Statistica Sinica 4(1994), 335-352

SEQUENTIAL UNBIASED ESTIMATION OF THE NUMBER OF CLASSES IN A POPULATION

Mary C. Christman and Tapan K. Nayak

George Washington University

Abstract: We consider the unbiased estimation of the number of classes, ν, in a population when the classes are equally likely to occur. Among the stopping rules based on a minimal sufficient statistic, the closed and complete plans are characterized. It is shown that ν cannot be estimated unbiasedly if the sample size is bounded; but unbounded, closed and complete plans admit best unbiased estimators of all functions of ν. A general rule for obtaining such estimators is given. It is also shown that without any assumptions about the class probabilities, ν cannot be estimated unbiasedly.

Key words and phrases: Sufficiency, completeness, stopping rule, closed sampling plan.

1. Introduction

Consider a population consisting of an unknown number, $\nu(\geq 1)$, of classes. Each member of the population belongs to one of the ν classes and any specific class is discovered when a member of that class is observed for the first time. The members of the population are selected at random, one at a time, and with replacement if the population is finite. We shall consider unbiased estimation of ν (or a function of ν) using the data, which consist of the class types of the successive members selected in the sample. Since ν cannot be estimated unbiasedly without any further assumptions (see Section 5) we shall assume as in Goodman (1953), Harris (1968), and others that all classes are equally likely to occur. The problem can also be stated in terms of an urn model where an urn contains equal numbers of ν differently colored balls. Balls are drawn at random and with replacement and the colors of the selected balls are noted. The problem is to estimate the number of colors using the data. For any given set of selections, let R (M) be the number of selections in which a new class is (not) discovered. Clearly, $R+M=N$ is the total number of selections. Harris (1968) showed that from a sample of fixed size n, ν can be estimated unbiasedly if and only if it is assumed that $\nu \leq n$. Goodman (1953) showed that ν is unbiasedly estimable when M is fixed, i.e., if sampling stops when m selections not discovering a new class are observed. Some other stopping rules appear in Darling and Robbins (1967), Samuel (1968, 1969) and Freeman (1972, 1973) but they were not concerned with unbiased estimation. The objective of this paper is to discuss a class of sampling plans for which a minimum variance unbiased estimator of ν exists without any further assumption. Thus, our parameter space is $\{\nu : \nu \geq 1\}$.

We shall consider the stopping rules for which the decision of whether or not to continue sampling, after N selections have been made, depends only on N and R, or equivalently on R and M, because for any fixed number of selections, R is a complete sufficient statistic for ν (Harris (1968), Nayak (1992)). Then, from Blackwell (1947), (N, R) or (R, M) are jointly sufficient for ν, where N is the sample size. Note that, N is the total number of selections made during the experiment and it can be a random quantity. For such plans, it is helpful to regard the outcome of a sequence of selections as a path in the (R, M) plane. The paths start at $(0,0)$ and at the jth ($j \geq 1$) step move one unit to the right if the jth selection discovers a new class; otherwise, move one unit up. Clearly, all paths move to the right at the first step. Hence, a sequence of points $\{\gamma_1, \gamma_2, \ldots\}$ in the (R, M) plane, where $\gamma_j = (R_j, M_j)$, $j \geq 1$, defines a path if and only if $\gamma_1 = (1,0)$ and, for all $j \geq 2$, γ_j is either $\gamma_{j-1} + (1,0)$ or $\gamma_{j-1} + (0,1)$, i.e., either $R_j = R_{j-1} + 1$ or $M_j = M_{j-1} + 1$ but not both.

A stopping rule can be regarded as a function ϕ defined on the (R, M) plane such that $\phi(r, m)$ is the probability of continuing sampling when the point (r, m) is reached. We shall consider the non-randomized rules, so that ϕ takes on the value 0 or 1. For a given stopping rule ϕ, a point (r, m) is *accessible* if there exists at least one path $\{\gamma_1, \gamma_2, ...\}$ such that $\gamma_{r+m} = (r, m)$ and $\phi(\gamma_j) = 1$ for $1 \leq j < r + m$. Otherwise, the point (r, m) is *inaccessible*. An accessible point γ is called a *boundary* point if $\phi(\gamma) = 0$; otherwise, it is called a *continuation* point. Then, a stopping rule is characterized by the set B of all of its boundary points.

The preceding development and definitions are similar to the developments in sequential unbiased estimation for binomial sampling, discussed by Girshick, et al. (1946), Lehmann and Stein (1950), DeGroot (1959), Gupta (1967), Sinha and Sinha (1975), Sinha and Bose (1985), and others. Some differences between the two problems may be noted. Unlike in binomial sampling, our sample space (e.g. the range of R) depends on ν, the parameter of interest; and the probabilities of vertical and horizontal movements of a path at any step depend on the point reached prior to that step.

In the next section, the necessary probability distributions are derived. The closure of stopping rules is investigated in Section 3. The closed plans, for which sampling stops with probability 1, are considered in the subsequent sections. Necessary and sufficient conditions for completeness of closed plans are discussed in Section 4. The results are similar to those for the binomial sampling plans. Unbiased estimation of ν is discussed in Section 5. It is shown that ν cannot be estimated unbiasedly if the sample size is bounded. Conversely, unbounded, closed, complete plans admit unique (and hence best) unbiased estimators of all functions of ν and the values of those estimators can be calculated recursively. It is also shown that ν cannot be estimated unbiasedly without any assumptions about the class probabilities.

