Statistica Sinica **6**(1996), 997-1004

FURTHER ASPECTS OF SEQUENTIAL UNBIASED ESTIMATION OF THE NUMBER OF CLASSES

Mary C. Christman

American University

Abstract: We consider unbiased estimation of the number of classes, ν, in a population where the classes are equally likely to occur and the parameter space of the number of classes is bounded from above. 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 unbounded, but some finite and closed plans admit unbiased estimators of all functions of ν. A general rule for obtaining such estimators is given.

Key words and phrases: Closed sampling plan, completeness, restricted parameter space, stopping rule.

1. Introduction

In a recent paper, Christman and Nayak (1994) considered sequential unbiased estimation of the number of classes, ν, in a population assuming that the parameter space of ν is unrestricted, i.e. $\{\nu : \nu \geq 1\}$. When sampling stops according to some non-randomized stopping rule based on a minimal sufficient statistic, they showed that the number of classes cannot be unbiasedly estimated if the sample size is bounded. On the other hand, unbounded sampling plans admit unbiased estimators of all functions of ν. They assumed throughout that all classes are equally likely to occur in a random selection.

In many applications, the unrestricted parameter space may be too general and a natural upper bound for the number of classes may be known. For example, in a study of the number of distinct animals in a particular species within a geographic region it may be known that the ecosystem can support no more than some given maximum number of animals, ν_0 say. (Note that here each animal is its own class.) Hence, it is of interest to consider how the restriction of the parameter space alters the characterization of plans that admit unbiased estimators of functions of ν.

Our development of unbiased estimation of ν follows that for the unrestricted parameter space as given in Christman and Nayak (1994). We refer the reader to their paper for the formulation of the problem and the relevant definitions.

We also assume, as in Goodman (1953), Harris (1968) and others, that all classes are equally likely to occur. This assumption is unlikely to hold in practice but is assumed here for mathematical tractability. For some stopping rules, it has been shown (Nayak and Christman (1992)) that unbiased estimators of ν derived under the assumption of equiprobability are negatively biased when the assumption fails. The magnitude of the bias depends on the severity of the deviation from equiprobability in that the more dissimilar the classes are in their "catchability" the more negatively biased the estimators are. Most of the previously studied estimators appear to behave reasonably well, i.e. are not too severely biased, when the classes have similar probabilities of being observed in a random selection (cf. Chao and Lee (1992), Nayak and Christman (1992)). Some preliminary investigations are currently underway to evaluate stopping rules that might yield estimators that are more robust to failure of the equally likely assumption than those previously studied.

The following notation is used. Let R (M) be the number of selections in which a new class is (not) discovered; then, $R + M = N$ is the total number of selections. Note that N can be a random quantity depending on the stopping rule used. Only non-randomized stopping rules based on the sufficient statistics (R, M) are considered here. The set of all points (r, m) where the rule is to stop sampling is referred to as the boundary, denoted by B, of the sampling plan.

In the next section, the relevant probability distributions are given and closure and completeness of stopping rules are briefly discussed. The results are similar to those for the unrestricted parameter space. Unbiased estimation of functions of ν is discussed in Section 3. It is shown that if the cardinality of the boundary set is less than ν_0 then ν is not unbiasedly estimable. Conversely, closed, complete plans with a boundary whose cardinality equals ν_0 admit a unique (and hence best) unbiased estimator of all functions of ν and such estimators can be calculated recursively. We also give conditions under which some sampling plans that are not complete admit unbiased estimators of functions of ν.

2. Closure and Completeness

Given a 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. Using the notation and definitions in Christman and Nayak (1994), the probability of observing $\gamma = (r, m) \in B$ given ν is

$$p_\nu(\gamma) = \frac{(\nu)_r}{\nu^{r+m}} K(r, m), \tag{2.1}$$

where $K(r,m)$ is a combinatorial quantity related to the set of paths that reach $\gamma = (r,m)$ and $(\nu)_r = \nu(\nu-1)\cdots(\nu-r+1)$ is defined to be zero if $r > \nu$. We note that $K(r,m)$ depends on γ and B but not on ν.

