Journal of the Korean
Data & Information Science Society
2001, Vol. 12, No.2, pp. 27 ~ 33

Asymptotic Distribution for Stopping Time in Estimating a Population Size [1]

Kiheon Choi [2]

Abstract

Suppose that there is a population of hidden objects of which the total number N is unknown. From such data, we derive an asymptotic distribution for stopping time.

Key Words and Phrases: Estimation population size, Exponential distribution, Stopping time, Anscombe's theorem, Asymptotic theory.

1 Introduction

Consider a problem which require us to find, observe, or catch some of or all of a group of hidden objects as prey. Examples of such prey are fish in a lake, potential voters in a voter registration drives, donors to charitable organizations, disintegrating atoms in a radioactive source, disease carriers, or relics at the site of an archaeological dig. This problem has been considered by several authors, including Starr(1974), Vardi(1980), Dalal and Mallows(1988).

Thus, consider an area containing N prey. Imagine the prey are labeled $1, \cdots, N$; let T_i denote the time at which we would capture the prey labeled i if we are to search indefinitely. We suppose throughout that $T_1, \cdots, T_N$ are independent and identically distributed with a continuous distribution function F for which $F(0) = 0$. The distribution function F may depend on an unknown parameter θ, or not. Let $t_1 \leq \cdots \leq t_N$ denote the order statistics of $T_1, \cdots, T_N$. If the search is continued for t units of times, then the available data consists of the number of objects found and the times at which they were found; in symbols,

$$K_t = \#\{k \leq N : T_k \leq t\}.$$

[1]This work was supported by the Institute of Natural Science Fund, Duksung Women's University, 2001

[2]Associate Professor, Department of Statistics, Duksung Women's University, Seoul, 132-714, Korea.

Let $\hat{N}_t$ denote an estimator of N, based on this data. Then we wish to find the distribution of stopping time

$$\tau = inf\left\{t > 0 : \hat{N}_t \geq \frac{4\sigma^2(t\hat{\theta}_t)}{h^2}\right\}.$$

2 Asymptotic Normality for Stopping Time

F is assumed to be known, continuous distribution function that is strictly increasing on the interval $(0, b_F)$, where $b_F = sup\{t : F(t) < 1\} \leq \infty$. Then the maximum likelihood estimator of F after t time units of observation is (an integer adjacent to)

$$\hat{N}_t = \frac{K_t}{F(t)},$$

for $0 < t < b_F$. Since K_t has binomial distribution, the mean and variance of $\hat{N}_t$ are $E_N[\hat{N}_t] = N$ and $D_N^2[\hat{N}_t] = N\sigma^2(t)$, where

$$\sigma^2(t) = \frac{1}{F(t)} - 1,$$

and $\hat{N}_t$ is asymptotically normal as $N \to \infty$ for fixed $t > 0$; that is

$$\frac{\hat{N}_t - N}{\sqrt{N\sigma^2(t)}} \Rightarrow Z \sim \Phi,$$

where $\Rightarrow$ denotes convergence in distribution and Φ denotes the standard normal distribution.

Next, we consider that F is assumed to be an exponential distribution with unknown failure rate θ. There are then two aspects to be the problem, estimating θ and then estimating N.

It is convenient to begin with the likelihood function: if $k \geq 1$ and $0 \leq t_1 \leq \cdots \leq t_k \leq t$, then

$$\begin{aligned} P_{\theta,N}\{K_t &= k, t_1 \leq T_1 \leq t_1 + dt_1, \cdots, t_k \leq T_k \leq t_k + dt_k\} \\ &\propto (N)_k \, e^{(N-k)\theta t}\theta e^{-\theta t_1} \times \cdots \times \theta e^{-\theta t_k} dt_1 \cdots dt_k \\ &= (N)_k \, \theta^k exp\{-(N-k)\theta t - \theta s_t\} dt_1 \cdots dt_k, \end{aligned}$$

where $(N)_k = N(N-1) \times \cdots \times (N-k+1)$ and $s_t = t_1 + \cdots + t_k$. It is easily seen that for each fixed t, the marginal distribution of K_t is

