N = 500$						
50	370.13	2396.86	315.09	1066.93	351.44	403.47
100	203.26	500.88	263.69	242.04	351.01	95.32
200	195.04	112.61	164.24	50.25	351.48	34.49
$N = 1000$						
50	923.29	–	935.90	–	581.80	4323.29
100	563.24	4791.02	517.28	2387.25	523.47	1070.50
200	302.99	1055.55	390.46	503.71	525.87	244.69

**Table 3** Expected sample size at the recapture stage of MRM and AS for  $N = 500, 1000$  with  $\alpha = 0.1, t = 50, 100, 200,$  and  $\delta = 1.1, 1.2, 1.3$

$\delta$	$t = 50$		$t = 100$		$t = 200$	
	MRM	AS	MRM	AS	MRM	AS
$N = 500$						
1.1	387.77	2753.33	332.16	1230.81	351.50	468.80
1.2	203.03	781.94	264.40	351.43	351.41	136.97
1.3	194.67	392.89	263.64	179.12	351.23	70.69
$N = 1000$						
1.1	602.93	5793.77	549.72	2752.96	528.49	1231.66
1.2	314.71	1656.19	390.54	780.45	528.25	352.11
1.3	285.91	836.96	390.78	392.70	528.99	179.26

**Table 4** Coverage probability at the recapture stage of MRM and AS for  $N = 500, 1000$  with  $\alpha = 0.1, t = 50, 100, 200,$  and  $d = 50, 100, 200$

$d$	$t = 50$		$t = 100$		$t = 200$	
	MRM	AS	MRM	AS	MRM	AS
$N = 500$						
50	0.90	0.89	0.91	0.89	0.94	0.88
100	0.93	0.72	0.95	0.78	0.97	0.79
200	0.98	0.65	0.99	0.66	0.99	0.82
$N = 1000$						
50	0.91	–	0.90	–	0.90	0.90
100	0.90	0.84	0.92	0.88	0.93	0.89
200	0.92	0.72	0.97	0.73	0.95	0.79

procedure requires one to do recapturing one-at-a-time with replacement for the first stage without changing the number of tagged items for the whole process.

Second, the coverage probability of MRM procedure is uniformly well for all values of  $N, t,$  and  $d,$  according to Tables 4 and 5. However, the AS procedure does not show the 90% coverage probability for the fixed-width confidence intervals when  $d = 100, 200,$  as per Table 4. The confidence intervals in (18) and (22) are known to have the property of

**Table 5** Coverage probability at the recapture stage of MRM and AS for  $N = 500, 1000$  with  $\alpha = 0.1, t = 50, 100, 200$ , and  $\delta = 1.1, 1.2, 1.3$ . The coverage probability estimates are shown in parentheses

$\delta$	$t = 50$		$t = 100$		$t = 200$	
	MRM	AS	MRM	AS	MRM	AS
$N = 500$						
1.1	0.90	0.91	0.91	0.90	0.98	0.90
1.2	0.90	0.92	0.98	0.90	0.99	0.90
1.3	0.97	0.92	0.99	0.91	1.00	0.91
$N = 1000$						
1.1	0.90	0.91	0.90	0.91	0.93	0.91
1.2	0.90	0.91	0.96	0.91	0.99	0.90
1.3	0.96	0.92	0.99	0.91	1.00	0.92

asymptotic consistency in the sense that

$$\lim_{d \rightarrow 0} P[N \in J_{T_d}] = 1 - \alpha, \quad \text{and} \quad \lim_{\delta \rightarrow 1} P[N \in J_{R_\delta}] = 1 - \alpha,$$

which means that the AS procedure holds the asymptotic consistency property only when  $d$  is small (close to 0) and  $\delta$  is small (close to 1).

Third, both of the procedures are giving at least 0.90 coverage probabilities for the chosen values of  $\delta$  in case of fixed-accuracy confidence intervals. However, the MRM outperforms the AS in many cases as is shown in Table 5.