For closure of plans, a stopping rule with boundary B is said to be *closed* if sampling stops with probability 1 for all $\nu \leq \nu_0$. The necessary and sufficient conditions for closure of boundaries that hold when the parameter space of ν is unrestricted, i.e. when $\{\nu : \nu \geq 1\}$, must also hold for each ν in the subset $1 \leq \nu \leq \nu_0$. Hence, the conditions for closure of plans for the restricted parameter space (Propositions 3.1 and 3.2 and associated corollaries, Christman and Nayak (1994)) are analogous to those for the unrestricted case with only minor modifications due to the finite parameter space. For example, statements such as "for all $r \geq 2$" are replaced with "for all $2 \leq r \leq \nu_0$".

We note that since the parameter space of ν is restricted from above, condition (ii) of Proposition 3.2 of Christman and Nayak (1994) implies that if B has infinitely many points then B cannot be closed. Hence, in the subsequent sections we consider only closed sampling plans with finitely many points.

For completeness of sampling plans, a closed plan with boundary B is said to be *complete* if for any function f defined on B, $\mathrm{E}_\nu[f(\gamma)] = 0$ for all $\nu \leq \nu_0$ implies that $f(\gamma) = 0$ for all $\gamma \in B$. For the following, let $|B|$ be the cardinality of closed B; since only boundaries with finitely many points can be closed, we must have $|B| < \infty$.

Proposition 2.1. *A closed sampling plan with boundary B is not complete if $|B| > \nu_0$.*

Proof. When $|B| = s > \nu_0$, $\mathrm{E}_\nu[f(\gamma)] = 0$ for all $\nu \leq \nu_0$ defines a set of ν_0 linear homogeneous equations with $s > \nu_0$ unknowns. Hence, an infinite number of solutions exist and B is not complete.

For further discussion, we need the following definitions. Any point (r,m) that could be observed during sampling and for which the stopping rule is to continue sampling is called a *continuation* point. The index of the point, (r,m), is $r+m$ and is the sample size necessary for observing the point. 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.

Similar to the case where the parameter space is unrestricted we can show that a closed finite boundary is complete if and only if it is both simple and minimal but, in view of Proposition 2.1, with the added proviso that the plan have $|B| \leq \nu_0$. We refer the reader to Propositions 4.2, 4.5 and 4.6 of Christman and Nayak (1994) which are also true in the restricted case as long as the boundary under consideration has cardinality of no more than ν_0.

3. Unbiased Estimation

For a sampling plan, the boundary B can be taken as the sample space so that an estimator of a function of ν is a function f defined on B. Given a function, $g(\nu)$ say, we address the question: Is $g(\nu)$ unbiasedly estimable from B, i.e. does there exist a function f defined on B such that $\mathrm{E}_\nu[f(\gamma)] = g(\nu)$ for all $\nu \leq \nu_0$? We begin by giving some conditions under which all functions of ν, $g(\nu)$, are unbiasedly estimable and give formulas for calculating the estimators. Let $|B| = s$ and define $B(r) = \{(r,m) : (r,m) \in B\}$, i.e. $B(r)$ is the set of boundary points with $R = r$. Also, let $m_r = \max\{m : (r,m) \in B(r)\}$.

Proposition 3.1. *Let B be a closed boundary with $s \geq \nu_0$ and $r_* = \max\{r : (r,m) \in B\} = \nu_0$. Then all functions of ν, $g(\nu)$, are unbiasedly estimable and can be calculated recursively by $f(1,m_1) = g(1)$ and*

$$\begin{aligned} f(r,m_r) &= \frac{1}{p_r(r,m_r)}\Big(g(r) - \sum_{i=1}^{r-1} f(i,m_i)p_r(i,m_i)\Big), \quad 2 \leq r \leq \nu_0, \\ f(r,m) &= 0, \quad 2 \leq r \leq \nu_0,\ m < m_r. \end{aligned} \tag{3.1}$$

Further, if B is complete (implying that $|B| = \nu_0$), the unbiased estimator of $g(\nu)$ is unique.

