Revisiting Estimation of Finite Population Size

NITIS MUKHOPADHYAY
University of Connecticut
Department of Statistics
215 Glenbrook Road Storrs
USA
nitis.mukhopadhyay@uconn.edu

DEBANJAN BHATTACHARJEE
Utah Valley University
Department of Mathematics
800 West University Parkway Orem
USA
stat.debanjan@gmail.com

Abstract: In this article we will discuss estimation of a closed population size under inverse binomial sampling with mark-recapture strategy. This talk is based on the methodology laid out by Mukhopadhyay and Bhattacharjee (2018). Under squared error loss (SEL) as well as weighted SEL, we propose sequential methodologies to come up with bounded risk point estimators of an optimal choice of s, the number of tagged items; leading to an appropriate sequential estimator of N. The sequential estimation methodologies are supplemented with first-order asymptotic properties, which are followed by extensive data analyses. We might also briefly discuss other inferential procedures on estimating N.

Key–Words: Asymptotics; Bounded-risk; Capture; First-order properties; Recapture; Release; Risk; Sequential methodology; Squared error loss; Tagging; Weighted squared error loss.

1 Introduction

The celebrated **capture-tag-release-recapture** (CTRR) methodology is often used to estimate the size of a closed finite population. Scheaffer et al. (Chapter 10, 2012) includes an eloquent presentation and there are other references with more technical descriptions. Here, in particular the indirect method of sampling is the main focus of discussion where, we gather random sample of size n that will be required to observe a prefixed number of tagged elements (s). This is also known as Inverse Binomial Sampling where an important task is to determine the number of tagged items or s. In this article we will discuss a purely sequential bounded-risk estimation strategy. The desirable asymptotic properties of the procedure are discussed. We will also present data analyses through simulation and with data from designed experiments.

2 Probability Distribution and Risk Function

Let us denote $p = t/N$ which remains unknown since N is unknown. 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, $i = 1, ..., s$.

Then, the X_i's are i.i.d. having the common geometric distribution, referred to as Geometric(p), with the following *probability mass function* (p.m.f.):

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

Consider the total number of items observed, namely,

$$Y \equiv X_1 + ... + X_s, \tag{2}$$

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

$$\widehat{N}_s \equiv t\widehat{p^{-1}} = t\overline{X}_s \tag{3}$$

The error in estimating N can be quantified by a squared error loss function given by,

$$L_s \equiv L_s(\widehat{N}_s, N) = (\widehat{N}_s - N)^2 \tag{4}$$

If $\omega(> 0)$ is the bound of the associated risk function then we have, $R_s = t^2 s^{-1}(p^{-2} - p^{-1}) \leq \omega$ which will lead to the expression of the optimal (fixed) choice of s to be

$$s^* = t^2\omega^{-1}(p^{-2} - p^{-1}). \tag{5}$$

3 Purely Sequential Methodology

Since p is unknown, s^* remains unknown in (5), we propose a sequential bounded-risk estimation strategy. We begin with the first observation X_1, that is, with the one recaptured tagged $s = 1$. Then, we successively consider $s = 2, 3, ...$ one by one and terminate sampling according to the following stopping rule: Let

$$S = \inf\left\{s \geq 1 : s \geq \tfrac{t^2}{\omega}\left(\overline{X}_s^2 - \overline{X}_s + s^{-\gamma}\right)\right\}, \tag{6}$$

with arbitrary $\gamma(> 0)$. Hence, upon termination the final estimator of N will be,

$$\widehat{N}_S \equiv t\overline{X}_S = tS^{-1}\Sigma_{s=1}^{S}X_s. \tag{7}$$

and the sequential risk will be,

$$R_S \equiv E_N[L_S] = t^2 E_N\left[(\overline{X}_S - p^{-1})^2\right] \tag{8}$$

Theorem 1 *For the purely sequential estimation strategy $(S, \widehat{N}_S)$ proposed via (6)-(7), for all fixed values of t and N, we have as $\omega \to 0$:*

(i) $S/s^* \to 1$ with probability 1.

(ii) $E\left[(S/s^*)^k\right] \to 1$ for fixed $k(> 0)$

(iii) $E\left[(S/s^*)^k\right] \to 1$ for fixed $k(< 0)$

(iii) $s^{*-1/2}\left(\hat{N}_S - N\right) \to N(0, t^2qp^{-2})$ in distribution.