As one last note, the AS procedure would perform well when one looks for confidence intervals that have high coverage probability but small width. One may note that the simulation results are missing when  $N = 1000, d = 50$ , and  $t = 50, 100$  in Tables 2 and 4. This is due to a limitation of the AS procedure. Let us consider the fixed-width procedure to briefly illustrate the limitation of AS. Because  $T_0$  is a random variable whose value is determined from the stopping rule in (14), the remaining  $T - T_0$  tagged item to be collected at the second stage may turn out to be greater than  $t$ , the actual number of tagged items in the population. For the practical sequential approach, the estimate of  $B_T$  (16) may become less meaningful in such a scenario and it would be absurd to do the recapturing for the second stage.

### 4 Illustrative examples

This section presents two applications of MRM to illustrate the dynamics of the mark–recapture–mark procedure, as well as how it will be used to construct fixed-length confidence interval for population sizes in real-life problems.

#### 4.1 Estimating the number of COVID-19 cases

We first consider an application of MRM procedure in estimating the number of COVID-19 cases. In the occurrence of any pandemic, it is critical to know the number of affected individuals residing in an area. It is crucial to estimate the true total number of COVID-19 infections, because it affects the decisions of policy-makers. However, due to a large number of asymptomatic cases, most of them are undetected. Recently, applications of capture–recapture procedures for estimating the number of COVID-19 cases were discussed by Böhning et al. (2020) and Kumar and Mishra (2020). Thus, we propose to use our newly developed MRM

procedure to estimate the total number of infected cases including both detected and undetected. The necessity of this is still being felt in many countries, since COVID-19 is very much present. Given that many countries have adopted strict lock-downs in many areas to arrest the spread of the disease, we can reasonably assume that, at least during the period of lock-down, the population is considered to be “near-closed”.

In the light of Tuite and Fisman (2020), we start with a brief discussion on the estimation of the true outbreak size of infected in Italy, where the reported case count is 1128 on Feb 29, 2020. It is impossible for us to go back to February of 2020 to conduct our MRM procedure, but we try to explain the procedure here to illustrate how one can use our procedure for estimating future outbreak sizes. A given population can be considered to comprise two groups, affected and unaffected. For the affected (cases) group, there are two sub-groups, detected (who are formally reported) and undetected (who will be detected when tested). To implement the MRM procedure on COVID-19 cases, one can treat the newly detected cases as newly added marked individuals. Thus, in this illustration, we will only collect data on the group of affected individuals, and implement the MRM procedure focusing on their status of being reported or undetected. Here, we simulate the data based on a true outbreak size of  $N = 3971$  cases, because Tuite and Fisman (2020) estimated a true outbreak size of 3971 cases, and  $t = 1128$ , since there are 1128 reported cases. For our reference, the 95% CI given by Tuite and Fisman (2020) is [2907, 5297]. After applying the MRM method for a fixed-accuracy ( $\delta = 1.2$ ) 95% CI, we have obtained the interval estimate [3433, 4944] and the MLE estimate  $\hat{N} = 4119$ , which required a sample size of 2482 affected individuals.

Naturally, it is questionable if an entire country could be put under a lock-down as it strongly depends on ideological, political, economic, logistical, and many other social conditions of each country. Anyways, it seems safe to propose that the method has a practical appeal if applied at the very beginning of the pandemic in a small area, such as a county or a well-delimited subregion of a city, when only a few dozens of individuals are reported by the time. With that in mind, now we consider a situation where  $t = 50$  infected individuals were initially tested and identified. This new simulation is based on a hypothetical outbreak size of  $N = 500$  in a region of 20,000 residents. This way, with  $s = 50$ , we applied MRM to the simulated data to obtain a fixed-accuracy ( $\delta = 1.1$ ) 90% CI for  $N$ , which returned the interval estimate [466, 565], and a point MLE estimate equal to  $\hat{N} = 513$ .

