International Journal of Statistics and Applications 2013, 3(3): 39-42
DOI: 10.5923/j.statistics.20130303.01

Estimating Clustered Population Size Using Two Stage Sampling when Capture Probabilities Vary among Individuals

Qian He, Naima Shifa*

Department of Mathematics, DePauw University, Greencastle, IN, 46135, USA

Abstract In real-world situations, researchers would like to estimate a characteristic of a population by observing and collecting information from only a part of the population. In recent years, many studies have been done in nature in an attempt to estimate the population total or population density. In ecological or biological studies, investigators might be interested in estimating the total number of animals in a huge study area where the animals live in clusters by nature and are rare in terms of huge territory and moving together for finding food or getting shelters from extreme weather. To obtain an estimator of the abundance of these clustered animals we propose a double stage sampling. This model is based on adaptive cluster sampling (ACS) to identify the location of the population and follow capture-recapture technique with unequal catchability to all units in a selected network to find the population abundance or the density. Some statistical properties of the propos ed estimator are also developed in this research.

Keywords Adaptive Cluster Sampling (ACS), Capture-recapture Sampling, Closed Population, Horvitz-Thompson Estimator, Jackknife Estimator, Two Stage Sampling

1. Introduction

The estimation of population size is of great importance in a variety of biological problems which are related to population growth, ecological adaptation, evolution and so on. The basic technique of capture-recapture (CR) was first introduced by Lincoln (1930)[9] to estimate the total number of duck in North America. The same method was adopted by Jackson (1933)[8] to estimate the true density of tsetse flies. Capture recapture methods have also been successfully applied to natural populations, like, moths by Fisher & Ford (1947)[4] caught, marked and released on several different days. Although in the various papers cited above effective use has been made of estimates of population size and of birth- and death-rates, there has been little discussion for varying capture probability which is obvious in natural population. In fact, equal capture probability is a convenient mathematical model with no empirical justification. Accurate estimation of population size requires with some degree of unequal probabilities of capture. The purpose of the paper is to produce a general model to estimate the size of a closed population allowing the variability of the capture probabilities among animal population. Actually, we present a two stage sampling procedure to estimate the total for a spatially aggregated, moving animal population in a huge study area, like, elephant-population in savanna forest in Africa or vampire bat- population in Central and South America. In this multistage sampling design, we first consider ACS and in the second stage we follow capture recapture (CR) sampling to obtain an estimate of the total of the population where the capture probabilities are varying among animals. In the second section, we discuss the idea of adaptive cluster sampling and then we derived the probability models for the animals with unequal capture probabilities. We derive the estimator for the animal total in each network and at last obtain the estimator of the population size in the study area. We also derive the variance and show the unbiasedness of the estimator.

2. Objective

There are many ways to modify cluster sampling for more complex sampling situations. One common modification is to take a sample of secondary units from within sampled clusters instead of inspecting every secondary unit within each sampled cluster. Estimating the wide variety of animal populations in nature is a complex situation. An efficient sampling procedure for estimating totals and means of rare and clustered populations was proposed by Thompson (1990)[13]. In this model, one is to take an initial sample by some ordinary sampling procedure, and, whenever the

* Corresponding author:
naimashifa@depauw.edu (Naima Shifa)
Published online at http://journal.sapub.org/statistics

variable of interest of a unit in the sample satisfies a previously specified condition, units in the neighbourhood of that unit are added to the sample. If any of the newly added units satisfy the condition, units in their neighbourhoods are also added until the sample includes all the neighbours of any unit satisfying the condition. The ACS technique is appropriate for sampling rare and clustered populations; one of its main drawbacks is the lack of control of the final sample size. Several studies have been done to control the final sample size well, and Salehi & Seber (2002)[12] have proposed a promising estimator of the mean. However, in this paper, we propose a two-stage version in which primary units are selected using a conventional design, and secondary units within the selected primary units are subsampled using adaptive cluster sampling designs. Then capture recapture method is applied. Our proposal, which we have called two stage sampling, requires the availability of an inexpensive and easy-to-measure auxiliary variable, which is used to select a first-phase adaptive cluster sample. The network structure of this first-phase sample is used to select the subsequent subsamples, which are selected using conventional designs. Only the values of the survey variable associated with the units in the final-stage subsample are recorded, and the population total is estimated by a capture recapture type of estimator. Our proposed two stage sampling design will allow the sampler to reach the following goals: to control the number of measurements of the variable of interest; to allocate the final-stage subsample near interesting places; to use the auxiliary variable at network selection; and to use different capture probability among animal population.

