Journal of the Korean
Data & Information Science Society
2000, Vol. 11, No. 2, pp. 327 ~ 334

Limiting Processes of 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 some properties of the limiting processes of stopping time in estimating a population size.

Key Words and Phrases: stopping time, limiting processes, Brownian motion.

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, 2000

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

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

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

for fixed $h > 0$ as in equation (2) and its properties.

2. Limiting Process 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 N 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 $D^2_N[\hat{N}_t] = N\sigma^2(t)$, 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

$$\frac{\hat{N}_t - N}{\sqrt{N\sigma^2(t)}} \Rightarrow Z, \tag{1}$$

where $\Rightarrow$ denotes convergence in distribution and Z denotes the standard normal distribution Φ. In fact, (1) holds for sequences $t = t_N > 0$ for which $N\sigma^2(t_N) \to \infty$ as $N \to \infty$. Using asymptotic normality to set an approximate to confidence interval and imposing the condition that the half width of the interval be at most $h\hat{N}_t$, as in Chow and Robbins(1965), suggests sampling until $\hat{N}_t \geq z^2\sigma^2(t)/h^2$, where $\Phi(z) = (1+\gamma)/2$ (and γ is the desired confidence coefficient).

Now $F = F_\theta$ is assumed to be an exponential distribution with unknown failure rate θ, $F_\theta = 1 - e^{-\theta x}$ for $x \geq 0$.

Let h be a fixed length, 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

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

Consider small t, say

$$t \downarrow 0.$$

Then

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

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

and

$$\frac{\hat{N}_t}{\sigma^2(t\hat{\theta}_t)} \approx \frac{(t\hat{\theta}_t)^3 \hat{N}_t}{12} \approx \frac{(t\hat{\theta}_t)^2 K_t}{12}.$$

Further, recalling that K_t has a binomial distribution, it is then easily seen that 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. Then, by Taylor expansion,

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

so that

$$t\theta_t \approx 6\left[1 - 2\mu(t\hat{\theta}_t)\right] = 6\left[1 - 2\frac{S_t}{tK_t}\right],$$

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

$$\frac{\hat{N}_t}{\sigma^2(t\hat{\theta}_t)} \approx 3K_t\left[1 - 2\frac{S_t}{tK_t}\right]^2.$$

Using these approximations and letting $x_+ = max(0, x)$, relation (2) may be rewritten: continue sampling until

$$3K_t\left[1 - 2\frac{S_t}{tK_t}\right]_+^2 \geq \frac{4}{h^2}.$$

This in turn suggests the stopping time,

$$\tau_h = \inf\left\{t > 0 : 3[tK_t - 2S_t]_+^2 \geq \frac{4t^2K_t + 1}{h^2}\right\},$$

where the term $1/h^2$ has been added to discourage early stopping.

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

$$\theta = 1.$$

and consider small t, say

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

where $0 < s < \infty$. Then

$$t^2 K_t \xrightarrow{p} s^3,$$

as $N \to \infty$. Let

$$K_N^0(s) = N^{-\frac{1}{3}}[K_t - N(1-e^{-t})],$$
$$K_N(s) = N^{-\frac{1}{3}}[K_t - Nt]$$

and

$$A_N(s) = \sqrt{3}[tK_t - 2S_t].$$

Thus,

$$N^{\frac{1}{3}}\tau_h = \inf\left\{s > 0 : A_N(s) > \frac{1}{h}\sqrt{4t^2K_t + 1}\right\}.$$

There is a simple relation between K_N and K_N^0,

$$K_N(s) - K_N^0(s) = N^{\frac{2}{3}}[1 - e^{-t} - t] \to \frac{1}{2}s^2$$

uniformly on compacts in $0 \le s < \infty$.

Next let $B(s)$, $0 \le s < \infty$, be a standard Brownian motion, and let

$$A(s) = \sqrt{3}\int_0^s (s-2u)dB(u) + \frac{\sqrt{3}}{6}s^3.$$

Observe that $A(s)$ is a Gaussian process with mean and covariance functions

$$E[A(s)] = \frac{\sqrt{3}}{6}s^3$$

and

$$\begin{aligned} Cov[A(s_1), A(s_2)] &= 3\int_0^{s_1}(s_1-2u)(s_2-2u)du \\ &= 3[s_1 s_2 u - (s_1+s_2)u^2 + \frac{4}{3}u^3]_{u=0}^{s_1} \\ &= s_1^3 \end{aligned}$$

for $0 \le s_1 \le s_2 < \infty$. Thus $\tilde{A}(s) := A(s^{1/3})$ is Brownian motion with drift parameter $\sqrt{3}/6$ and unit diffusion parameter. Next, let $D[0,c]$ denote the space of càdlàg functions on the interval $[0,c]$, as described by Billingsley(1968, Ch.3), for example.

Lemma 2.1 *As* $N \to \infty$, $K_N^0 \Rightarrow B$ *in* $D[0,\infty)$.