To better explain the MRM procedure for the COVID-19 problem, consider Table 6, which shows the details on the number of tagged individuals for the first and last six withdrawals. In the beginning, we know that  $t = 50$  infected individuals have been detected (tagged) by tests. Also, remind that we fixed  $s = t = 50$ . Now, one would start the MRM procedure by randomly selecting an individual in the area of interest. If the person was not among the 50 tested individuals, then a test for COVID-19 will be performed. If the test is negative, we would ignore the data and consider the person to be unaffected. If the test is positive, then we would record the data and that means marking one more person and we would have  $X_1 = 0$ , and then, there will be 51 identified positive (tagged) individuals in the area. Then, we would randomly select another individual. Suppose the test was positive and the person was not one of the previously tested individuals. Then,  $X_2 = 0$ , and there will be 52 tagged individuals after this step. Suppose that the third person turned out to be one of the previously tested (tagged) persons. Then, we still have 52 tested individuals with  $X_3 = 1$ . This process will continue until we have observed 50 tested individuals. After that, the sequential procedure would stop if (5) holds. Otherwise,  $s$  is increased in on unity and one more person is randomly selected for continuing with the same resampling scheme. The experiment that we performed demanded 91 updates of  $s$ , i.e., it reached up to  $s = 50 + 91 = 141$  when the sequential

**Table 6** Detailed output of MRM based on the simulated data with  $N = 500$  and  $t = 50$  in a hypothetical region with 20,000 residents. The MRM parametrization was  $s = t = 50$ ,  $\delta = 1.1$ , and  $\alpha = 0.05$ 

Draw no. ( $i$ )	Tagged (Y/N)	$X_i$	$s$	Total tagged after the $i$ th draw
1	N	0	50	51
2	N	0	50	52
3	Y	1	50	52
4	N	1	50	53
5	N	1	50	54
6	Y	2	50	54
...	...	...	...	...
379	Y	137	140	293
380	Y	138	140	293
381	Y	139	140	293
382	N	139	141	294
383	Y	140	141	294
384	Y	141	141	294

process finished, and it has identified  $294 - 51 + 1 = 244$  new infected persons among the 384 draws.

Naturally, the assumption of confinement is fragile, and this application would work only if the time between mark and recapture is sufficiently short. Another relevant limitation is the fact that the individuals exposed to the virus may not be infectious by the time of the test for the disease. Also, we have to take into account the false-positive and false-negative rates of the COVID-19 tests. Hence, while the present example aims to start the elaboration on using mark–recapture sampling for epidemiological applications, these practical issues should be addressed for avoiding underestimate the size of the affected population.

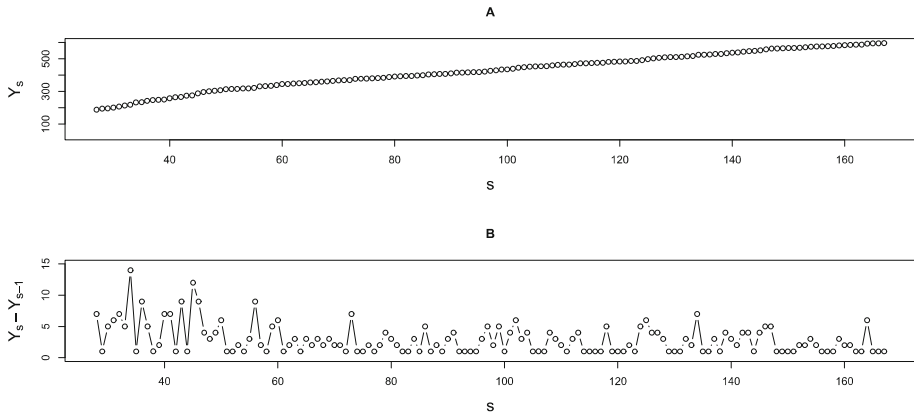
## 4.2 Population size of endangered species

We exemplify the applicability of the MRM method using the context of estimating the population size of endangered golden lion tamarin (*Leontopithecus rosalia*), here called GL, and construct a 90% fixed-accuracy confidence interval with the accuracy parameter  $\delta$  chosen to be 1.1.