2. Probability Distribution

In this section we derive the probabilities of the boundary points for any given sampling plan. Given any stopping rule ϕ with boundary B and a point $\gamma \in B$, let $p_\nu(\gamma)$ be the probability, given ν, of observing the point γ. The point $\gamma = (r, m)$ is observed if and only if it is the first boundary point encountered in the sample path. Further, such paths are characterized by the first $(r+m)$ points in the path. Let $A(r,m) = \{\boldsymbol{a} = (\gamma_1, \dots, \gamma_{r+m}) : \gamma_{r+m} = (r,m) \text{ and } \phi(\gamma_j) = 1 \text{ for } 1 \leq j < r+m\}$ so that

$$p_\nu(\gamma) = p_\nu(A(r,m)) = \sum_{\boldsymbol{a} \in A(r,m)} p_\nu(\boldsymbol{a}). \tag{2.1}$$

Since the number of classes discovered cannot be greater than ν, we have $p_\nu(\gamma) = 0$ whenever $\nu < r$. So, consider the case $\nu \geq r$. Any given $\boldsymbol{a} \in A(r,m)$ has exactly r steps to the right and r runs of vertical steps. Let a_j, $1 \leq j \leq r-1$, be the length of the vertical run between the jth and $(j+1)$th discoveries and let a_r be the length of the vertical run between the rth discovery and reaching the point (r,m). Some of the a_j, $1 \leq j \leq r$, may be zero, but $\sum_j a_j = m$. Then

$$\begin{aligned} p_\nu(\boldsymbol{a}) &= p_\nu(\gamma_1) \prod_{j=2}^{r+m} p_\nu(\gamma_j | \gamma_1, \dots, \gamma_{j-1}) \\ &= \left(\frac{\nu}{\nu}\right)\left(\frac{1}{\nu}\right)^{a_1}\left(\frac{\nu-1}{\nu}\right)\left(\frac{2}{\nu}\right)^{a_2} \cdots \left(\frac{\nu-r+1}{\nu}\right)\left(\frac{r}{\nu}\right)^{a_r} \\ &= \frac{\nu(\nu-1)\cdots(\nu-r+1)}{\nu^{r+m}} \prod_{j=1}^{r} j^{a_j}, \end{aligned} \tag{2.2}$$

because $p_\nu(\gamma_1) = 1$ and, for $j \geq 2$,

$$\begin{aligned} p_\nu(\gamma_j | \gamma_1, \dots, \gamma_{j-1}) &= p_\nu(\gamma_j | \gamma_{j-1}) \\ &= \begin{cases} (\nu - r_{j-1})/\nu, & \text{if } \gamma_j = \gamma_{j-1} + (1,0); \\ (r_{j-1})/\nu, & \text{if } \gamma_j = \gamma_{j-1} + (0,1). \end{cases} \end{aligned}$$

Letting $(\nu)_r = \nu(\nu-1)\cdots(\nu-r+1)$ and

$$K(\gamma) = K(r,m) = \sum_{\boldsymbol{a}\in A(r,m)} \left(\prod_{j=1}^{r} j^{a_j}\right),$$

we get from (2.1) and (2.2) that, if $\gamma = (r,m) \in B$,

$$p_\nu(\gamma) = \frac{(\nu)_r}{\nu^{r+m}} K(\gamma). \tag{2.3}$$

It may be noted that $K(\gamma)$ depends on γ and the given boundary but not on ν.

Example 2.1. Suppose the experiment stops after n (fixed) selections so that $B = \{(r,m) : r+m = n\}$. Then, for any $(r,m) \in B$, the sum in $K(r,m)$ is over the set $\{(a_1,\ldots,a_r) : a_j \geq 0, \sum_{j=1}^{r} a_j = m\}$ and it follows that $K(r,m) = S(r+m,r) = S(n,r)$, where $S(n,r)$ is a Stirling number of the second kind and is defined as (Abramowitz and Stegun (1972))

$$S(n,r) = \frac{1}{r!}\sum_{i=0}^{r}(-1)^{r-i}\binom{r}{i} i^n.$$

Example 2.2. For Goodman's (1953) stopping rule with fixed m, $B = \{(r,m) : r \geq 1\}$ and the sum in $K(r,m)$ is over $\{(a_1,\ldots,a_r) : a_r \geq 1,\ a_j \geq 0,\ j = 1,\ldots,r-1, \text{ and } \sum_{j=1}^{r} a_j = m\}$. Here $K(r,m) = rS(r+m-1,r)$ (Harris (1968)).

3. Closed Plans

A stopping rule ϕ with boundary B is said to be *closed* if

$$p_\nu(B) = \sum_{\gamma\in B} p_\nu(\gamma) = 1 \quad \text{for all} \quad \nu \geq 1, \tag{3.1}$$

i.e., the sampling stops with probability 1 for all $\nu \geq 1$. Obviously, only the closed plans are of interest in most applications. To characterize closure, first consider the boundaries with finitely many points. If B is finite, $S = \max\{r+m : (r,m) \in B\}$ exists and is called the *index* of the boundary. The *index* of a point (r,m) is defined as $r+m$, which is the sample size necessary for reaching the point (r,m). The following result can be proved easily.

Proposition 3.1. *Let B be a finite boundary of index S. Then B is closed if and only if there does not exist any continuation point of index S.*

Let C_i and B_i be the sets of continuation and boundary points, respectively, of index i. The continuation sets C_i can be obtained recursively by

$$C_i = \{(r,m) : \text{ either } (r-1,m) \text{ or } (r,m-1) \text{ in } C_{i-1}\} - B_i.$$

To use Proposition 3.1 one needs to obtain C_S and check whether it is empty or not.

For further discussion of closure, let $B(r) = \{(r,m) : (r,m) \in B\}$ and $C(r) = \{(r,m) : (r,m)$ is a continuation point$\}$, i.e., $B(r)$ and $C(r)$ are respectively the sets of boundary and continuation points with $R = r$. Further, let $m_r = \max\{m : (r,m) \in B(r)\}$ and $m_r^* = \max\{m : (r,m) \in C(r)\}$ when $B(r)$ and $C(r)$, respectively, are non-empty and finite. Now, note that when $\nu = 1$, the path $\{\gamma_1, \gamma_2, \ldots\}$ with $\gamma_i = (1, i-1)$, $i \geq 1$, occurs with probability 1. So, (3.1) is true for $\nu = 1$ if and only if $B(1)$ is non-empty. Further, as we require the boundary points to be accessible (by definition), $B(1)$ can have at most one element. Thus, (3.1) is true for $\nu = 1$ if and only if $B(1)$ contains exactly one point, $(1, m_1)$ say, in which case $C(1) = \{(1,m) : m < m_1\}$, and $\{(1,m) : m > m_1\}$ are inaccessible. We assume that $m_1 > 0$ because for $m_1 = 0$ sampling stops after the first selection.