3. Methods and Models

We assume that we have a huge study area of *N* units of same size. Suppose we take a random sample of size *n* with or without replacement. If an observed sampled unit satisfies the condition of existence of a particular habitat, then the additional units of the neighborhood will be added to the sample. If any of these additional units satisfy the condition again, then the units in their neighborhoods are added to the sample also. The process continues until a cluster of units is obtained that contains a "boundary" of edge units that do not satisfy the condition.

We will follow the notations of Thompson and Seber (1996)[15]. Let A_i be the network for sampling unit *i*, that is, selection of any unit in A_i leads to the selection of all of A_i. Let m_i be the number of sampling units in A_i. Also let a_i be the total number of sampling units in networks of which sampling unit *i* is an edge unit. If the initial sample is selected without replacement, the probability that unit is included in the sample becomes,

$$\pi_i = 1 - \frac{\binom{N - m_i - a_i}{n}}{\binom{N}{n}} \tag{3.1}$$

If we do not consider the edge units, the partial inclusion probability (1) becomes,

$$\pi_i = 1 - \frac{\binom{N - m_i}{n}}{\binom{N}{n}} \tag{3.2}$$

If the probability, π_i (2) is known for all sampled units, we can use Horvitz-Thompson estimator (1952)[7] to estimate the population total, τ, namely,

$$\hat{\tau} = \sum_{i=1}^{n} \frac{\hat{\tau}_i}{\pi_i} = \sum_{i=1}^{N} \frac{\hat{\tau}_i I_i}{\pi_i} \tag{3.3}$$

In the above expression, $\hat{\tau}_i$ is the estimator of the total in the *ith* network and I_i takes 1 when the unit *i* is included in the sample, another words, if the initial sample intersects A_i (with probability π_i) and 0 otherwise.

If the initial sample is selected with replacement, then Hansen-Hurwitz estimator of the population total is suggested, see, Hansen and Hurwitz (1943)[6]. Now the probability of selecting ith unit, p_i is known and the inclusion probability becomes,

$$\pi_i = 1 - (1 - p_i)^n \tag{3.4}$$

The Hansen-Hurwitz estimator of the population total is given by,

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\tau}_i}{p_i} = \frac{1}{n} \sum_{i=1}^{N} \frac{\hat{\tau}_i J_i}{p_i} \tag{3.5}$$

In Hansen-Hurwitz estimator[11], J_i is the number of times unit I is selected and $J_i \sim Bin(n, p_i)$. The final sample then consists of *n* clusters, one for each unit selected in the initial sample. We apply a variable capture probability model in each network.

Notation and the model

In second stage of the model, we consider a closed population and derive an estimator to estimate the animal population size in a single network. This model allows variability in capture probabilities among animals[1]. The source of variation in the capture probabilities is the

heterogeneity among individuals. This model is applicable when the time difference between two the trapping occasions is short, such as consecutive days. Here the population size in the *ith* network is N_i.

p_{kit} = The capture probability of the *kth* animal at *t* capture occasion in the *ith* network.

$x_{kit} = 1$, If the *kth* animal is caught at *t* capture occasion in the *ith* network *and 0* otherwise.

$t = 1, \ldots, T, \quad i = 1, \ldots, n; \quad k = 1, \quad \ldots, N_i$.

p_{ki} : A random sample from *F*.