Proof. Since B is closed with $s \geq \nu_0$ and $r_* = \nu_0$, $B(r)$ is non-empty and finite and m_r exists for each $r \leq \nu_0$. Then it is easy to show that $f(\gamma)$ as given in the Proposition is an unbiased estimator of $g(\nu)$.

Further, if B is complete, $|B| = \nu_0$ so that $B(r)$ contains exactly one point (r, m_r) for all $r \leq \nu_0$. Hence, f as defined in the proposition is the unique unbiased estimator of $g(\nu)$.

These results are similar to some of those for the case of the unrestricted parameter space with one important difference. When the parameter space is unrestricted, only infinite closed boundaries admit an estimator which is unbiased for $g(\nu)$. The condition of infinitely many points for the unrestricted space is replaced here with the conditions that $\nu \leq |B| < \infty$ and $r_* = \nu_0$.

For further discussion of unbiased estimation we require some definitions and notation. Uniquely number each $\gamma \in B$ so that $\gamma_1 \in B(1)$, $B(2) = \{\gamma_2, \ldots, \gamma_k\}$ where $k = |B(2)| + 1$, $B(3)$ contains $\gamma_{k+1}, \ldots, \gamma_l$ where $l = |B(3)| + k$, and the last $|B(r_*)|$ points in B are in $B(r_*)$. Define the $\nu_0 \times s$ matrix of probabilities $\mathbf{P} = \{p_\nu(\gamma_j)\}$, $\nu = 1, \ldots, \nu_0$, $j = 1, \ldots, s$ where $p_\nu(\gamma_j)$ is the probability given ν of reaching γ_j and its formula is as given in (2.1). Also, define $\mathbf{f}_{s\times 1} = [f(\gamma_1)\, f(\gamma_2) \cdots f(\gamma_s)]'$ and $\mathbf{g}_{\nu_0 \times 1} = [g(1)\, g(2) \cdots g(\nu_0)]'$. Then, $g(\nu)$ is unbiasedly estimable if there is a solution, $\mathbf{f}_0$ say, to the equation $\mathbf{P}\mathbf{f}_0 = \mathbf{g}$. We shall derive further results concerning unbiased estimation in the form of propositions whose proofs are based on some well-known results in linear algebra (see Edelen and Kydoniefs (1976) for details).

Proposition 3.2. *If a closed boundary B has $s < \nu_0$, then ν is not unbiasedly estimable.*

Proof. For each $\nu \leq \nu_0$, $\mathrm{E}_\nu[f(r,m)] = \nu$ defines a system of linear equations in $f(\gamma)$, $\gamma \in B$, and an unbiased estimator of ν must satisfy ν_0 such equations. This is not possible as $\nu_0 > s$. Thus, an unbiased estimator of ν does not exist when $|B| < \nu_0$.

The following example demonstrates that some functions of ν are unbiasedly estimable using Harris's (1968) ratio estimator of ν for the fixed sample size rule.

Example 1. Consider the fixed sample size rule, i.e. $B = \{(r,m) : r + m = n\}$ where $n < \nu_0$. Then, $|B| = n < \nu_0$, B is complete, and $rank(\mathbf{P}) = n$. Now, partition $\mathbf{P}$ so that

$$\mathbf{P} = \left[\frac{\mathbf{P}_1}{\mathbf{P}_2}\right],$$

where $\mathbf{P}_1 = \{p_\nu(\gamma_j)\}, \nu = 1, \ldots, n,\ j = 1, \ldots, \nu$, is an $n \times n$, lower triangular matrix of rank n and $\mathbf{P}_2 = \{p_\nu(\gamma_j)\}, \nu = n{+}1, \ldots, \nu_0,\ j = 1, \ldots, n$ is a $(\nu_0{-}n){\times}n$ matrix. Partition $\mathbf{g}$ similarly so that

