The Annals of Statistics
1979, Vol. 7, No. 3, 644–652

LINEAR ESTIMATION OF THE PROBABILITY OF DISCOVERING A NEW SPECIES[1]

BY NORMAN STARR

University of Michigan

A population consisting of an unknown number of distinct species is searched by selecting one member at a time. No a priori information is available concerning the probability that an object selected from this population will represent a particular species. Based on the information available after an n-stage search it is desired to predict the conditional probability that the next selection will represent a species not represented in the n-stage sample. Properties of a class of predictors obtained by extending the search an additional m stages beyond the initial search are exhibited. These predictors have expectation equal to the unconditional probability of discovering a new species at stage $n + 1$, but may be strongly negatively correlated with the conditional probability.

1. Introduction. We consider a nonempty population π composed of (possibly countably many) distinct species, which we imagine to be labelled with the integers $1, 2, \cdots$, in some arbitrary fashion. Let p_i denote the probability that an object chosen from π is a representative of species i; we suppose that there is no a priori information available concerning either the number of species in the population or the vector of *search probabilities* $\mathbf{p} = (p_1, p_2 \cdots)$ except that $\mathbf{p} \in S$, where

$$S = \{(p_1, p_2, \cdots) : 0 \leqslant p_i \leqslant 1 \ \forall i \quad \text{and} \quad \Sigma_i p_i = 1\}.$$

We may search the population by selecting one member of π at a time, noting the species to which it belongs, and returning it to the population. (If $p_i > 0 \ \forall i$ and π is infinite, then an equivalent search may proceed nonsequentially—without replacement). If n independent selections are made then we say the search has size n (or is n-stage), let X_i^n denote the random number of representatives of species i that will be found in the search, $i = 1, 2, \cdots$, and say that species i has been discovered if X_i^n assumes a positive value.

The quantity of interest in this note is the realization of the unobservable random variable

$$U_n = \Sigma_i p_i I[X_i^n = 0],$$

the sum of the unknown probabilities associated with species which will not be discovered in a search of size n. U_n may be regarded as the *random* conditional probability that we will discover a new species at the last stage of an $n + 1$ stage search; that is, given the values $X_i^n = x_i^n$, $i = 1, 2, \cdots$ resulting from a search of

Received April 1977; revised October 1977.
[1]Research supported by U.S. Army Grant DAAG 29-76-B0302.
AMS 1970 *subject classifications.* Primary 62F10; secondary 2A99.
Key words and phrases. Linear unbiased estimation, prediction, search probabilities, species, Vandermonde determinant.

size n the realization

$$u_n = \Sigma_i p_i I[x_i^n = 0]$$

of U_n is the conditional probability that if the search were extended one more stage we would discover a new species.

In this note we discuss the problem of estimating u_n. Bear in mind that the available data comprise only sample frequencies for those species which have been discovered and that the labelling is that of the searcher. To put this in perspective, suppose that at the conclusion of a search of size n a total of d species have been discovered, and that their frequencies are $X_i^n = x_i^n$, $i = 1, \cdots, d$, where the indices are imposed by the searcher in some arbitrary manner; for example, the order in which the species were discovered. Then our problem may be formulated in the following way.

Imagine that there are a total of $d + 1$ species in the population with search probabilities $(p_1, \cdots, p_d, u_n)$ and with corresponding sample frequencies $(x_1^n, \cdots, x_d^n, 0)$. How may we use this data set to estimate u_n? Two observations are immediate. If we view the data from this perspective (that is, conditionally), then knowledge of the number of species in the population, say k, is irrelevant to estimation, unless $d = k$ in which case we know certainly that $u_n = 0$. Moreover, it is apparent that standard procedures, such as maximum likelihood (which estimates u_n to be zero for every n), are inadequate.

However, an indirect method has surfaced in the literature, apparently suggested by A. M. Turing (see [2]) and discussed in a variety of detail and perspective by Good [2, 3], Good and Toulmin [4], Harris [5], Knott [6], Robbins [7], and their bibliographies. Consider the quantity

$$\theta_n = E(U_n) = \Sigma_i p_i E I[X_i^n = 0] = \Sigma_i p_i q_i^n,$$

where we have set $q_i = 1 - p_i$, $\forall i$. θ_n denotes the unconditional probability that at the last stage of an $n + 1$ stage search we will discover a new species. To see this directly let A be the event that a species will be discovered at stage $n + 1$, and let A_i denote the event that species i will be discovered at stage $n + 1$; then the A_i are mutually exclusive with geometric probability $p_i q_i^n$ for each i, and $A = \cup_i A_i$, so that

$$P(A) = \Sigma_i P(A_i) = \theta_n.$$

Suppose now that we can develop an estimator V of θ_n for which $E(V) = \theta_n$. Then, since $E(U_n) = \theta_n$, there is some reason to hope that realizations of both V and U_n will be close to θ_n with high frequency, and hence close to one another, so that a given realization of V may represent a useful estimate of u_n. Thus, common to the papers cited above is the attempt to develop and study estimators of θ_n. Unfortunately, the problem of judging the goodness of such estimators when they are utilized to predict U_n appears to have received only modest attention, and that from a single perspective.