$$K_t \sim Binomial(N, 1 - e^{-\theta t}).$$

and the conditional density of $X_j = t_j/t, j = 1, \cdots, k$, given that $K_t = k$, is the same as the distribution of the order statistics of a sample of size k from the density

$$f_w(x) = \frac{we^{-wx}}{1-e^{-w}}, \quad 0 \leq x \leq 1,$$

where $w = \theta t$. Let $\mu(w)$ denote the mean of f_w. Further, we know that the family f_ω, $\omega > 0$, is an exponential family, $f_\omega(x) = exp[-\omega x - \psi(\omega)]$, $0 \leq x \leq 1$, with cumulant generating function

$$\psi(\omega) = log(1-e^{-\omega}) - log(\omega).$$

So the mean and variance of f_ω are

$$\mu(\omega) = -\psi'(\omega) = \frac{1}{\omega} - \frac{1}{e^\omega - 1} = \frac{e^\omega - 1 - \omega}{\omega(e^\omega - 1)}$$

and

$$\psi''(\omega) = \frac{1}{\omega^2} - \frac{e^\omega}{(e^\omega-1)^2} = \frac{(e^\omega-1)^2 - \omega^2 e^\omega}{\omega^2(e^\omega-1)^2}.$$

Now, we estimate N and θ by the method of moments. Let

$$S_t = t_1 + \cdots + t_{K_t}.$$

Then

$$E(S_t|K_t = k) = tk\mu(t\theta).$$

Let $\hat{\theta}_t$ solve the equation

$$\mu(t\hat{\theta}_t) = \frac{S_t}{tK_t};$$

that is,

$$\hat{\theta}_t = \frac{1}{t}\mu^{-1}\left(\frac{S_t}{tK_t}\right).$$

Also let

$$\hat{N}_t = \frac{K_t}{1-e^{-t\hat{\theta}_t}}.$$

Theorem 2.1 *For fixed $\theta > 0$, $t > 0$, $\hat{N}_t$ is asymptotically normal with mean N and variance $N\sigma^2(t\theta)$ as $N \to \infty$, where*

$$\sigma^2(t\theta) = \frac{1}{e^{t\theta}-1}\left[1 + \frac{e^{t\theta}}{(e^{\theta t}-1)^2\psi''(t\theta)}\right].$$

Proof. See Choi(1999).

Now we consider the problem of obtaining asymptotic normality for stopping time.

Let h be a fixed length. From Theorem 2.1, we have

$$P_{N,\theta}\left(\left|\frac{\hat{N}_t}{N}-1\right|\leq h\right)\approx 2\Phi\left[\frac{h\sqrt{N}}{\sigma(t\theta)}\right]-1$$

as $N\to\infty$, and we need

$$\frac{h\sqrt{N}}{\sigma(t\theta)}\geq 2.$$

This suggests that we continue sampling until

$$\frac{\hat{N}_t}{\sigma^2(t\hat{\theta}_t)}\geq\frac{4}{h^2}.$$

For the asymptotic, suppose (without essential loss of generality) that

$$\theta=1.$$

Consider small t, say

$$t\downarrow 0.$$

Then

$$1-e^{-t}\sim t,$$

$$\sigma^2(t)\sim\frac{12}{t^3}.$$

We define the stopping time

$$\tau=inf\left\{t:\hat{N}_t\geq\frac{4}{h^2}\sigma^2(t).\right\}$$

First, We need to show that

Lemma 2.1 *If*

$$t=\frac{S_t}{N^{\frac{1}{3}}},$$

then

$$\frac{\hat{N}_t}{N}\xrightarrow{P}1.$$

Proof. Since

$$\frac{\hat{N}_t}{N} = \frac{K_t}{N}\frac{1}{1-e^{-\hat{\theta}t}}$$

and

$$\frac{\hat{N}_t}{N} - 1 = \frac{1}{N}\left[\frac{K_t}{1-e^{-\hat{\theta}t}} - \frac{N(1-e^{-t})}{1-e^{-t}}\right].$$