$$\mathbf{Pf} = \left[\frac{\mathbf{P}_1}{\mathbf{P}_2}\right]\mathbf{f} = \left[\frac{\mathbf{g}_1}{\mathbf{g}_2}\right] = \mathbf{g},$$

where $\mathbf{g}_1 = [g(1) \cdots g(n)]'$ and $\mathbf{g}_2 = [g(n+1) \cdots g(\nu_0)]'$. Then for any $\mathbf{g}_1$, $\mathbf{f}_0 = \mathbf{P}_1^{-1}\mathbf{g}_1$, so that $\mathbf{g}_2 = \mathbf{P}_2\mathbf{f}_0 = \mathbf{P}_2\mathbf{P}_1^{-1}\mathbf{g}_1$. Such $\mathbf{g}$ have unbiased estimators and further, by the Lehmann-Scheffe theorem (Bickel and Doksum (1977), p. 122), the estimator is unique since B is complete.

For example, let $\mathbf{g}_1 = [1 \cdots n]'$, i.e. $g(\nu) = \nu$. Now, for the fixed sample size rule,

$$p_\nu(\gamma_j) = p_\nu(r, n-r) = \frac{(\nu)_r}{\nu^n} S(n,r)$$

for $r \leq \nu$ where $S(n,r)$ is a Stirling number of the second kind (Harris (1968)). Let $p^*_\nu(\gamma)$ denote the probability distribution for the boundary $B^* = \{(r, m{+}1) : (r,m) \in B\}$ so that

$$p^*_\nu(r, n-r+1) = \frac{(\nu)_r}{\nu^{n+1}} S(n+1, r)$$

for $r \leq \nu$. Note that $p_\nu(B^*) = \sum_{r \in B^*} p_\nu(r) = 1$ for $\nu \leq n$ but B^* is not closed when $n < \nu \leq \nu_0$.

Then, the solution $\mathbf{f}_0 = \mathbf{P}_1^{-1}\mathbf{g}_1$ is given by

$$f(r, n-r) = \frac{S(n+1, r)}{S(n, r)},$$

$r \leq n$, since, for $\nu \leq n$, $\mathrm{E}_\nu[f] = \nu$. So, for $\nu > n$ we have

$$\mathrm{E}_\nu[f(r, n-r)] = \nu \sum_{r=1}^{n} p_\nu^*(r, n-r+1) = \nu[1 - p_\nu^*(n+1, 0)],$$

where $p_\nu^*(n+1, 0)$ is defined to be the probability under B^* of reaching the point $(n+1, 0)$, i.e. it is the probability of not stopping sampling under B^* for given ν. Note that $p_\nu^*(n+1, 0) = 0$ if $\nu \leq n$. Hence $f(r, n-r) = S(n+1, r)/S(n, r)$ is the UMVU estimator of $g(\nu) = \nu[1 - p_\nu^*(n+1, 0)]$.

Our next proposition gives conditions for unbiased estimation of any function of ν when $r_* = \max\{r : (r, m) \in B\} < \nu_0$.

Proposition 3.3. *Let B be a closed boundary with $s \geq \nu_0$ and $r_* < \nu_0$. Then, all functions of ν have unbiased estimators if either* (i) *B is complete; or,* (ii) *$d = s - \nu_0$ where d is the dimension of the solution space of $\mathbf{Pf} = \mathbf{0}$. Further, if B is complete then the unbiased estimator is unique.*

Proof. (i) If B is complete, $\mathbf{P}$ is a square matrix of size $\nu_0 \times \nu_0$ and $rank(\mathbf{P}) = \nu_0$. Then, $\mathbf{Pf} = \mathbf{g}$ is always consistent. So, for any $\mathbf{g}$, the unique unbiased estimator is given by $\mathbf{f} = \mathbf{P}^{-1}\mathbf{g}$.