The goal is to illustrate how the method works in practice. For this, we have simulated one trajectory of the variable  $Y_s$  for a fixed  $N = 1000$ . The real empirical part is the number of tagged animals,  $t = 27$ , from the study by Screnci-Ribeiro (2016).

In the work of Screnci-Ribeiro (2016),  $t = 27$  individuals in the population of GL were tagged and monitored in the year of 2014. The study was performed in the farm Fazenda Boa Esperança, Silva Jardim city, Rio de Janeiro, Brazil. GL is a neotropical callitrichid monkey endemic to the Atlantic forest of Brazil, a global biodiversity hotspot.

As the real data related to the mark–remark experiment of Screnci-Ribeiro (2016) did not involve marking the non-marked captured animals, it is not possible to actually apply the MRM procedure to that real data. Therefore, aiming to show the status of  $Y_s$  each time that  $s$  is updated in the course of the sequential procedure, here, we simulate the data based on a population size of  $N = 1000$ . This way, the dynamic of the  $Y_s$  process can be accompanied.



**Fig. 1** Values of  $Y_s$  and  $Y_s - Y_{s-1}$  for the sequential estimation of  $N$  using MRM for a fixed-accuracy ( $\delta = 1.1$ ) 90% confidence interval. The turning parameters

Figure 1a shows the entire trajectory of  $Y_s$ , from the starting point where  $s = 27$  tagged animals were captured, which occurred for  $Y_{27} = 188$ , to the stopping point, where the final adaptive value was  $s = 167$  with  $Y_{167} = 596$ . Note the decreasing rate at which  $Y_s$  evolves over the  $s$  updates. It seems that the procedure behaves as if it is ‘learning’ with the proportions of marked and unmarked animals, decreasing the variability of  $Y_s$  in the direction of requiring less recaptures to reach the next tagged animal. This behavior is more evident with Fig. 1b, where the first difference ( $Y_s - Y_{s-1}$ ,  $s = 28, \dots, 167$ ) highlights the smaller variability of the required captures among two consecutive  $s$  updates, just like it is expected to happen when the number of marked items is dense enough in the population. With these very simulated data, the estimate of  $N$  was  $\hat{N} = 942.5$ , leading to the confidence interval  $[942.5/1.1, 942.5 \times 1.1] = [856.82, 1036.75]$ .

### 5 Concluding remarks

The MRM method presents a satisfactory performance as it demands much less resampling effort in the recapture stage than one of the most prominent methods for population size estimation, the AS. Moreover, MRM keeps the coverage probability over the nominal confidence coefficient.

The drawback of MRM relates to the necessity of using Monte Carlo approximation to make the method computationally viable. While the exact solution would probably present a superior performance, the Monte Carlo version can, sometimes, lead to the inconvenience of keeping recapturing items from the population even when all the items are already tagged. For instance, take Tables 2 and 3. There, we can see that the expected recaptures dropped down as  $d$  or  $\delta$  gets larger for  $t = 50$  and  $t = 100$ . However, when  $t = 200$ , the expected recaptures are very close for different values of  $d$  and  $\delta$ . The Monte Carlo approximation in (9) is conservative by construction. Therefore, depending on the choices of  $d$ ,  $\delta$ , and  $t$ , it can require more recaptures than necessary before reaching the stopping point. However, MRM keeps the nominal confidence coefficient under control and it saves many resampling efforts comparing with AS. Thus, MRM is a powerful tool for sequential population size estimation.