(iv) $s^{*-1/2}(S - s^*) \to N(0, (1+q)^2q^{-1})$ in distribution.

(v) $\omega^{-1}R_S \to 1$ with $\gamma > 1/2$, s^* and R_S come from (5) and (8) respectively.

Proof: Part (i) follows from the basic inequality, $t^2\omega^{-1}\{(\overline{X}_S^2 - \overline{X}_S) + S^{-\gamma}\} \leq S < t^2\omega^{-1}\{(\overline{X}_{S-1}^2 - \overline{X}_{S-1}) + (S-1)^{-\gamma}\} + 1$ and then taking the limit $\omega \to 0$.

Part (ii) is known as asymptotic efficiency. Note that $S < t^2\omega^{-1}\{4\overline{X}_S^2 + 1\}I(S \geq 2) + 1 \Rightarrow Ss^{*-1} < p^2q^{-1}(4U^2 + 1) + 1$, where $U = \sup_{s\geq 1}\overline{X}_s$. The result follows from (i) and dominated convergence theorem.

Part (iii) will adopt the techniques from Ghosh and Mukhopadhyay (1979) and Sen and Ghosh (1981). One can show that $P_N[S \leq (1-\epsilon)s^*] = O\left(s^{*-r/(2+2\gamma)}\right)$ for $0\epsilon < 1$ and some $r \geq 2$. The result will follow from the fact that $E_N\left[(s^*/S)^k I(S > \frac{1}{2}s^*)\right] = 1 + o(1)$.

Part (iv) can be easily shown from Anscombe's Random Central Limit Theorem.

Part (v) will follow from delta method, Slutsky's theorem and Ghosh-Mukhopadhyay (1975) theorem.

Part (vi) is called *asymptotic risk efficiency*. To prove this we see that, $E_N\left[(\overline{X}_S - p^{-1})^2\right] = E_N\left[s^{*-1}Y_{S\omega} + ((s^{*2}/S^2) - 1)s^{*-1}Y_{S\omega}\right]$, where $Y_{S\omega} \equiv s^{*-1}\left(\Sigma_{i=1}^{S}X_i - p^{-1}S\right)^2$.

Then we can show that, $Y_{S\omega}$ is uniformly integrable and $E_N\left[\left(s^{*2}S^{-2} - 1\right)Y_{s,\omega}\right] \to 0$ as $\omega \to 0$. The result will then follow by observing that, $\omega^{-1}R_S = E_N[Ss^{*-1}] + p^2q^{-1}E_N\left[((s^{*2}/S^2) - 1)Y_{S\omega}\right]$

□

4 Data Analysis

In this section first we briefly describe the simulation and comment on the results. Then we describe an experiment that was performed to collect data.

4.1 Simulation Process

First, we generate ID labels $1, 2, 3, ..., N-1, N$ corresponding to pretend-animals $\#1, \#2, \#3, ..., \#N-1$, $\#N$ in the full population under consideration. We pick t of these N animals (labels) selected by SRSWOR and tag them by turning them **bold**. Then, in the whole population, we place these t tagged labels back so that the whole population continues to include N labels of which t are tagged (that is, they are **bold**). Now, we permute the population and implement the sampling strategy proposed by the purely sequential stopping rule (6) by observing $X_1, X_2, ...$ until termination with the final data $\{S = s, x_1, ..., x_s\}$. During one run, when we move from stage $i-1$ to i, we have the full population on hand to sample from (using SRSWR) with exactly t tagged items. One should note that at each stage i, we permute the whole population and then go to the next stage $i+1$. Once we stop according to (6), that amounts to first full replication. From this first single replication, one will obtain a value of $(S = s, \widehat{N}_s)$. Data analysis is summarized for a number of choices of (N, t, ω). During simulation, we arbitrarily fixed $\gamma = 0.7$. Other values of $\gamma > 0.5$ gave similar results. For all choices of (N, t, ω) the results of Theorem 1 were empirically verified.

4.2 Experimental Data

In this section, we summarize the performances of a random experiment where we implemented the estimation methodology from Section 3 thereby collecting data sequentially in order to estimate population size, N. We had a jar containing a large unknown number (N) of silver coins.