Now suppose $\nu = 2$ and $(1, m_1) \in B$ with $m_1 \geq 1$. Then, $B(2)$ can have at most m_1 points as it can have at most one point $(2,m)$ with $m \geq m_1 - 1$. If $B(2)$ is empty or if $B(2)$ is non-empty but $m_2 < m_1 - 1$, then

$$\begin{aligned}
& P_{\nu=2}(\text{not stopping sampling}) \\
= \; & P_{\nu=2}(\text{reaching } (2, m_1 - 1)) \\
\geq \; & P_{\nu=2}\Big(\{(1,0),(1,1),\ldots,(1,m_1-1),(2,m_1-1)\}\Big) \\
= \; & \left(\frac{1}{2}\right)^{m_1} > 0.
\end{aligned}$$

Conversely, if $B(2)$ is non-empty and $m_2 \geq m_1 - 1$, sampling stops with at most $(m_2 + 2)$ selections when $\nu = 2$. Thus, (3.1) holds for $\nu = 2$ if and only if $B(2)$ is non-empty and $m_2 \geq m_1 - 1$.

Suppose that (3.1) is true for $\nu = 1$ and 2. If $C(2)$ is empty, geometrically (and also by Proposition 3.1) it is clear that the boundary is closed also for all $\nu \geq 2$. So, suppose $C(2)$ is non-empty. Then, $B(3)$ must be finite as it can contain at most one point, $(3,m)$, with $m \geq m_2^*$. As in the previous paragraph, it can now be seen that B is closed for $\nu = 3$ if and only if $B(3)$ is non-empty and $m_3 \geq m_2^*$. Using induction on ν, we can prove the following result.

Proposition 3.2. *A stopping rule ϕ with a boundary B is closed if and only if*
(i) *there exists a unique value m_1 such that $(1, m_1) \in B$, and*
(ii) *$C(r-1)$ is finite, $B(r)$ is non-null and finite, and $m_r \geq m_{r-1}^*$ for all $r \geq 2$ such that $C(r-1)$ is non-empty.*

The following corollaries are easily derived from our preceding discussion.

Corollary 3.1. *Let B be a finite boundary with $r_* = \max\{r : (r,m) \in B\}$. Then B is closed if and only if $C(r_*)$ is empty and conditions* (i) *and* (ii) *of Proposition 3.2 are true for $2 \leq r \leq r_*$.*

Corollary 3.2. *If for all $r \geq 1$, $B(r)$ contains exactly one point, then B is closed if and only if* (i) *holds and $m_r \geq m_{r-1} - 1$ for all $r \geq 2$.*

4. Completeness

A closed stopping rule ϕ with boundary B is said to be *complete* if for any function f defined on B,

$$E_\nu[f(\gamma)] = 0, \quad \text{for all} \quad \nu \geq 1, \tag{4.1}$$

implies that $f(\gamma) = 0$ for all $\gamma \in B$. Completeness of B implies that, based on (R, M), there can be at most one unbiased estimate of ν (or a function of ν). So, if $E_\nu[f(R,M)] = g(\nu)$ for all $\nu \geq 1$, $f(R,M)$ is the minimum variance unbiased estimator of $g(\nu)$. In this section we characterize the closed plans which are complete.

The following definitions are needed: A boundary B is said to be *simple* if for each $n \geq 1$, the continuation points of index n form an interval. A closed boundary B is said to be *minimal* if changing a boundary point to a continuation point destroys closure. A boundary point (r,m) is a *lower boundary point* (LBP) if for some $a > 0$, $(r-a, m+a)$ is a continuation point and is an *upper boundary point* (UBP) if for some $a > 0$, $(r+a, m-a)$ is a continuation point. A boundary point can be either a LBP or an UBP but it is possible that a point in B is neither a LBP nor an UBP; for example, the boundary points in $B = \{(r,m) : r+m = n\}$ are neither LBPs nor UBPs. It is also possible for a point in some B to be both a LBP and an UBP but, if B is simple, then no point is both a LBP and an UBP. For any boundary B, for all $r \geq 1$, $B(r)$ can contain at most one point which is not a LBP since if B contains (r, m_1) and (r, m_2) with $m_1 < m_2$ and both are not LBPs, (r, m_2) would not be accessible. Similar to binomial sampling plans (Girshick, et al. (1946), Savage (1947), Lehmann and Stein (1950)) we shall prove that a closed boundary B is complete if and only if it is both simple and minimal. For convenience and clarity, the results will be presented in several propositions.

Proposition 4.1. *A closed sampling plan ϕ with boundary B is complete only if it is minimal.*

Proof. It is shown that a non-minimal boundary is not complete. Suppose B is not minimal and a boundary point $\alpha = (x,y)$ can be deleted without destroying closure. Let $B^* = B - \alpha$ be the boundary of the plan obtained by deleting α from B. Given ν and $\gamma \in B^*$, let $p_\nu^*(\gamma)$ be the probability of reaching γ under

B^*. Let the probability distribution for B be $p_\nu(\gamma)$, $\gamma \in B$. Note that, if $\gamma \in B^*$, the probability of reaching γ without passing through α under B^* is $p_\nu(\gamma)$. So, for all $\gamma \in B^*$,

$$p^*_\nu(\gamma) = p_\nu(\gamma) + p_\nu(\alpha)p^*_\nu(\gamma|\alpha), \tag{4.2}$$

where $p^*_\nu(\gamma|\alpha)$ is the conditional probability under B^* that a path reaches γ given that it reached α. Hence, the product term in (4.2) is the probability under B^* of reaching γ by passing through α.

To calculate the conditional probability, suppose that $\gamma = (r, m) \in B^*$ can be reached from α. Then $r \geq x$ and $m \geq y$. Let $n = x + y$, $s = r + m$, and

$$B_\alpha(\gamma) = \Big\{\boldsymbol{b} = (\gamma_{n+1}, \ldots, \gamma_s) : \gamma_{n+1} \text{ is } (x+1, y) \text{ or } (x, y+1),\ \gamma_s = (r, m), \\ \text{and } \phi(\gamma_i) = 1, n+1 \leq i < s\Big\}.$$