To solve the disadvantage of MRM, one could try a simpler decision rule for the exact solution to make the calculations simpler too. Another possibility would be an asymptotic



solutions in the spirit of AS, which could lead to gains comparable to those with the MRM. The efforts of the authors to find such alternative solutions were unsuccessful by the moment; therefore, this is a pending issue for future investigations.

## References

- Anscombe FJ (1952) Large-sample theory of sequential estimation. *Proc Camb Phil Soc* 48:600–607
- Bailey TJ (1951) On estimating the size of mobile populations from recapture data. *Biometrika* 38:293–306
- Banerjee S, Mukhopadhyay N (2016) A general sequential fixed-accuracy confidence interval estimation methodology for a positive parameter: illustrations using health and safety data. *Ann Inst Stat Math* 68:541–570
- Böhning D, Rocchetti I, Maruotti A, Holling H (2020) Estimating the undetected infections in the Covid-19 outbreak by harnessing capture–recapture methods. *Int J Infect Dis* 97:197–201
- Bonner S, Holmberg J (2013) Mark-recapture with multiple, non-invasive marks. *Biometrics* 69(3):766–775
- Burden RL, Faires JD (1985) *Numerical analysis*. PWS Publishers, Worcester
- Chapman D (1952) Inverse, multiple and sequential sample censuses. *Biometrics* 8(4):286. <https://doi.org/10.2307/3001864>
- Chattopadhyay B, Mukhopadhyay N (2013) Two-stage fixed-width confidence intervals for a normal mean in the presence of suspect outliers. *Seq Anal* 32:134–157
- Chow YS, Robbins H (1965) On the asymptotic theory of fixed width sequential confidence intervals for the mean. *Ann Math Stat* 36:457–462
- Creel S, Spong G, Sands J, Rotella J, Zeigle J, Joe L, Murphy K, Smith D (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Mol Ecol* 12(7):2003–2009. <https://doi.org/10.1046/j.1365-294x.2003.01868.x>
- Dantzig GB (1940) On the non-existence of tests of “Student’s” hypothesis having power functions independent of  $\sigma$ . *Ann Math Stat* 11:186–192
- Darroch J (1958) The multiple-recapture census: I. Estimation of a closed population. *Biometrika* 45(3/4):343. <https://doi.org/10.2307/2333183>
- Epopa P, Millogo A, Collins C (2017) The use of sequential mark-release-recapture experiments to estimate population size, survival and dispersal of male mosquitoes of the *Anopheles gambiae* complex in Bana, a west African humid savannah village. *Parasites Vectors* 10(376):1–15
- Grimm A, Gruber B, Henle K (2014) Reliability of different mark-recapture methods for population size estimation tested against reference population sizes constructed from field data. *PLoS One* 9(6):e98840
- Hu J, Zhuang Y, Goldiner C (2021) Fixed-accuracy confidence interval estimation of  $p(x < y)$  under a geometric-exponential model. *Jpn J Stat Data Sci*. <https://doi.org/10.1007/s42081-021-00122-2>
- Jackson CHN (1939) The analysis of an animal population. *J Anim Ecol* 8(2):238–246
- Jennison C, Turnbull BW (1999) *Group sequential methods with applications to clinical trials*. Chapman and Hall/CRC, London
- Khalifeh A, Mahmoudi E, Chaturvedi A (2020) Sequential fixed-accuracy confidence intervals for the stress-strength reliability parameter for the exponential distribution: two-stage sampling procedure. *Comput Stat* 35:1553–1575
- Kumar R, Mishra R (2020) Estimation of Covid-19 prevalence under closed population using capture–recapture approach. *Int J Sci Res* 9:451–457
- Laplace PS (1786) Sur les naissances, les mariages et les morts. *Histoire de l’Académie Royale des Sciences*, p 693
- Leslie P, Chitty D (1951) The estimation of population parameters from data obtained by means of the capture–recapture method: I. The maximum likelihood equations for estimating the death-rate. *Biometrika* 38(3, 4):269–292
- Lincoln F (1930) Calculating waterfowl abundance on the basis of banding returns. US Department of Agriculture, Washington
- Moore JA, Grant T, Brown D, Keall SN, Nelson NJ (2010) Mark-recapture accurately estimates census for tuatara, a burrowing reptile. *J Wildl Manag* 74(4):897–901
- Mukhopadhyay N, Bhattacharjee D (2018) Sequentially estimating the required optimal observed number of tagged items with bounded risk in the recapture phase under inverse binomial sampling. *Seq Anal* 37(3):412–429
- Mukhopadhyay N, de Silva BM (2009) *Sequential methods and their applications*. CRC, Boca Raton