We have a data matrix of dimension, $N^* = N_i \times T$.

y_{ki} = The number of times the *kth* animal is captured in the *ith* network.

$$y_{ki} = \sum_{t=1}^{T} x_{kit}$$

f_{it} = The number of animals captured exactly *t*-times in the *ith* network.

f_{i0} = The number of animals never capture in the *ith* network.

S_i = The number of animals has been seen at least once during the trapping occasion in the *ith* network.

$$S_i = \sum_{t=1}^{T} f_{it}$$

Assumptions

1. The population at risk of capture is closed and is of size N_i.

2. $p_{ki} = p_i; k = 1, \cdots N_i$ is a random sample from a probability distribution *F*.

3. The random variables x_{kit} $(k = 1,2,...,N_i; i = 1,2,...,n, t = 1, \cdots, T)$ are mutually independent for given p_k.

In *N** matrix, we can only observe *Si* rows and it allows the calculations of the capture-recapture statistics of the unobserved rows are all zeros. The joint conditional distribution of x_{kit} is[1],

$$P(x \mid F) = \prod_{k=1}^{N_i} \prod_{t=1}^{T} p_k^{x_{kii}} (1 - p_k)^{1 - x_{iik}}$$

$$= \prod_{k=1}^{N_i} p_k^{y_{ki}} (1 - p_k)^{T - y_{ik}}$$

Since this probability distribution is not useful for estimation of N_i, we consider $p_i = p$ as a random sample and average over it to obtain the capture distribution of x_{kit} ,

$$P(x \mid F) = \prod_{k=1}^{N_i} \left[\int_0^1 p^{y_{ki}} (1-p)^{T - y_{ik}} \, dF(p) \right]$$

$$= \left[\int_0^1 (1-p)^t \, dF(p) \right]^{N_i - S_i}$$

$$\prod \left[\frac{\int_0^1 p^t (1-p)^{T-t} \, dF(p)}{\int_0^1 (1-p)^{f_{it}} \, dF(p)} \right]^{f_{it}}$$

For this model, the capture frequencies is the set of sufficient statistics and sufficiency holds for the entire class of distributions *F* of capture probabilities. Because of this, nonparametric method is applicable to estimate the population size. The unconditional distribution of the capture frequencies is,

$$P(f_{i0}, f_{i1}, \cdots, f_{iT}) = \frac{N_i!}{f_{i0}! \cdots f_{iT}!} \prod_{t=1}^{T} (\eta_t(F))$$

$$\eta_t(F) = \int_0^1 \binom{T}{t} p^t (1-p)^{T-t} \, dF(p)$$

Application of Jackknife estimator to estimate the population total

This method was first introduced by Gray and Schucany (1972). Let the initial estimator $\hat{N}_{0i} = S_i$, the number of animals captured in the *ith* network. Here S_i is the nonparametric maximum likelihood estimator of N_i. Again S_i is biased and the bias decreases as *T* increases.

$$E(S_i) = N_i + \frac{\alpha_1}{T} + \frac{\alpha_2}{T^2} + \cdots$$

Here $\alpha_1, \alpha_2, \cdots$ are constants. Here $\hat{N}_{kt}$ is a linear combination of capture frequencies and it is minimal sufficient statistic. It follows from elementary properties of the multinomial distribution[7] that

$$E\left[\hat{N}_{kt}\right] = N_i \sum_{t=1}^{T} \alpha_{kt} \eta_t(F)$$

$$Var\left[\hat{N}_{kt}\right] = \sum_{t=1}^{T} \alpha_{kt}^2 \, E\left[f_{it}\right] - \left[E\left(\hat{N}_{kt}\right)\right]^2 \Big/ N_i$$

After finding the U-statistics, the Jackknife estimators becomes,

$\hat{N}_{1i} = S_i + \frac{T-1}{T} f_{i1}$, $j = 1$, the order of Jackknife estimator.