Note that any $\boldsymbol{b} \in B_\alpha(\gamma)$ has exactly $(r-x)$ steps to the right and $(r-x+1)$ runs of vertical steps, some of which may be of length zero. For any given $\boldsymbol{b} \in B_\alpha(\gamma)$, let $b_x, \ldots, b_r$ be the lengths of the successive runs. Then, similar to the derivation of (2.2), if $\nu \geq r$,

$$p^*_\nu(\boldsymbol{b}|\alpha) = \left(\frac{x}{\nu}\right)^{b_x}\left(\frac{\nu - x}{\nu}\right)\left(\frac{x+1}{\nu}\right)^{b_{x+1}} \cdots \left(\frac{\nu - r + 1}{\nu}\right)\left(\frac{r}{\nu}\right)^{b_r}.$$

Let $L_\alpha(\gamma)$ be 0 if γ cannot be reached from α and otherwise let

$$L_\alpha(\gamma) = \sum_{b \in B_\alpha(\gamma)} \left(\prod_{i=x}^{r} i^{b_i}\right). \tag{4.3}$$

Then, if $\nu \geq r$,

$$\begin{aligned} p^*_\nu(\gamma|\alpha) &= \sum_{b \in B_\alpha(\gamma)} p^*_\nu(\boldsymbol{b}|\alpha) \\ &= \frac{(\nu - x)(\nu - x - 1)\cdots(\nu - r + 1)}{\nu^{s-n}} L_\alpha(\gamma). \end{aligned} \tag{4.4}$$

Combining (2.3), (4.2), and (4.4), we get for all $\nu \geq 1$ and $\gamma = (r, m) \in B^*$,

$$p^*_\nu(\gamma) = g_\nu(\gamma)\Big\{K(\gamma) + K(\alpha)L_\alpha(\gamma)\Big\}, \tag{4.5}$$

where $g_\nu(\gamma) = (\nu)_r/\nu^{r+m}$. Since both B and B^* are closed,

$$\sum_{\gamma \in B} p_\nu(\gamma) = 1 = \sum_{\gamma \in B^*} p^*_\nu(\gamma) \quad \text{for all} \quad \nu \geq 1,$$

which implies, in view of (4.5), that

$$g_\nu(\alpha) = \sum_{\gamma \in B^*} L_\alpha(\gamma) g_\nu(\gamma) \quad \text{for all} \quad \nu \geq 1. \tag{4.6}$$

Now, for $\gamma \in B$, let

$$f(\gamma) = \begin{cases} -1, & \text{if } \gamma = \alpha; \\ \dfrac{K(\alpha) L_\alpha(\gamma)}{K(\gamma)}, & \text{if } \gamma \neq \alpha. \end{cases}$$

Then, by (4.6),

$$E_\nu(f) = \sum_{\gamma \in B^*} K(\alpha) L_\alpha(\gamma) g_\nu(\gamma) - K(\alpha) g_\nu(\alpha) = 0$$

for all $\nu \geq 1$. However, $P_\nu(f \neq 0) > 0$ for all $\nu \geq x$ and hence B is not complete.

Proposition 4.2. *A closed sampling plan ϕ with boundary B is complete only if it is simple.*

Proof. Suppose B is not simple and for some $a \geq 2$, $b \geq 1$, and $k \geq 0$, $(a-1, b+1)$ and $(a+k+1, b-k-1)$ are two continuation points of index $n = a+b$ separated by the boundary points $A = \{\gamma_i = (a+i, b-i), i = 0, 1, \ldots, k\}$. Then $\alpha' = (a, b+1)$ and $\alpha'' = (a+k+1, b-k)$ are accessible points. We shall construct a nonzero function on B with expected value 0 for all $\nu \geq 1$. To do that first note that the quantity $L_\alpha(\gamma)$ in (4.3) is defined for any continuation point α and all $\gamma \in B$. Since α', α'' may be either continuation or boundary points, if $\alpha \in B$, further define $L_\alpha(\alpha) = 1$ and $L_\alpha(\gamma) = 0$, $\gamma \in B$. Let $p_\nu(\alpha)$ denote the probability that a path passes through (or reaches) α. Since B is closed we have

$$p_\nu(\alpha) = \sum_{\gamma \in B} p_\nu(\alpha \cap \gamma), \tag{4.7}$$

where $p_\nu(\alpha \cap \gamma)$ is the probability under B that a path passes through α and reaches γ. Then, (4.7) and arguments similar to those given in Section 2 and in the proof of Proposition 4.1 give us

$$g_\nu(\alpha) = \sum_{\gamma \in B-A} L_\alpha(\gamma) g_\nu(\gamma) \tag{4.8}$$

for all $\nu \geq 1$ and $\alpha = \alpha', \alpha''$.

Now, for $\gamma_i \in A$, let

$$f(\gamma_i) = \frac{(-1)^i}{(a+i)! K(\gamma_i)}, \quad i = 0, \ldots, k, \tag{4.9}$$

and for $\gamma \in B - A$, let

$$f(\gamma) = \frac{-L_{\alpha'}(\gamma)}{(a-1)!K(\gamma)} + \frac{(-1)^{k+1}L_{\alpha''}(\gamma)}{(a+k)!K(\gamma)}. \tag{4.10}$$

Then, using (4.8)–(4.10),

$$\begin{aligned}
E_\nu[f] &= \sum_{\gamma \in B} f(\gamma)K(\gamma)g_\nu(\gamma) \\
&= \sum_{i=0}^{k} \frac{(-1)^i g_\nu(\gamma_i)}{(a+i)!} - \frac{1}{(a-1)!}\sum_{\gamma \in B-A} L_{\alpha'}(\gamma)g_\nu(\gamma) + \frac{(-1)^{k+1}}{(a+k)!}\sum_{\gamma \in B-A} L_{\alpha''}(\gamma)g_\nu(\gamma) \\
&= \sum_{i=0}^{k} \frac{(-1)^i g_\nu(\gamma_i)}{(a+i)!} - \frac{g_\nu(\alpha')}{(a-1)!} + \frac{(-1)^{k+1}g_\nu(\alpha'')}{(a+k)!} \\
&= \frac{1}{\nu^n}\left\{\sum_{i=0}^{k}(-1)^i\binom{\nu}{a+i} - \binom{\nu-1}{a-1} + (-1)^{k+1}\binom{\nu-1}{a+k}\right\}
\end{aligned} \tag{4.11}$$

for all $\nu \geq 1$. If $\nu < a$, (4.11) is clearly 0. For $\nu \geq a$, (4.11) can be seen to be zero by replacing $\binom{\nu}{a+i}$ with $\binom{\nu-1}{a+i-1} + \binom{\nu-1}{a+i}$. This completes the proof of the proposition.