- Mukhopadhyay N, Zhuang Y (2017) Purely sequential and two-stage bounded-length confidence interval estimation problems in Fisher's "nile" example. *Jpn Stat Soc* 47(2):237–271. <https://doi.org/10.14490/jjss.47.237>
- Peterson CGJ (1896) The yearly immigration of young plaice into the Limfjord from the German sea. *Rep Danish Biol Stn* 6:1–48
- Sathasivam S, Alzaeemi S (2022) Studies in systems, decision and control: trisection and pentasection method: a modification of the bisection method for solving non-linear algebraic equations. In: Abdul Karim SA, Shafie A (eds) *Towards intelligent systems modeling and simulation*. Springer, Cham
- Schnabel Z (1938) The estimation of total fish population of a lake. *Am Math Mon* 45(6):348–352. <https://doi.org/10.2307/2304025>
- Schwarz C, Dempson J (1994) Mark-recapture estimation of a salmon smolt population. *Biometrics* 50(1):98–108. <https://doi.org/10.2307/2533200>
- Screnci-Ribeiro R (2016) *Uso de corredor ecologico por mico leoes dourados (leontopithecus rosalia, linnaeus, 1766) e saguis invasores na mata atlantica de baixada*. Master's thesis, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Rio de Janeiro Brazil
- Seber GAF, Schofield MR (2019) *Capture–recapture: parameter estimation for open animal populations*. Springer, Berlin
- Seber G, Huakau J, Simmons D (2000) Capture–recapture, epidemiology, and list mismatches: two lists. *Biometrics* 56(4):1227–1232. <https://doi.org/10.1111/j.0006-341x.2000.01227.x>
- Sekar C, Deming W (1949) On a method of estimating birth and death rates and the extent of registration. *J Am Stat Assoc* 44(245):101–115. <https://doi.org/10.1080/01621459.1949.10483294>
- Smith P (1988) Bayesian methods for multiple capture–recapture surveys. *Biometrics* 44(4):1177–1189. <https://doi.org/10.2307/2531745>
- Smout S, King R, Pomeroy P (2010) Estimating demographic parameters for capture–recapture data in the presence of multiple mark types. *Environ Ecol Stat* 18(2):331–347. <https://doi.org/10.1007/s10651-010-0135-y>
- Tuite NVREAR, Fisman D (2020) Estimation of Covid-19 outbreak size in Italy. *Lancet Infect Dis* 20(5):537
- Van Der Burg J (1995) Bisection is best. *Math Gaz* 79(486):575–577
- Villella RF, Smith DR, Lemarié DP (2004) Estimating survival and recruitment in a freshwater mussel population using mark-recapture techniques. *Am Midl Nat* 151(1):114–133
- Zhuang Y, Bhattacharjee D (2021) Minimum risk point estimation of the size of a finite population under mark-recapture strategy. *Seq Anal*. <https://doi.org/10.1080/07474946.2021.1912522>
- Zhuang Y, Hu J, Zou Y (2020) Fixed-accuracy confidence interval estimation of  $p(x > c)$  for a two-parameter gamma population. *Commun Stat Appl Methods* 27:625–639

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.