From the jar, we selected a fistful of silver coins. We happened to pick 117 silver coins, but we replaced each with a gold coin and put these back inside the jar. In other words, the jar contained an unknown number of coins, N, but $t = 117$ of them were tagged as gold coins. All the coins were similar with regard to size, shape, and texture. We had at our disposal the full jar (= population) to sample from. We fixed $\omega = 2500$ and $\gamma = 0.7$. The data collection proceeded as follows:

Step 1: The entire jar was vigorously shaken and coins were selected from the full jar (obviously without looking) by *simple random sampling with replacement* (SRSWR). This process continued until a gold coin was observed. We kept track of the sequential number of tries to draw the first gold coin. As soon as the first gold coin appeared, we recorded the number of trials (X) and the run # (S).

Step 2: The first entry ($S = s = 1$) corresponded to the first run which terminated with $4(= X_1 = x_1)$ successively observed coins, but the first 3 were silver coins followed by the 4^{th} one which happened to be a gold coin. This meant that at the fourth draw, we observed the first gold coin. Then, we checked with the criterion for stopping defined through (6) to determine if the sampling could be terminated. Then, the random draws from the full jar began all over again and we continued drawing coins until we observed the next gold coin. The entries of the second row ($s = 2$) shows the second run which terminated with $7(= X_2 = x_2)$ successively observed coins, but the first 6 were silver coins followed by the 7^{th} one which happened to be a gold coin. Thus far, we had ($s = 1, x_1 = 4$) and ($s = 2, x_2 = 7$) successively. The stopping rule (6) asked us not to stop with $s = 2$.

Step 3: With each value of $X_1, X_2, X_3, ...$ observed in succession, the average $\overline{X}_s = s^{-1}\Sigma_{i=1}^{s} X_i$ was calculated sequentially, namely we came up with the observed values of $\overline{X}_1, \overline{X}_2, \overline{X}_3, ...$

Step 4: Steps 1-3 continued as we kept on checking successively for stopping. Continuing that way, it was $s = 91$ when we stopped according to (6).

Step 5: Proceeding as above, $s = 91$ was our final value of S in order to estimate the size N of the population, the total number of coins inside the jar. Thus, we came up with the following estimate of N: $\widehat{N}_{91} \equiv t\overline{X}_{91} = 117 \times 4.59 = 537.03$.

5 Conclusion

This article addresses an important question on determining the optimal number of tagged items to be observed in the CTRR sampling methodology. The purely sequential bounded-risk strategy provides excellent estimates of the population size.

But we also note that from a practical point of view, specifying appropriate ω by a practitioner under the formulation in Section 2 may be a bit concerning just because $\widehat{N}_s$ is highly variable for fixed and small s. In order to make life a bit "simpler", we may consider a weighted squared error loss function of the form

$$L_s \equiv L_s(\widehat{N}_s, N) = N^{-1}(\widehat{N}_s - N)^2. \qquad (9)$$

The loss quantified here will be considerably smaller compared with that shown in Section 2. One may adopt appropriate sequential methodology to study relevant properties and accuracy of estimating N under this weighted squared error loss function.

References:

[1] M. Ghosh, and N. Mukhopadhyay, Asymptotic Normality of Stopping Times in Sequential Analysis, Unpublished Report, 1975.

[2] M. Ghosh, and N. Mukhopadhyay, Sequential Point Estimation of the Mean when the Distribution is Unspecified, *Communications in Statistics-Theory & Methods* 8, 1979, pp. 637-652.

[3] N. Mukhopadhyay and D. Bhattacharjee, Sequentially Estimating the Required Optimal Observed Number of Tagged Items with Bounded Risk in the Recapture Phase Under Inverse Binomial Sampling *Sequential Analysis,* 37, 2018, pp. 412-429.

[4] R.–L. Scheaffer, W. Mendenhall, R.–L Ott and K.–G Gerow, *Elementary Survey Sampling* Boston: Brooks/Cole. 2012.

[5] S. Sen and M. Ghosh, Sequential Point Estimation of Estimable Parameters Based on U-Statistics, *Sankhya, Series A* 43, 1981, pp. 331–344.