To prove the converses of Propositions 4.1 and 4.2, we shall consider infinite and finite boundaries separately and in that process explain some unique properties of the two types of boundaries.

Proposition 4.3. *A closed infinite boundary is simple and minimal if and only if for all $r \geq 1$, $B(r)$ contains exactly one point.*

Proof. Since B is infinite and closed, Proposition 3.2 implies that for all $r \geq 1$, $B(r)$ and $C(r)$ must be non-empty but finite and $m_r \geq m^*_{r-1}$. Further, if $B(r)$ contains exactly one point for all $r \geq 1$, we must have $m_r \geq m_{r-1} - 1$ for all $r \geq 2$, in which case it is geometrically clear that B is minimal and simple.

To prove the "only if" part, let B be closed, infinite, simple and minimal, and if possible, for some $a \geq 2$, let $B(a)$ contain more than one point. Let $m_0 = \min\{m : (a, m) \in B(a)\}$. Then, (a, m_a) and (a, m_0) are in $B(a)$ and $m_0 < m_a$. Further, (a, m_0) is a LBP, otherwise (a, m_a) is not accessible. So (a, m_0) cannot be an UBP as B is simple. This implies that for all $r \geq a$, the m-coordinates of the points in $B(r)$ and $C(r)$ are greater than m_0. So, if we delete the point (a, m_0), the values of m_r and m^*_r remain unchanged for all $r \geq 1$ and hence the plan remains closed. This implies that B is not minimal, which is a contradiction and the proposition is proved.

Proposition 4.4. *An infinite closed boundary B is complete if it is simple and minimal.*

Proof. In view of Proposition 4.3, $B = \{(r, m_r) : r \geq 1\}$. Let $f(r, m_r)$ be a function on B such that

$$E_\nu(f) = \sum_{r=1}^{\nu} f(r, m_r) \frac{(\nu)_r}{\nu^{r+m_r}} K(r, m_r) = 0 \quad \text{for all} \quad \nu \geq 1. \tag{4.12}$$

Clearly, (4.12) is true for $\nu = 1$ if and only if $f(1, m_1) = 0$. Now the proposition can be proved using induction on ν.

From Propositions 4.1–4.4 and Corollary 3.2, an infinite boundary is closed and complete if and only if for all $r \geq 1$, $B(r)$ contains exactly one point, say (r, m_r), and $m_{r+1} \geq m_r - 1$. It is shown in the next section that such plans admit best unbiased estimators of ν. Clearly, Goodman's (1953) stopping rule (Example 2.2) satisfies these conditions. Two other stopping rules falling into this category use, respectively, $m_r = \inf\{m : m \geq rC\}$ and $m_r = \inf\{m : m \geq \max(1, r \ln r + rD)\}$, where $C > 0$ and $-\infty < D < \infty$ are fixed numbers. Motivation and some properties of these rules are discussed in Darling and Robbins (1967) and Samuel (1968, 1969).

Proposition 4.5. *A necessary condition for a closed finite boundary to be simple is that it be minimal.*

Proof. Let B be closed, finite and simple and let $r_* = \max\{r : (r, m) \in B\}$. If possible, suppose B is not minimal and remains closed if $\alpha = (x, y) \in B$ is deleted. Let $B^* = B - \alpha$. Then there exists $m > y$ such that $\gamma_1 = (x, m) \in B$; otherwise for $\nu = x$, the sampling will not stop under B^* whenever α is reached. Also, there must exist $r > x$ such that $\gamma_2 = (r, y) \in B$; otherwise, when $\nu > r_*$, the probability under B^* of not stopping sampling is greater than

$$\begin{aligned} & p_\nu(\alpha) p_\nu(\{\gamma_{x+y+i} = (x + i, y), i = 1, \ldots, r_* - x + 1\} | \alpha) \\ = \ & p_\nu(\alpha) \prod_{i=0}^{r_*-x} \left(\frac{\nu - x - i}{\nu} \right) > 0. \end{aligned}$$

However, accessibility of γ_1 and γ_2 implies that α is both LBP and UBP. Hence, B is not simple. This contradiction proves the result.

Proposition 4.6. *If a closed finite boundary is simple then it is complete.*

Proof. If possible, suppose B is a closed, finite and simple boundary but it is not complete. Let f be a function on B such that $f \neq 0$ for some $\gamma \in B$ and (4.1) is satisfied. First we show that f must be 0 for every LBP in B. If not, let n be the smallest index for which a LBP with $f \neq 0$ exists and let $\alpha = (x, y)$ be the lowest LBP of index n for which $f \neq 0$, i.e., $y = \min\{m : (n-m, m)$ is a LBP and $f(n-m, m) \neq 0\}$ and $x = n - y$. Then $f(r, m) = 0$ if $(r, m) \in B$ and $(r+m) < n$. To see this, suppose $A = \{(r, m) : f(r, m) \neq 0,\ (r, m) \in B$, and $(r+m) < n\}$ is non-empty and let $a = \min\{r : (r, m) \in A\}$. Within A there can exist only one point with $R = a$ because, from the definition of α, the points in A are not LBPs, and there can exist at most one point with $R = a$ which is not a LBP. Let that point be $\gamma_1 = (a, b)$. Since γ_1 is not a LBP, $\{(r, m) : r \leq a, r + m \geq a + b\}$ and $\{(a, m) : m > b\}$ are inaccessible. So, the definitions of α and γ_1 imply that, in $\cup_{r \leq a} B(r)$, f is nonzero only at γ_1. Then, for $\nu = a$, (4.12) can be true only if $f(\gamma_1) = 0$. So, $f(r, m) = 0$ if $(r + m) < n$.

