Journal of the Korean
Data & Information Science Society
1999, Vol. 10, No. 2, pp. 313 ~ 318

Asymptotic Distribution 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.

Key Words and Phrases: Estimation population size, Exponential distribution, 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 radioative source, diesease 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 labelled $1, \cdots, N$; let T_i denote the time at which we would capture the prey labelled 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,

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

Let $\hat{N}_t$ denote an estimator of N. We wish to find the distribution of $\hat{N}_t$.

[1]This work was supported by the Institute of Natural Science Fund, Duksung Women's University, 1999

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

2. The Model

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 $D_N^2[\hat{N}_t] = N\sigma^2$, 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

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

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

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

Also,

Lemma 2.1 *The conditional distribution of $X_j = t_j/t$, $j = 1, ..., k$, given $K_t = k$ has the same distribution as the order statistics corresponding the k independent random variable's with density $f_{\theta t}$, where*

$$f_\omega(x) = \frac{\omega e^{-\omega x}}{1 - e^{-\omega}},\ 0 \leq x \leq 1,$$

for $\omega > 0$.

Proof. Now

$$P(T \leq x|T \leq t) = \frac{1 - e^{-\theta x}}{1 - e^{-\theta t}},\ 0 \leq x \leq t.$$

So

$$P\left(\frac{T}{t} \leq u|T \leq t\right) \ = \ \frac{1 - e^{-\theta tu}}{1 - e^{-\theta t}},\ 0 \leq u \leq 1.$$

□

In particular, the conditional distribution depends only on θt and not on N. This provides a method for estimating θ.

Lemma 2.2 *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*

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

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

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

Proof.

$$f_\omega(x) = e^{-\omega x + log\frac{\omega}{1-e^{-\omega}}} = e^{-\omega x - \psi(\omega)}.$$

So

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

and

$$\mu(\omega) = E_\omega X = -E_\omega(-X) = -\psi'(\omega).$$

□

Lemma 2.3 *$\mu(\omega)$ is strictly decreasing in ω. Moreover,*

$$\lim_{\omega \to 0} \psi''(\omega) = \frac{1}{12}$$

and

$$\mu(\omega) = \frac{1}{2} - \frac{1}{12}\omega + O(\omega^2)$$

as $\omega \to 0$.

Proof. By Taylor expansion.

□

So, (conditional) maximum likelihood estimators may be obtained from the method of moments.

Theorem 2.1 *For fixed $\theta > 0,\ t > 0$, the conditional distribution of $\hat{\theta}_t$ given $K_t = k$ is asymptotically normal with mean $t\theta$ and variance $1/k\psi''(t\theta)$ as $k \to \infty$, and the unconditional distribution is asymptotically normal with mean $t\theta$ and variance $1/[N(1-e^{-t\theta})\psi''(t\theta)]$ as $N \to \infty$.*

Proof.
These assertions follow immediately from the asymptotic properties of maximum likelihood estimators and the law of large numbers for K_t.

3. Asymptotic Normality of Moments Estimators

First, 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 3.2 *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.

Theorem 3.2 follows from Theorem 2.1 and the asymptotic normality of K_t. So

$$K_t \Rightarrow Normal[N(1-e^{-t\theta}), N(1-e^{-t\theta})e^{-t\theta}].$$

Write

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

Then

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

Here

$$\frac{K_t - N(1-e^{-t\theta})}{\sqrt{N}} \Rightarrow Normal\left[0, (1-e^{-t\theta})e^{-t\theta}\right]$$

and

$$e^{-t\hat{\theta}_t} - e^{-t\theta} \Rightarrow Normal\left[0, \frac{e^{-2t\theta}}{N(1-e^{-t\theta})\psi''(t\theta)}\right].$$

Also using the delta method. Now

$$\frac{\hat{N}_t - N}{\sqrt{N}} \Rightarrow Normal[0, \sigma^2(t\theta)],$$

where

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

□

References

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

2. Dalal, S.C. and Mallows, C.L. (1988), When should on stop testing software. *Journal of the American Statistical Association*, Vol. 83, No. 403, 872-879.

3. Starr, Norman (1972), Optimal and adaptive stoping based on capture times. *Journal of Applied Probability*, 11, 294-301.

4. Vardi, Yehuda (1980), On a stopping time of Starr and its use in estimating the number of tranmission sources, *Journal of Applied Probability*, 17, 235-242.