On some aspects of estimation methods based on capture alone with applications

Sameer Y. Desale
Dhirendra N. Ghosh
Synectics for Management Decisions Inc.

Bimal K. Sinha
University of Maryland, Baltimore County

Abstract

In this paper we review and extend an earlier method of Ghosh (1985) to estimate the total number of a *special* type of population which is augmented constantly by a fixed number but are deflated by a fixed proportion (less than one) on a regular basis. The subsample that constitute fixed proportion of the existing population and is expunged provide an input namely the point in time when they entered the system. Both frequentist and Bayesian methods are described. Some generalizations are also mentioned. Two obvious examples are:

1 A species of fish in a closed environment, which are caught and removed at a regular interval and the age of the fish caught is noted.

2 Illegal immigrants to any country who are regularly apprehended and deported. They provide their time of arrival to the country.

1. Introduction

It is a fact that each year many foreign citizens enter the United States as illegal immigrants. However, in spite of serious efforts of the U.S. Immigration and Naturalization Service (INS), an exact or even an approximate count of the total number of such illegal immigrants is lacking due to obvious reasons. It is the object of this paper to review and extend an earlier attempt by Ghosh (1985) to solve this problem.

The capture recapture methodology is well known for estimating the total size of a population. However, in this case we have a devise a methodology for estimating the total size from capture alone. It is possible to interview a stratified sample of the apprehended illegal immigrants before deportation and determine their year of arrival to U.S. Assuming this information is available, it is possible to estimate the total size of illegal immigrants living in U.S. using statistical model.

Assume that a fixed number N of illegal immigrants enter every year and also that a fixed proportion P of them are apprehended every year. This means out of N immigrants who entered illegally last year, NP of them would be apprehended this year and so N(1-P) = NQ would be left over to join the new incoming N illegal immigrants this year. Likewise, those illegal immigrants who are apprehended this year but entered two years back belong to a group of NPQ persons. Based on these simple facts, Ghosh (1985)

provided *moment* estimates of both N and P on the basis of the number of people apprehended this year but who actually entered either last year (n_1) or the year before last (n_2). Since an estimate of P is given by $(n_1-n_2)/n_1$, it is obviously biased.

We examine this setup more closely in section 2

2. Main Methods

It is clear that n_1~ *Bino* (N, P), independently of n_2~ *Bino* (NQ, P). The likelihood of our observed data pair (n_1, n_2) can then be written as

$$L(N,P \mid n_1,n_2) = N_{C_{n_1}}.P^{n_1}.Q^{N-n_1}.NQ_{C_{n_2}}.P^{n_2}.Q^{NQ-n_2}. \quad (1)$$

While the method of moments for estimation of N and P is based on equating n_1 and n_2 with their respective expectations, NP and NPQ, we can try the method of maximum likelihood and also Bayesian estimation in this problem. For the MLE, we need to maximize the above likelihood with respect to both N and P. Because of obvious complexity of the likelihood, exact analytical expressions of the MLEs are impossible. However, numerical computation can be carried out to determine the MLEs fairly accurately for any given data set.

Since N and NQ are expected to be fairly large, one can use Stirling's approximation for factorials to approximate the likelihood given in (1) by

$$L^* = \frac{e^{-n_1-n_2} N^{N+\frac{1}{2}} (NQ)^{NQ+\frac{1}{2}} P^{n_1+n_2} Q^{N+NQ-n_1-n_2}}{n_1! n_2! (N-n_1)^{N-n_1+\frac{1}{2}} (NQ-n_2)^{NQ-n_2+\frac{1}{2}}} \quad (2)$$

The approximate maximum likelihood estimates of N and P are then obtained by solving the two equations

$$\frac{1}{N} + \ln N + Q\ln(NQ) + (1+Q)\ln Q + \frac{1}{2(N-n_1)} + \frac{Q}{2(NQ-n_2)} = \ln(N-n_1) + Q\ln(NQ-n_2) \quad (3)$$

$$\frac{(n_1+n_2)}{P} + N\ln(NQ-n_2) + \frac{N}{2(NQ-n_2)} = 2(N\ln Q) + N\ln N + \frac{N+NQ+1/2-n_1-n_2}{Q} \quad (4)$$

Using the simple fact that *ln(1+x) ~ x* for small *x*, it is possible to simplify the above two equations further, and one gets:

$$\frac{n_1+n_2+1}{N} + (1+Q)\ln Q + \frac{1}{2(N-n_1)} + \frac{Q}{2(NQ-n_2)} = 0 \quad (5)$$

$$\frac{n_1+n_2}{P} + \frac{N}{2(NQ-n_2)} = N\ln Q + \frac{N+NQ+\frac{1}{2}-n_1}{Q} \quad (6)$$

We present two examples to show this computation. As regards Bayesian estimation of N and P, we assume that, a priori, N has a uniform distribution over (L, U) where L and U are two reasonable limits, and P, independent of N, has beta prior $\Pi(P/\alpha, \beta)$ with known parameters α and β. Specific choices of alpha and beta reflecting one's prior belief about N and P with the likelihood given above to derive the joint posterior of N and P, given the data (n_1, n_2), which is given as

$$\pi(N, P \mid n_1, n_2) = KL(N, P \mid n_1 n_2)\pi(N)\pi(P) \qquad (7)$$

Where K represents the norming constant. It should be noted that, a posteriori, N and P are always dependent! Bayes estimates of N and P can then be obtained from the above joint posterior distribution. For example, $E(N/n_1, n_2)$ and $E(P/n_1, n_2)$ can be used as standard estimates of N and P, respectively.

We present below two examples showing our computations. An extensive study is also performed to compare the three proposal estimates of N and P in terms of their bias and mean squared error.

Table 1: Estimates of P and N using Method of moments

n_1	n_2	P	N
10	9	0.1	100
100	98	0.02	5000
200	192	0.04	5000
300	282	0.06	5000
400	368	0.08	5000
600	564	0.06	10000
700	672	0.04	17500
800	784	0.02	40000
900	864	0.04	22500
1000	940	0.06	16666.67

Table 2: Estimates of P and N using Maximum Likelihood Estimation

n_1	n_2	P	N
10	9	0.1	105
100	98	0.02	5000
200	192	0.04	4925
300	282	0.06	4866
400	368	0.08	4810
600	564	0.06	9714
700	672	0.04	17173
800	784	0.02	39649
900	864	0.04	22072
1000	940	0.06	16178

Reference: Ghosh, Survey Research Methods, ASA Proceedings 1985