Of special interest to us here are estimators of the type proposed by Herbert Robbins [7]. Suppose that an initial search of size n is completed, at which time the random variable U_n assumes the unobservable value u_n. Assume, however, that with the objective of improving our chances of accurately predicting U_n, we extend the search one additional stage, and let

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

denote the number of species with exactly k representatives in the extended search of total size $n+1$. Robbins proposed as a predictor (he regards it as an "estimator") of U_n the random variable

$$V_1 = \frac{q_1(n+1)}{n+1},$$

the proportion of species with exactly one representative in the extended search. V_1 represents a good predictor of U_n in the sense that

$$(1) \qquad E(V_1) = E(U_n) = \theta_n \quad \text{and} \quad E(V_1 - U_n)^2 < \frac{1}{n+1}$$
$$\text{for every} \quad \mathbf{p} \in S. \text{ (See [7]).}$$

Indeed, we shall prove that V_1 is the unique linear combination of $q_1(n+1), q_2(n+1), \cdots, q_{n+1}(n+1)$ with expectation θ_n. However, V_1 *does not follow* U_n in a sense that might reasonably be demanded of a predictor; viz., that realizations u_n of U_n larger than θ_n be accompanied with high frequency by realizations v_1 of V_1 larger than θ_n, and vice versa. In particular, we shall prove that if the *search probabilities* p_i are equal and if the size of the search is of the same order as the number of species, then V_1 and U_n are strongly *negatively correlated.* We conclude that although V_1 will be close to U_n in the average sense of (1), that this is largely a result of the fact that the random variables have a common mean and modest variances, rather than a consequence of their being positively related or associated in any commonly understood sense of predictive inference. On the other hand, we hasten to observe that we do not yet know how to do better (and perhaps cannot).

In the next section we shall consider a class of predictors of U_n obtained by extending an initial search of size n by an additional m stages, and call it the class of Robbins-type predictors (Robbins studied the case $m = 1$). One of the referees has envisioned the following kind of conversation that could result at the conclusion of an n-stage search from the use of Robbins-type prediction. Paraphrased it goes:

Searcher: "I am considering making one more search. If I do so, am I likely to discover a new species?"

Statistician: "Make the search and then I will tell you."

The reference becomes less awkward if the problem is developed in a design context; for example:

Searcher: "I am contemplating extending my initial search an additional large number M of stages, and will so do if the expected number $M \cdot u_n$ of individuals I will select in the second search who do not represent species discovered in my initial search is large. What do you recommend?"

Statistician: "Make one more search and then I will tell you."

2. Results. In this section we shall suppose that an initial search of size n has been extended an additional m stages, $m = 1, 2, \cdots$ and let

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

denote the number of species for which there will be exactly k representatives in the search of total size $n + m$. Our immediate objective is to use the values $q_k(n + m)$, $k = 1, 2, \cdots, n + m$, to estimate the parametric function

$$\theta_n = \Sigma_i p_i q_i^n.$$

THEOREM 1. *Let $\alpha_0, \alpha_1, \cdots, \alpha_{n+m}$ be constants and define*

$$W_m = \alpha_0 + \Sigma_{k=1}^{n+m} \alpha_k q_k(n + m).$$

Then $E(W_m) = \theta_n$ identically in $\mathbf{p} \in S$ if and only if $\alpha_0 = \alpha_{m+1} = \cdots = \alpha_{m+n} = 0$ and

$$\alpha_k = \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}} \qquad \text{for} \quad k = 1, \cdots, m.$$

That is, the estimator defined by

$$V_m = \Sigma_{k=1}^{m} \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}} q_k(n + m)$$

is the unique linear form in $\{q_k(n + m), k = 1, \cdots, n + m\}$ with expectation θ_n.

PROOF. Observe that the random variables $\{q_k(n + m), k = 1, \cdots, n + m\}$ are constrained by the condition

$$\Sigma_{k=1}^{n+m} k q_k(n + m) = n + m. \tag{2}$$

By direct computation

$$\begin{aligned} EW_m &= \alpha_0 + \Sigma_{k=1}^{n+m} \alpha_k \Sigma_i E\{I[X_i^{n+m} = k]\} \\ &= \alpha_0 + \Sigma_{k=1}^{n+m} \alpha_k \binom{n+m}{k} \Sigma_i p_i^k q_i^{n+m-k} \\ &= \alpha_0 + \Sigma_{k=1}^{n+m} \alpha_k \binom{n+m}{k} \Sigma_i p_i \Sigma_{j=0}^{k-1} (-1)^j \binom{k-1}{j} q_i^{n+m-k+j}. \end{aligned}$$

Interchanging the order of summation and making use of symmetry in the arguments of the binomial coefficients yield