(ii) If $d = s - \nu_0$, $rank(\mathbf{P}) = \nu_0$ and consistency of $\mathbf{Pf} = \mathbf{g}$ follows.

Finally, it is of interest to consider boundaries that are not complete but admit unbiased estimators of ν and to ask whether such plans admit unique minimum variance unbiased estimators of ν or functions of ν even though they are not complete.

First note that the plans of interest are such that $|B| > \nu_0$ (by Proposition 3.1), $rank(\mathbf{P}) = \nu_0$ (by Proposition 3.3), and $|B(r)| > 1$ for some $r \leq r_* \leq \nu_0$. Let $\widehat{\nu}(\gamma)$ be an unbiased estimator of ν and let $h(\gamma)$ be unbiased for 0. Further, let H be the set of all unbiased estimators of 0 except for $h(\gamma) = 0 \ \ \forall \gamma$. Now, there exists a subset of B, A' say, where $h(\gamma) = 0$ if $\gamma \in A'$ for all $h \in H$. That A' is not empty is easily shown. Note that when $\nu = 1$, $\mathrm{E}_1[h] = 0$ implies that $h(1, m_1) = 0$. Now suppose $\nu = 2$ and $h(1, m_1) = 0$. Then $\mathrm{E}_2[h] = 0$ implies that if $|B(2)| = 1$ then $h(2, m_2) = 0$. Using induction on ν we have $h(1, m_1) = 0 \ \ \forall h \in H$ and $h(r, m_r) = 0 \ \ \forall r \leq r'$ where r' is the lowest value of r such that $|B(r)| > 1$. Hence, at a minimum A' contains the boundary points with r-coordinate values less than r'.

Proposition 3.4. *A boundary B that is not simple admits a UMVU estimator of $g(\nu)$ if $g(\nu)$ is of the form*

$$g(\nu) = \mathrm{E}_\nu[\widehat{\nu}(\gamma)|\gamma \in A']p_\nu(A') + cp_\nu(B - A'),$$

where c is an arbitrary constant that does not depend on ν and $p_\nu(X) = \sum_{r\in X} p_\nu(r)$.

Proof. Let a closed boundary B with an unbiased estimator, $\widehat{\nu}$, of ν be not simple. Since B is not simple, then, for some $x \geq 2$, $y \geq 1$, and $k \geq 0$, $(x-1, y+1)$ and $(x+k+1, y-k-1)$ are two continuation points separated by the points $D = \{\gamma_i = (x+i, y-i) : i = 0, \ldots, k, \gamma_i \in B\}$. Further, let $\alpha' = (x, y+1)$ and $\alpha'' = (x+k+1, y-k)$; note that these two points are accessible. It can be shown that, for non-simple B, $h(\gamma)$ is an unbiased estimator of 0 if it is of the form

$$\begin{aligned} h(\gamma_i) &= \frac{a(-1)^i}{(x+1)!K(\gamma_i)}, \quad \gamma_i \in D, \\ h(\gamma) &= \frac{a}{K(\gamma)}\Big[\frac{-L_{\alpha'}(\gamma)}{(x-1)!} + \frac{(-1)^{k+1}L_{\alpha''}(\gamma)}{(x+k)!}\Big], \quad \gamma \in B - D, \end{aligned} \tag{4.3}$$

where a is an arbitrary constant, $L_\alpha(\alpha) = 1$ and $L_\alpha(\gamma) = 0$ if $\alpha, \gamma \in B$, and otherwise $L_\alpha(\gamma)$ is the K-function for path segments from α to γ when α is a continuation point (Christman and Nayak (1994)). Now, a necessary and sufficient condition for $\hat{\nu}$ to be UMVU for its expectation is $\mathrm{E}_\nu[\hat{\nu}h] = 0$ for all $h \in H$ and for all $\nu \leq \nu_0$ (Lehmann (1983), Theorem 1.1, p. 77), i.e. for $\widehat{\nu}(\gamma)h(\gamma) \in H$ so that it satisfies (4.3). This condition is satisfied provided $\widehat{\nu}(\gamma) = c$, a constant that does not depend on ν, for all $\gamma \in B - A' = \{\gamma : \gamma \in B$ such that $h(\gamma) \neq 0\}$ and $\widehat{\nu}$ is otherwise a finite function of $\gamma \in B$. Hence, $\widehat{\nu}(\gamma)$ is a UMVU estimator of its expectation

