The Annals of Statistics
1981, Vol. 9, No. 6, 1339–1342

ON ESTIMATING THE PROBABILITY OF DISCOVERING A NEW SPECIES

BY ANNE CHAO

National Tsing Hua University, Taiwan

We search a population by selecting one member at a time with replacement and observing the species of each selected member. We are interested in predicting the conditional probability of discovering a new species in the next selection after n observations. The existence and asymptotic behavior of the uniformly minimum variance unbiased estimator of the unconditional probability is investigated under the condition that the species are uniformly distributed. We also compare the performance of the estimator as a predictor of the conditional probability with that of a linear unbiased predictor.

1. Introduction. In this note, we study mainly the same problem as in Starr (1979). Consider a population which consists of an unknown number of distinct species, possibly countably many. We search this population by selecting one member at a time, noting its species identity and returning it to the population. A search is called an n-stage search if n selections are made. Imagine that the species are labeled 1, 2, $\cdots$ in any arbitrary fashion. Let p_i denote the probability that a randomly selected member belongs to the ith species, $i = 1, 2, \cdots$ and let X_i^n be the number of representatives of the species i in the n-stage search. As indicated in Starr (1979), the conditional probability that we will discover a new species in the n + 1st selection given the X_i^n is

$$U_n = \sum_i p_i I[X_i^n = 0].$$

The unconditional probability that at the last stage of an $n + 1$ stage search we will find a new species is equal to

$$\theta_n = EU_n = \sum_i p_i(1 - p_i)^n.$$

We are particularly interested in finding estimators of θ_n, which are to be used as predictors of U_n. An estimator obtained by extending the initial search an additional stage has been discussed extensively in a number of previous papers, including Starr (1979) and Robbins (1968). Based on the search of size $n + 1$, the estimator is

$$V_1 = q_1(n + 1)/(n + 1),$$

where

$$q_k(n + 1) = \sum_i I[X_i^{n+1} = k]$$

denotes the number of species which have k representatives in the $n + 1$ stage search, $k \geq 1$. Robbins (1968) has shown that V_1 is a good predictor of U_n in the sense that

$$EV_1 = EU_n = \theta_n, \text{ and } E(V_1 - U_n)^2 < (n + 1)^{-1}.$$

Starr (1979) generalized V_1 to a class of estimators which he called Robbins-type estimators. In his study, the original search was extended by an additional m stages, $m \geq 1$. Starr

Received February 1980; revised January 1981.

AMS 1970 *subject classifications.* Primary 62F10, secondary 62E20.

Key words and phrases. Discovering species, asymptotic behavior, uniformly minimum variance unbiased estimator.

showed that

$$V_m = \textstyle\sum_{k=1}^{m} \dfrac{\binom{m-1}{k-1}}{\binom{n+m}{k}} q_k(n+m)$$

is that unique linear combination of

$$q_k(n+m) = \textstyle\sum_i I[X_i^{n+m} = k], \qquad k = 1, \cdots, n+m,$$

which has expectation θ_n.

Starr (1979, page 650) conjectured that V_m is the uniformly minimum variance unbiased estimator (UMVUE) of θ_n. In Section 2 we shall disprove Starr's conjecture by obtaining the UMVUE of θ_n for a special case. The UMVUE of θ_n is further compared in Section 3 with Robbins' estimator in the associated prediction problem.

2. Results. Assume that the initial search has been extended an additional m stages, $m \geq 1$. Let

$$d = d(n+m) = \textstyle\sum_{k=1}^{n+m} q_k(n+m) \tag{1}$$

represent the number of observed species in the search of size $n+m$.

THEOREM 1. *Suppose there are μ species, $\mu \leq n+m$, and $p_1 = p_2 = \cdots = p_\mu = \mu^{-1}$. Then the UMVUE of θ_n based on a search of size $n+m$ is*

$$W_m = W_m(d) = \textstyle\sum_{k=0}^{m-1} \binom{m-1}{k} \alpha_{d-1,n+k}/\alpha_{d,n+m},$$

where $\alpha_{p,q}$ are the Stirling numbers of the second kind, $p \leq q$, defined by $x^q = \sum_{p=1}^{q} \alpha_{p,q} x^{(p)}$; if $p > q$, we define $\alpha_{p,q} = 0$.

PROOF. Under the specified condition, Harris (1968) showed that d in (1) is the complete sufficient statistic for μ. Thus, from the Lehmann-Scheffé theorem, we can conclude after some manipulations that the unique UMVUE $W_m(d)$ of θ_n is given by

$$W_m(d) = \textstyle\sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \alpha_{d,k+m}/\alpha_{d,n+m}.$$

The result follows directly from the following identity:

$$\textstyle\sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \alpha_{d,k+m} = \sum_{k=0}^{m-1} \binom{m-1}{k} \alpha_{d-1,n+k}.$$

See Chao (1980) for further details.

REMARKS.

A. We can show that, based on the original search, no unbiased estimator which is a function of the complete sufficient statistic exists. Similarly, no such estimator can be obtained from a search of size less than n.

B. If $m = 1$, the UMVUE of θ_n can be written as

$$W_1 = 1 - d\Big/\left(\frac{\alpha_{d,n+1}}{\alpha_{d,n}}\right),$$

with $d = d(n+1)$ defined by (1). It is interesting to find that W_1 is analogous in form to U_n for the uniform case, since $U_n = 1 - d(n)/\mu$, where $d(n)$ is the number of observed species in the initial search. Actually, according to a result provided in Harris (1968) for a sample of size $n + 1$, $\alpha_{d,n+1}/\alpha_{d,n}$ is asymptotically the UMVUE of μ.