Since B is simple, α is not an UBP and the m-coordinates of the boundary points of index $> n$ are greater than y. Further, since $f(r, m) = 0$ whenever $(r + m) < n$, $f(r, m)$ can be non-zero only if $(r + m) \geq n$ and $m > y$. So, (4.1) implies that

$$-f(x, y)\frac{(\nu)_x}{\nu^{x+y}}K(x, y) = \sum_{\substack{(r,m)\in B\\ m>y,\, r+m\geq n}} f(r, m)\frac{(\nu)_r}{\nu^{r+m}}K(r, m),$$

or

$$-f(x, y)\frac{(\nu)_x}{\nu^{x}}K(x, y) = \frac{1}{\nu} \sum_{\substack{(r,m)\in B\\ m>y, r+m\geq n}} f(r, m)\left\{\frac{(\nu)_r}{\nu^{r}}\right\}\frac{1}{\nu^{m-y-1}}K(r, m), \quad (4.13)$$

for all $\nu \geq 1$. The right side of (4.13) converges to 0 as $\nu \to \infty$ and the left side converges to $-f(x, y)K(x, y)$. So, (4.13) can be true for all $\nu \geq 1$ only if $f(x, y) = 0$. This contradiction in the definition of n shows that f is 0 for all LBPs.

Since, for each $r \geq 1$, $B(r)$ can contain at most one point which is not a LBP, now the fact that f must be 0 also for all boundary points which are not LBPs can be established by induction on ν as in Proposition 4.4.

Remark. The boundary of a fixed sample size experiment (Example 2.1) is clearly simple and satisfies the conditions of Proposition 3.2. So, it is closed and complete.

5. Unbiased Estimation

In view of sufficiency, given any closed sampling plan, its boundary B can be taken as the sample space for estimation purposes. Then, an estimator is a function f defined on B. Given a boundary B and a function $g(\nu)$ of ν, we address the question: Is $g(\nu)$ unbiasedly estimable from B, i.e. does there exist a function f defined on B such that $E_\nu[f(\gamma)] = g(\nu)$ for all $\nu \geq 1$?

Proposition 5.1. *If B is finite, an unbiased estimator of ν does not exist.*

Proof. Let S be the index of B, i.e., let $S = \max\{r+m : (r,m) \in B\}$. If possible, let $f(\gamma)$ be an unbiased estimator of ν. Then,

$$E_\nu[f(\gamma)] = \sum_{(r,m)\in B} f(r,m)\frac{(\nu)_r}{\nu^{r+m}}K(r,m) = \nu$$

or

$$\sum_{(r,m)\in B} f(r,m)(\nu)_r \nu^{S-r-m}K(r,m) = \nu^{S+1} \quad \text{for all} \quad \nu \geq 1. \tag{5.1}$$

However, (5.1) cannot be true because the left side is a polynomial in ν of degree at most S. So, ν cannot be estimated unbiasedly.

Proposition 5.1 also follows from Engen (1978, Theorem 2.2, p.28). More generally, if B is finite, any unbounded function of ν is not unbiasedly estimable.

Proposition 5.2. *If B is infinite, all functions of ν are unbiasedly estimable. Further, if B is complete (implying that $B = \{(r, m_r) : r \geq 1\}$), the unbiased estimator of any given function $g(\nu)$ is unique and can be calculated recursively by $f(1, m_1) = g(1)$ and*

$$f(r,m_r) = \frac{1}{p_r(r,m_r)}\left(g(r) - \sum_{i=1}^{r-1} f(i,m_i)p_r(i,m_i)\right), \quad r \geq 2. \tag{5.2}$$

Proof. Since B is closed and infinite, for all $r \geq 1$, $B(r)$ is non-empty and finite and $m_r = \max\{m : (r,m) \in B(r)\}$ exists. For estimating $g(\nu)$ let $f(r,m) = 0$ if $m \neq m_r$ and define $f(r, m_r)$, $r \geq 1$, recursively by $f(1, m_1) = 1$ and (5.2). Then it can be seen that $f(\gamma)$ is an unbiased estimator of $g(\nu)$.

Further, if B is complete, from Section 4, $B(r)$ contains exactly one point (r, m_r) for all $r \geq 1$ and hence $f(\gamma)$ as defined in the proposition is the unique unbiased estimator of $g(\nu)$.

In the following, we show that the recursive formulas reduce to ratios of some $K(r,m)$ values for certain powers of ν.

Proposition 5.3. *Let B be an infinite, closed, complete boundary and let $\tilde{m} = \min\{m_r : r \geq 1\}$. Then for all integers $p > -\tilde{m}$, the minimum variance unbiased estimator of ν^p is*

$$f(r,m_r) = \frac{K^*(r,m_r+p)}{K(r,m_r)}, \quad r \geq 1, \tag{5.3}$$

where K^* is the K function for the boundary $B^* = \{(r, m_r + p) : r \geq 1\}$ obtained by shifting up all the points in B by p units.

Proof. Let $p_\nu^*(\gamma)$, $\gamma \in B^*$, denote the probability distribution for B^*, so that

$$p_\nu^*(r, m_r + p) = \frac{(\nu)_r}{\nu^{r+m_r+p}} K^*(r, m_r + p), \quad r \geq 1.$$

From Corollary 3.2 it follows that B^* is closed. Hence,

$$\begin{aligned} E_\nu[f(r, m_r)] &= \sum_{r=1}^{\nu} \frac{K^*(r, m_r + p)}{K(r, m_r)} \frac{(\nu)_r}{\nu^{r+m_r}} K(r, m_r) \\ &= \nu^p \sum_{r=1}^{\nu} \frac{(\nu)_r}{\nu^{r+m_r+p}} K^*(r, m_r + p) \\ &= \nu^p \end{aligned}$$

for all $\nu \geq 1$. The rest follows from the Lehmann-Scheffe theorem (Bickel and Doksum (1977, p.122)).

Remarks. 1. By Proposition 5.3, an unbiased estimate of the variance of the best unbiased estimator of ν is given by

$$\left(\frac{K^*(r, m_r + 1)}{K(r, m_r)}\right)^2 - \frac{K^*(r, m_r + 2)}{K(r, m_r)}.$$

2. For $p = -\tilde{m}$ the boundary of the shifted plan is not really $\{(r, m_r - \tilde{m}) : r \geq 1\}$ as some of these points are inaccessible. The true boundary is $\{(r, m_r - \tilde{m}) : r \leq r_0\}$ where $r_0 = \min\{r : m_r = \tilde{m}\}$. Proposition 5.3 remains valid for $p = -\tilde{m}$ if (5.3) is replaced by $f = 0$ for $r > r_0$.

3. When B is closed and infinite but not complete the approach in Proposition 5.3 yields an unbiased estimator of ν^p if $p > -\min\{m : (r, m) \in B$ for some $r \geq 1\}$. There, we define the ratio (5.3) for all $(r, m) \in B$ after shifting all the points in B up by p units.