Proof.

$$K_N^0(s) = N^{1/6} \cdot N^{-1/2}(K_t - N(1-e^{-t})) \Rightarrow N(0,s),$$

we interpret $N(0,s)$ as a normal random variable with mean 0 and variance s and for $s_1 < s_2$ we have

$$\begin{aligned} Cov(K_N^0(s_1), K_N^0(s_2)) &= E(K_N^0(s_1)K_N^0(s_2)) \\ &= N^{-2/3}E\{\sum_{i=1}^{N}[1_{X_i \le z_1} - (1-e^{-z_1})]\sum_{j=1}^{N}[1_{X_j \le z_2} - (1-e^{-z_2})] \\ &= N^{1/3}e^{-z_2}(1-e^{-z_1}) \\ &\to s_1. \end{aligned}$$

Theorem 2.1 *As* $N \to \infty$,

$$A_N(s) \Rightarrow A(s)$$

in $D[0,\infty)$.

Proof. First observe that

$$S_t = \sum_{k=1}^{N} T_k 1_{\{T_k \le t\}} = \int_0^t u dK_u.$$

So,

$$A_N(s) = \sqrt{3}\int_0^t (t-2u)dK_u = \sqrt{3}\int_0^s (s-2u)dK_N(u),$$

and

$$\begin{aligned} A_N(s) &\Rightarrow \sqrt{3}\int_0^s (s-2u)[dB(u) - udu] \\ &= \sqrt{3}\int_0^s (s-2u)dB(u) + \frac{\sqrt{3}}{6}s^3, \end{aligned}$$

as asserted.

Corollary 1 *As* $N \to \infty$,

$$N^{\frac{1}{3}}\tau_h \Rightarrow \eta_h = \inf\{s > 0 : A_+(s) > \frac{1}{h}\sqrt{4s^3+1}\}.$$

Proof. By Theorem 2.1 and $t^2 K_t \xrightarrow{p} s^3$.

Theorem 2.2 *As* $N \to \infty$,

$$\hat{\theta}_t \Rightarrow \frac{2\sqrt{3}}{s^3}A(s)$$

in $D(0,\infty)$.

Proof.

First recall that $t^2K_t \xrightarrow{p} s^3$ and that

$$\frac{S_t}{tK_t} = \mu(t\hat{\theta}_t) = \frac{1}{2} - \frac{1}{12}t\hat{\theta}_t + O(t^2\hat{\theta}_t^2).$$

So,

$$\begin{aligned}
\hat{\theta}_t &= \frac{6}{t}(1 - 2\frac{S_t}{tK_t}) + O(t\hat{\theta}_t^2) \\
&= \frac{6}{t^2K_t}[tK_t - 2S_t] + O(t\hat{\theta}_t^2) \\
&= \frac{2\sqrt{3}}{t^2K_t}A_N(s) + O(t\hat{\theta}_t^2) \\
&\Rightarrow \frac{2\sqrt{3}}{s^3}A(s),
\end{aligned}$$

as $N \to \infty$.

Next let

$$Z_N(s) = \frac{N}{\hat{N}_t}.$$

Corollary 2 *As* $N \to \infty$,

$$Z_N(s) \Rightarrow \frac{2\sqrt{3}}{s^3}A(s)$$

in $D(0, \infty)$.

Proof.

We have

$$\begin{aligned}
\frac{N}{\hat{N}_t} &= \frac{1 - e^{-t\hat{\theta}_t}}{1 - e^{-t}} \times \frac{N(1 - e^{-t})}{K_t} \\
&= \hat{\theta}_t + O(t^2) \Rightarrow \frac{2\sqrt{3}}{s^3}A(s),
\end{aligned}$$

as $N \to \infty$.

Recall the definition of η_h in the Corollary 1.

Theorem 2.3

$$lim_{h\downarrow 0}h^{\frac{2}{3}}\eta_h = 2 \times 6^{\frac{1}{3}}$$

Proof. For fixed h,

$$\frac{\sqrt{3}}{6}\eta_h^3 = \frac{1}{h}\sqrt{4\eta_h^3 + 1}$$

So

$$\begin{aligned}\frac{h^2}{12}\eta_h^6 &= 4\eta_h^3+1\\ h^2\eta_h^3 &= 48+\frac{12}{\eta_h^3}\\ &\to 48\end{aligned}$$

as $h \downarrow 0$.

References

1. Billingsley, P. (1968). *Convergence of Probability and Measure*, Wiley, New York.
2. Billingsley, P. (1986). *Probability and Measure*, 2nd edition, Wiley, New York.
3. Chow, Y. S. and Robbins, H. (1965). Asymptotic theory of fixed width confidence intervals for the mean. *Ann. Math. Statist.*, 36: 457-462.
4. Dalal, S.C. and Mallows, C.L. (1988). When should on stop testing software. *Journal of American Statistical Association*, Vol. 83, No. 403, 872-879.
5. Starr, Norman (1972). Optimal and adaptive stopping based on capture times. *Journal of Applied Probability*, 11, 294-301.
6. Vardi, Yehuda (1980). On a stopping time of Starr and its use in estimating the number of transmission sources, *Journal of Applied Probability*, 17, 235-242.