We now proceed to examine the asymptotic behavior of W_m.

THEOREM 2. *If $n \to \infty$ and $\mu \to \infty$ in such a way that $n/\mu \to \alpha$, $0 < \alpha < \infty$, then with probability one,*

$$W_m = \exp(-R_m) + O(n^{-1}),$$

where R_m is the unique solution of $f(R) = R/\{1 - \exp(-R)\} = (n + m)/d$, $m = 1, 2, \cdots$.

PROOF. It follows from the recursive formula of the Stirling numbers of the second kind (Jordan, 1950, page 169) that

$$W_m = \sum_{k=0}^{m-1} \binom{m-1}{k} (\alpha_{d,n+k+1} - d\alpha_{d,n+k})/\alpha_{d,n+m}$$
$$= 1 - \frac{d\alpha_{d,n+m-1}}{\alpha_{d,n+m}} + Q,$$

where

$$Q = \sum_{k=1}^{m-1} \binom{m-1}{k-1} \frac{\alpha_{d,n+k}}{\alpha_{d,n+m}} - \sum_{k=0}^{m-2} \binom{m-1}{k} \frac{d\alpha_{d,n+k}}{\alpha_{d,n+m}}.$$

Applying the results

$$\frac{\alpha_{d,n+m-1}}{\alpha_{d,n+m}} = \frac{R_m}{n+m}\{1 + O(n^{-1})\},$$
$$\frac{d\alpha_{d,n+m-1}}{\alpha_{d,n+m}} = \{1 - \exp(-R_m)\}\{1 + O(n^{-1})\},$$

which are deduced from a theorem of Harris (1968, page 841), we can establish that $Q = O(n^{-1})$. The result is immediate.

Starr (1979) found that the Robbins predictor V_1 has an unattractive property, namely that V_1 is strongly negatively correlated with U_n. If we employ $\exp(-R_m)$ as a predictor of U_n, a similar drawback exists. The negative correlation can be explained in the following intuitive way: the more species we found in the search, the more likely we are to discover a new species in the next selection. Note that the negative correlations are asymptotic results. Therefore, we are essentially assuming that there are many species, so that the negative correlations can be reasonably understood. Numerical results indicate that W_m increases from 0 to 1 as d is increased from 1 to $n + m$ for any μ. Then the negative correlation is still valid even if there are few species.

3. Comparison.

We now compare the performance of W_1 as a predictor of U_n with that of V_1. It will be shown that W_1 is the better predictor in the sense that $E(W_1 - U_n)^2$ is asymptotically uniformly smaller than $E(V_1 - U_n)^2$. Robbins (1968) showed that

$$(n + 1)E(V_1 - U_n)^2 \to f_1(\alpha) = e^{-\alpha}(1 + \alpha) - e^{-2\alpha},$$

if $n, \mu \to \infty$ such that $n/\mu \to \alpha$, $0 < \alpha < \infty$. Under the same conditions, we can establish that

$$(n + 1)E(W_1 - U_n)^2 \to f_2(\alpha), \tag{2}$$

where

$$f_2(\alpha) = \alpha^3 e^{-3\alpha}(1 - e^{-\alpha} - \alpha e^{-\alpha})^{-1} + \alpha e^{-\alpha}(1 - e^{-\alpha} - \alpha e^{-\alpha}) + 2\alpha^2 e^{-2\alpha}.$$

The proof of (2) is omitted, although the derivation is indirect. The reader is referred to Chao (1980, pages 13-16) for details.

We show that the UMVUE is also superior in the associated prediction problem by claiming that $f_2(\alpha) < f_1(\alpha)$ for all $0 < \alpha < \infty$. It is equivalent to verify that

$$g(\alpha) = e^{-\alpha}(2 + \alpha^2 - e^{-\alpha}) < 1,$$

which follows from the fact that $g(\alpha)$ is a strictly decreasing function on $[0, \infty)$ and consequently $g(\alpha) < g(0) = 1$, for all $0 < \alpha < \infty$.

Although $E(W_1 - U_n)^2$ is uniformly smaller than $E(V_1 - U_n)^2$, we still do not know whether $E(W_1 - U_n)^2$ will attain the minimum in the class of all unbiased estimators of θ_n. We finally remark that $E(V_1 - \theta_n)^2 - E(W_1 - \theta_n)^2$ is asymptotically equal to $E(V_1 - U_n)^2 - E(W_1 - U_n)^2$. This fact reveals that the difference in variance when we employ the UMVUE, instead of V_1, to estimate θ_n is essentially the reduction of mean square error if W_1, rather than V_1, is used to predict U_n.

Acknowledgment. The author would like to thank referees for helpful comments.

REFERENCES

CHAO, A. (1980). Estimation of the probability of discovering a new species. Institute of Applied Mathematics Technical Report No. 19, National Tsing Hua University, Taiwan.

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

JORDAN, C. (1950). *Calculus of Finite Differences.* Chelsea Publishing Co., New York.

ROBBINS, H. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Statist.* **39** 256-257.

STARR, N. (1979). Linear estimation of the probability of discovering a new species. *Ann. Statist.* **7** 644-652.

INSTITUTE OF APPLIED MATHEMATICS
NATIONAL TSING HUA UNIVERSITY
HSIN-CHU, TAIWAN