4. For Goodman's (1953) stopping rule, given in Example 2.2, not only ν but all functions of ν are unbiasedly estimable by Proposition 5.2. Further, for all integers $p > -m$, the best unbiased estimator of ν^p, by Proposition 5.3, is $S(r + m + p - 1, r)/S(r + m - 1, r)$; and an unbiased estimate of the variance of the best unbiased estimator of ν is given by

$$\frac{S(r + m, r)}{S(r + m - 1, r)} \left(\frac{S(r + m, r)}{S(r + m - 1, r)} - \frac{S(r + m + 1, r)}{S(r + m, r)}\right).$$

Berg (1975) gives recursive formulas for calculating these ratios.

5. Propositions 5.1 and 5.2 show that for unbiased estimation the boundary needs to be infinite. That is also true for maximum likelihood estimation. The MLE of ν is ∞ if and only if the observed value of M is 0 (e.g., Samuel (1969)). So, the MLE of ν is finite with probability 1 for all $\nu \geq 1$ only if M is nonzero for all boundary points in which case the boundary must be infinite if it is closed.

Throughout this paper we have assumed that all the classes are equally likely. This assumption is made not only for simplicity but also for some necessity, as discussed in the following. In the unequal case, for any given $\nu \geq 1$, let $p_1, p_2, \ldots, p_\nu$ be the probabilities of the ν classes. We shall show that ν cannot be estimated unbiasedly without any restrictions on $p_1, p_2, \ldots, p_\nu$. Now, for a fixed number of selections, R is not sufficient and hence we shall consider a more general setup. As in Nayak (1992), the outcome of a sequence of selections can be described using $G_1, G_2, \ldots$, where $G_i = j$ if the ith observation is a member of the jth discovered class. We shall consider all closed non-randomized stopping rules where, at the nth stage, the stopping probability is a function of $(G_1, G_2, \ldots, G_n)$. Thus, the stopping rules may be path dependent. The sample space $\mathcal{S}$ is the collection of all observable sequences, possibly of different lengths. We shall use R to denote the number of discovered classes in an observable sequence. Now we state our main result.

Proposition 5.4. *Without any assumption about the class probabilities, ν cannot be estimated unbiasedly.*

Proof. If possible, let f be a function on $\mathcal{S}$ such that

$$E[f] = \nu \quad \text{for all } \nu \geq 1 \text{ and } p_1, \ldots, p_\nu. \tag{5.4}$$

Let $\mathcal{S}_r$, $r \geq 1$, be the set of all observable sequences for which $R = r$. Considering $\nu = 1$ the closure of the plan implies that $\mathcal{S}_1$ contains a unique element, say δ_1. Let the sample size for δ_1 be N_1, i.e., if the first N_1 selections result in the same class type, the experiment stops. Again considering $\nu = 1$, (5.4) implies that $f(\delta_1)$ must be 1.

Now consider the case $\nu = 2$. Let p and $q = (1-p)$, $0 < p < 1$, denote the probabilities of the two classes and let $p(\delta)$, $\delta \in \mathcal{S}_1 \cup \mathcal{S}_2$, be the corresponding probability distribution. Note that the points outside $\mathcal{S}_1 \cup \mathcal{S}_2$ cannot be observed when $\nu = 2$. Since $p(\delta_1) = p^{N_1} + (1-p)^{N_1}$, (5.4) implies that

$$p^{N_1} + (1-p)^{N_1} + \sum_{\delta \in \mathcal{S}_2} f(\delta) p(\delta) = 2$$

or

$$p^{N_1} + \sum_{i=1}^{N_1} \binom{N_1}{i} (-1)^i p^i + \sum_{\delta \in \mathcal{S}_2} f(\delta) p(\delta) = 1 \tag{5.5}$$

for all $0 < p < 1$. The probability of observing any given $\delta \in S_2$ is $p^a q^b + p^b q^a$, where a (b) is the number of 1s (2s) in δ. Since a, $b \geq 1$, the probability is a polynomial in p without a constant term. Thus, the left side of (5.5) is a polynomial (or a power series) in p with the constant term being equal to 0. So, (5.5) cannot be true for all $0 < p < 1$ and hence ν cannot be estimated unbiasedly. It is clear from the proof that the proposition remains true even when ν has an upper bound $\nu_0 \geq 2$.

6. Discussion

A continuous analogue of our discrete model, discussed in Nayak (1991), is a superposition of an unknown number (ν) of independent homogeneous Poisson processes with unknown rates $\lambda_1, \lambda_2, \ldots, \lambda_\nu$ where, for each event, the component process in which it occurred can be identified. The observable process can be described as a marked point process $\{W_i, G_i\}$ where $\{W_i\}$ are the event times and the marks $\{G_i\}$ are as defined in Section 5, i.e., $G_i = j$ if the ith event occurs in the jth detected process. It follows that $\{W_i\}$ is a homogeneous Poisson process with rate $\Lambda = \sum \lambda_i$ and $\{W_i\}$ and $\{G_i\}$ are independent. In the discrete case only the mark process $\{G_i\}$ is relevant (see Section 5) and its distribution is the same for both the discrete and the continuous models, with $p_i = \lambda_i/\Lambda$, $i = 1, \ldots, \nu$. The equiprobability case corresponds to the case of $\lambda_1 = \cdots = \lambda_\nu$. If the total number of events N is determined only by the $\{G_i\}$, Nayak and Christman (1992) showed that the inference problems about ν coincide in the two cases. Thus, the results of this paper apply also to the continuous model if the experiment is stopped at the epoch of an event determined only by the $\{G_i\}$ and the decision to stop depends only on the sufficient statistics (R, M). An example of such an experiment is the likelihood-based stopping rule suggested by Goudie (1990). The boundary of Goudie's rule is given by $B = \{(r, m_r) : r \geq 1, m_r = \text{integer value of } \min(m : m \geq 1 - r + \ln(rA)/\ln(1 + r^{-1}))\}$ where A is a given constant. Since B contains one point for each $R = r$ and m_r is a nondecreasing function of r, by Corollary 3.2 and Propositions 4.3 and 4.4, B is closed and complete. Hence, although originally derived for a different purpose, Goudie's stopping rule does admit best unbiased estimators of functions of ν.