$$\begin{aligned} \mathrm{E}_\nu[\widehat{\nu}(\gamma)] &= \sum_{\gamma\in A'} \widehat{\nu}(\gamma)p_\nu(\gamma) + c\sum_{\gamma\in B-A'} p_\nu(\gamma) \\ &= \mathrm{E}_\nu[\widehat{\nu}(\gamma)|\gamma \in A']p_\nu(A') + cp_\nu(B - A'). \end{aligned} \tag{4.4}$$

So $g(\nu)$ has a UMVU estimator if and only if it is of the form given in (4.4).

Example 2. Let B be a boundary for the fixed sample size stopping rule with $n = \nu_0$ where B is not simple, e.g. let $B = \{(3+i, y-i) : i = 0, 1; 3+y = \nu_0 - 1\} + \{(r, m) : r+m = \nu_0, r = 1, 2, 3, 5, \ldots, \nu_0\}$, say. Here, $\alpha' = (3, \nu_0 - 3) \in B$ and $\alpha'' = (5, \nu_0 - 5) \in B$. Further, $A' = \{(r, \nu_0 - r) : r < 3$ or $r > 5\}$. Let $\widehat{\nu}(r, m) = K^*(r, m+1)/K(r, m)$ for all $(r, m) \in A'$; otherwise set $\widehat{\nu}(r, m) = c$, a constant. Then, the expected value of $\widehat{\nu}$ is

$$\begin{aligned} \mathrm{E}_\nu[\widehat{\nu}] &= \nu\sum_{(r,m)\in A'} \frac{(\nu)_r}{\nu^{\nu_0+1}}K^*(r, m+1) + c\sum_{(r,m)\in B-A'} \frac{(\nu)_r}{\nu^{r+m}}K(r, m) \\ &= \nu p^*_\nu((A')^*) + cp_\nu(B - A'), \end{aligned}$$

where $(A')^* = \{(r, m+1) : (r, m) \in A'\}$ and $p^*_\nu((A')^*)$ is the probability under $B^* = \{(r, m+1) : (r, m) \in B\}$ of reaching any boundary point in $(A')^*$ given ν. Note that setting c as

$$c = \nu \frac{[1 - p^*_\nu((A')^*)]}{p_\nu(B - A')}$$

yields an unbiased estimator of ν but which is a function of ν so that $\hat{\nu}$ as given here is only locally minimum variance unbiased. On the other hand, setting $c = 0$ yields a UMVU estimator of $\nu p^*_\nu((A')^*)$.

Finally, we mention that unlike the unrestricted parameter space, it is possible in the restricted case to have a boundary that is simple but not minimal. This occurs if the boundary points that can be deleted without destroying closure of B have R-coordinate values of ν_0. In that case, if the boundary without these points is simple then the locally minimum variance unbiased estimator of $g(\nu)$ is also the UMVU estimator since c depends on the constant ν_0.

Acknowledgement

We thank Dr. Tapan K. Nayak for some discussions that led to this research. We also thank the referees for some helpful suggestions.

References

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

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

Christman, M. C. and Nayak, T. K. (1994). Sequential unbiased estimation of the number of classes in a population. *Statist. Sinica* **4**, 335-352.

Edelen, D. G. B. and Kydoniefs, A. D. (1976). *An Introduction to Linear Algebra for Science and Engineering.* Elsevier, New York.

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

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

Lehmann, E. L. (1983). *Theory of Point Estimation.* John Wiley and Sons, New York.

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.

Department of Mathematics and Statistics, American University, Washington, DC 20016-8050, U.S.A.

(Received January 1995; accepted October 1995)