Now

$$\begin{aligned} V\left[\frac{K_t}{(1-e^{-t})N}\right] &= \frac{1}{(1-e^{-t})^2N^2}N(1-e^{-t})e^{-t} \\ &= \frac{1}{N(e^t-1)} = O\left(\frac{1}{N^{\frac{2}{3}}}\right). \end{aligned}$$

So

$$\frac{K_t}{(1-e^{-t})N} \xrightarrow{p} 1.$$

Since $S_t = t_1 + \cdots + t_{K_t}$. Then

$$E(S_t|K_t = k) = kt\mu(t).$$

Let $\hat{\theta}$ solve the equation

$$\mu(\hat{\theta}) = \frac{S_t}{tK_t}.$$

that is,

$$\hat{\theta} = \frac{1}{t}\mu^{-1}\left(\frac{S_t}{tK_t}\right).$$

Also

$$V(S_t|K_t = k) = k\psi''(t).$$

and

$$V\left(\frac{S_t}{tK_t}|K_t = k\right) = \frac{\psi''(t)}{t^2k}.$$

Now, we know that given $K_t = k$, $t_1, \cdots, t_k$ are order statistics from

$$f(x) = \frac{e^{-x}}{1-e^{-t}}, \qquad 0 \le x \le t.$$

But

$$\begin{aligned} V\left(\frac{S_t}{tK_t}|K_t = k\right) &\approx \frac{1}{12K_t} \\ &\xrightarrow{p} 0. \end{aligned}$$

So, $\hat{\theta}$ is consistent estimator, i.e.,

$$\hat{\theta} \longrightarrow 1 (= \theta).$$

Therefore,

$$\begin{aligned} \frac{\hat{N}_t}{N} &= \frac{K_t}{N(1-e^{-\hat{\theta}t})} \\ &= \left[\frac{1-e^{-t}}{1-e^{-\hat{\theta}t}}\right] \frac{K_t}{N(1-e^{-t})}. \end{aligned}$$

Thus

$$\frac{\hat{N}_t}{N} \xrightarrow{p} 1.$$

Next, we recall that the stopping time is

$$\tau = inf\left\{t : \hat{N}_t \geq \frac{4}{h^2}\sigma^2(\hat{\theta}t)\right\}.$$

Set

$$t = \frac{s}{N^{\frac{1}{3}}}.$$

Suppose $\tau > t$, then

$$\hat{N}_t < \frac{4}{h^2}\sigma^2(\hat{\theta}t) \quad \approx \quad \frac{48}{h^2\hat{\theta}^3 t^3}.$$

So

$$\frac{\hat{N}_t}{N} < \frac{48}{h^2\hat{\theta}^3 s^3} \quad \xrightarrow{p} \quad \frac{48}{h^2 s^3}.$$

Thus

$$N^{\frac{1}{3}}\tau \xrightarrow{p} \left(\frac{48}{h^2}\right)^{\frac{1}{3}}$$

Theorem 2.2

$$\frac{\hat{N}_\tau - N}{\sqrt{N\sigma^2(\hat{\theta}_\tau \tau)}} \Longrightarrow N(0,\ 1)$$

Proof. Theorem follows from Theorem 2.1, and Anscombe's theorem(see Gut(1988), Theorem 3.1).

References

1. Billingsley, P. (1986), *Probability and Measure*, 2nd edition, Wiley, New York.

2. Choi, K. (1999), Asymptotic Distribution in Estimating a Population Size. *J. Korean Data & Information Science Society*, Vol. 12, No.2, 313-318.

3. Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton Univ. Press.

4. Dalal, S.C. and Mallows, C.L. (1988), When should on stop testing software. *J. Amer. Statist. Assn.,* Vol. 83, No. 403, 872-879.

5. Gut, A. (1988), *Stopped Random Walks, Limit Theorems and Applications*, Springer-Verlag, New York.

6. Starr, Norman (1974), Optimal and adaptive stopping based on capture times. *J. Appl. Probab.*, 11, 294-301.

7. Vardi, Yehuda (1980), On a stopping time of Starr and its use in estimating the number of transmission sources, *J. Appl. Probab.*, 17, 235-242.