In this paper we consider the set of stopping rules which depend only on the sufficient statistics, R and N. Some existing stopping rules, although designed to achieve some sense of optimality, do not belong to this set. For example, the 'step-wise minimal' rule of Darling and Robbins (1967) and Nayak's (1988) rule, derived for detecting all classes with a prespecified probability, depend on R and the lengths of runs of vertical movements (or waiting times for discovering a new class). Specifically, if the rth ($r \geq 1$) discovery is followed by a certain number, k_r, of selections representing already discovered classes, the experiment stops.

The values of k_r, $r \geq 1$, are calculated differently in these two rules. For these plans, the concept of a boundary does not apply because any fixed point (r, m) may or may not be a stopping point depending on how that point is reached. Investigation of such path dependent plans is a topic of future research. The adaptive stopping rule of Rasmussen and Starr (1979) depends on R and the number of classes observed exactly once. This plan utilizes some information (viz. number of classes observed exactly once) not contained in the path (as defined in this paper). For a formal discussion of such plans, the general framework that preceeds our Proposition 5.4 is needed.

Our results are derived without imposing any restrictions on ν. In some applications, however, ν may have natural upper and/or lower bounds. In restricted problems where the range of ν is a proper subset of the positive integers, the definitions of closure, completeness and unbiasedness will naturally be modified. Some restricted problems are currently under investigation. Comparison of various sampling plans is also an important topic for future research. The variance of estimators and the average sample size both should be taken into consideration when evaluating a stopping rule. In view of Samuel's (1969) work we anticipate that to increase the accuracy of the estimators it will be necessary to increase the average sample size. So, some criteria of combining these two performance measures will be necessary if one wants to select an optimum stopping rule.

The unbiased estimators of ν derived under the equiprobability assumption are usually negatively biased when that assumption fails (Nayak and Christman (1992)). So, it is important to derive more robust estimators with smaller bias in non-equiprobable cases. Chao (1984, 1987), Chao and Lee (1992), and others have made some significant contributions in that direction. But, as Proposition 5.4 shows, an unbiased estimator of ν for the general case cannot be obtained. As an alternative to unbiased estimation, estimators from any given sampling plan may be compared by their mean square errors. Interestingly, Bai and Chow (1991) have proved recently that, for Goodman's plan, the maximum likelihood estimator of ν is inadmissible.

Acknowledgement

We thank Professor Bikas K. Sinha for some discussions that inspired us to do this research. We also thank a referee for some helpful suggestions.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables.* Dover Publ. Inc., New York.

Bai, Z. D. and Chow, M. (1991). Inadmissibility of the maximum likelihood estimator in the sequential estimation of the size of a population. *Biometrika* 78, 817–823.

Berg, S. (1975). Some properties and applications of a ratio of Stirling numbers of the second kind. *Scand. J. Statist.* **2**, 91–94.

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics.* Holden-Day, California.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Ann. Math. Statist.* **18**, 105–110.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Statist.* **11**, 265–270.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–791.

Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.* **87**, 210–217.

Darling, D. A. and Robbins, H. (1967). Finding the size of a finite population. *Ann. Math. Statist.* **38**, 1392–1398.

DeGroot, M. H. (1959). Unbiased sequential estimation for binomial populations. *Ann. Math. Statist.* **30**, 80–101.

Engen, S. (1978). *Stochastic Abundance Models, With Emphasis on Biological Communities and Species Diversity.* Halsted Press, John Wiley, New York.

Freeman, P. R. (1972). Sequential estimation of the size of a population. *Biometrika* **59**, 9–17.

Freeman, P. R. (1973). A numerical comparison between sequential tagging and sequential recapture. *Biometrika* **60**, 499–508.

Girshick, M. A., Mosteller, F. and Savage, L. J. (1946). Unbiased estimates for certain binomial sampling problems with applications. *Ann. Math. Statist.* **17**, 13–23.

Goodman, L. A. (1953). Sequential sampling tagging for population size problems. *Ann. Math. Statist.* **24**, 56–69.

Goudie, I. B. J. (1990). A likelihood-based stopping rule for recapture debugging. *Biometrika* **77**, 203–206.

Gupta, M. K. (1967). Unbiased estimate for $1/p$. *Ann. Inst. Statist. Math.* **19**, 413–416.

Harris, B. (1968). Statistical inference in the classical occupancy problem: Unbiased estimation of the number of classes. *J. Amer. Statist. Assoc.* **63**, 837–847.

Lehmann, E. L. and Stein, C. (1950). Completeness in the sequential case. *Ann. Math. Statist.* **21**, 376–385.

Nayak, T. K. (1988). Estimating population size by recapture sampling. *Biometrika* **75**, 113-120.

Nayak, T. K. (1991). Estimating the number of component processes of a superimposed process. *Biometrika* **78**, 75–81.

Nayak, T. K. (1992). On statistical analysis of a sample from a population of unknown species. *J. Statist. Plann. Inference* **31**, 187–198.

Nayak, T. K. and Christman, M. C. (1992). Effect of unequal catchability on estimates of the number of classes in a population. *Scand. J. Statist.* **19**, 281–287.

Rasmussen, S. L. and Starr, N. (1979). Optimal and adaptive stopping in the search for new species. *J. Amer. Statist. Assoc.* **74**, 661-667.

Samuel, E. (1968). Sequential maximum likelihood estimation of the size of a population. *Ann. Math. Statist.* **39**, 1057–1068.

Samuel, E. (1969). Comparison of sequential rules for estimation of the size of a population. *Biometrics* **25**, 517–527.

Savage, L. J. (1947). A uniqueness theorem for unbiased sequential binomial estimation. *Ann. Math. Statist.* **18**, 295–297.

Sinha, B. K. and Bose, A. (1985). Unbiased sequential estimation of $1/p$: Settlement of a conjecture. *Ann. Inst. Statist. Math.* **37**, 455–460.

Sinha, B. K. and Sinha, B. K. (1975). Some problems of unbiased sequential binomial estimation. *Ann. Inst. Statist. Math.* **27**, 245–258.

Department of Statistics/Computer and Information Systems, George Washington University,
Washington, DC 20052, U.S.A.

(Received April 1992; accepted July 1993)