$$\hat{N}_{2i} = S_i + \frac{2T-3}{T} f_{i1} - \left\{ \frac{(T-2)^2}{T(T-1)} \right\} f_{i2}, \quad j = 2$$

$$\hat{N}_{3i} = S_i + \frac{3T-5}{T} f_{i1} - \left\{ \frac{3T^2 - 15T + 19}{T(T-1)} \right\} f_{i2}$$

$$\text{[illegible]} - \left\{ \frac{(T-3)^2}{T(T-1)(T-2)} \right\} f_{i3}, \quad j = 3$$

Actually, $\hat{N}_{ji} = \sum_{t=1}^{T} a_{jit} f_{it}$ is a linear combination of the capture frequencies which is are minimal sufficient statistics.

Now the estimated animal total in the study area becomes,

$$\hat{\tau} = \sum_{i=1}^{n} \frac{\hat{N}_{ji}}{\pi_i}$$

Here the initial sample is selected by SRS without replacement with inclusion probability,

If the initial sample is selected by SRS with replacement, the estimator becomes,

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{N}_{ji}}{\pi_i}$$

Properties

In this two stage sampling, $\hat{\tau}$ is a biased estimator with population total with variance,

$$Var(\hat{\tau}) = \sum_{i=1}^{N} v_i + \sum_{i=1}^{N} \sum_{l=1}^{N} D_i D_l \left(\frac{\pi_{il} - \pi_i \pi_l}{\pi_i \pi_l} \right)$$

Where,

$$v_i = Var\left(\hat{N}_{ji}\right) = \sum_{t=1}^{T} a_{ij}^2 E\left(f_{ti}\right) - \frac{\left(E\left(\hat{N}_{ji}\right)\right)^2}{N}$$

And, $D_i = N_i \sum_{t=1}^{T} a_{jt} \eta(F)$

4. Discussion

This paper is a kind of outline of an ongoing research work. We still need to check the mathematical properties of the developed models by simulation study. We find that is would be extremely appropriate if we could show a real life application of this model. If it is not possible to collect data in from the real world, we are planning to perform a simulation study.

REFERENCES

[1] Burnham, K.P. and Overton, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65, 625-633.

[2] Burnham, K.P. and Overton, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. Ecology, 60, 927-936.

[3] Carothers, A.D. (1973). Capture-recapture methods applied to a population with known parameters. Journal of Animal Ecology 42, 125-146.

[4] Fisher, R. A. & Ford, E. B. (1947). The spread of a gene in natural conditions in a colony of the moth Panaxia dominula L. Heredity, 1, 143-74.

[5] Gray, H. L. & Schucany, W. R. (1972). The Generalized Jackknife Statistic. New York: Marcel Dekker.

[6] Hansen, M.M. and Horwitx, W.N. (1953). Sample Survey Methods and Theory Vol. 1 341-345. New York: Wiley.

[7] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association , 47, 663-685

[8] Jackson, C. H. N. (1933). On the true density of tsetse flies. Journal of Animal Ecology, 2, 204-9.

[9] Lincoln, F.C. (1930) Calculating waterfowl abundance on the basis of banding returns. Cir. U.S. Department of Agriculture, Vol. 118, 1-4, 1930.

[10] Rao, C. R. (1973). Linear Statistical Inference and its Applications. 2nd edition. New York: Wiley.

[11] Richard, J. L. and Morris, L. M (2006). An Introduction to Mathematical Statistics and its Applications. Pearson

[12] Salehi, M. M. & Seber, G. A. F. (2002). Unbiased estimators for restricted adaptive cluster sampling. Aust. New Zeal. J. Statist.. 44, 63-74.

[13] Thompson, S. K. (1990). Adaptive cluster sampling. J. Am. Statist. Assoc. 85, 1050-9.

[14] Thompson, S. K. (1991). Adaptive cluster sampling: Designs with primary and secondary units. Biometrics, 47, 1103-15.

[15] Thompson, S.K., and Seber, G.A.F. Adaptive Sampling . New York: Wiley, 1996.