(3)

$$EW_m = \alpha_0 + \Sigma_i p_i \Sigma_{j=1}^{n+m} \Sigma_{k=0}^{n+m-j} (-1)^k \alpha_{k+j} \binom{n+m}{k+j} \binom{k+j-1}{k} q_i^{n+m-j}.$$

To see that $EV_m = \theta_n$, set $\alpha_0 = \alpha_{m+1} = \cdots = \alpha_{m+n} = 0$ and

$$\alpha_k = \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}}, \qquad k = 1, \cdots, m$$

in (3). Then

$$\begin{aligned} EV_m &= \Sigma_i p_i \Sigma_{j=1}^{m} \Sigma_{k=0}^{m-j} (-1)^k \binom{m-1}{k+j-1} \binom{k+j-1}{k} q_i^{n+m-j} \\ &= \Sigma_i p_i q_i^n + \Sigma_i p_i \Sigma_{j=1}^{m-1} \Sigma_{k=0}^{m-j} (-1)^k \binom{m-1}{k+j-1} \binom{k+j-1}{k} q_i^{n+m-j} \\ &= \Sigma_i p_i q_i^n + \Sigma_i p_i \Sigma_{j=1}^{m-1} \binom{m-1}{j-1} q_i^{n+m-j} \Sigma_{k=0}^{m-j} (-1)^k \binom{m-j}{k} = \theta_n \end{aligned}$$

where the last equality follows from the Binomial theorem.

To prove uniqueness, set

$$a_j = \Sigma_{k=0}^{n+m-j} (-1)^k \alpha_{k+j} \binom{n+m}{k+j} \binom{k+j-1}{k}$$

for each $j = 1, \cdots, n+m-1$. Then from (3) we have for every $\mathbf{p} \in S$

(4) $$E(W_m) = \alpha_0 + \alpha_{n+m} + \Sigma_{j=1}^{n+m-1} a_j \Sigma_i p_i q_i^{n+m-j}.$$

Thus, setting $b_j = a_j$ for $j \neq m$ and $b_m = (a_m - 1)$, it is easily seen that $E(W_m) = \theta_n$ identically in $\mathbf{p} \in S$ only if

(5) $$\alpha_0 + \alpha_{n+m} + \Sigma_{j=1}^{n+m-1} b_j \Sigma_i p_i q_i^{n+m-j} = 0$$

identically in $\mathbf{p} \in S$. Clearly, (5) can hold only if $\alpha_{n+m} = -\alpha_0$, where α_0 is arbitrary. Let $\mathbf{b}$ denote the column vector whose jth component is b_j, and for given $\mathbf{p} \in S$ let $\mathbf{d}(\mathbf{p})$ denote the column vector with jth component $\Sigma_i p_i q_i^{n+m-j}$, $j = 1, \cdots, n+m-1$. Thus from (5) we have that $E(W_m) = \theta_n$ identically in $\mathbf{p} \in S$ only if

(6) $$\mathbf{b}'\mathbf{d}(\mathbf{p}) = 0 \qquad \text{for every} \quad \mathbf{p} \in S.$$

In the sequel we shall show that

(7) $$\text{span}\{\mathbf{d}(\mathbf{p}), \mathbf{p} \in S\} \qquad \text{has dimension} \quad n+m-1$$

so that (6) holds only if $\mathbf{b}$ is the null vector; that is, only if

$$a_j = 0, j = 1, \cdots, n+m, j \neq m \qquad \text{and} \quad a_m = 1.$$

To summarize, $EW_m = \theta_n$ identically in $\mathbf{p} \in S$ only if

$$\Sigma_{k=0}^{n+m-j}(-1)^k\alpha_{k+j}\binom{n+m}{k+j}\binom{k+j-1}{k} = 0, j = 1, \cdots, n+m-1, j \neq m,$$

$$\Sigma_{k=0}^{n}(-1)^k\alpha_{k+m}\binom{n+m}{k+m}\binom{k+m-1}{k} = 1, \alpha_{n+m} = -\alpha_0$$

and α_0 is arbitrary. Thus, if $EW_m = \theta_n$, any choice of α_0 uniquely determines the other coefficients $\alpha_1, \alpha_2, \cdots, \alpha_{n+m}$. But for an arbitrary choice of a constant α it follows from (2) that

$$\begin{aligned} v_m &= (1-\alpha)v_m + \alpha v_m \\ &= (1-\alpha)\Sigma_{k=1}^{m}\frac{\binom{m-1}{k-1}}{\binom{n+m}{k}} + \alpha\Sigma_{k=1}^{m}\frac{\binom{m-1}{k-1}}{\binom{n+m}{k}}\left\{\frac{1}{k}\left[(n+m)\right.\right. \\ &\quad \left.\left. -\Sigma_{j=1;j\neq k}^{n+m} jq_j(n+m)\right]\right\} \\ &= \alpha_0 + \Sigma_{k=1}^{n+m}\alpha_k q_k(n+m), \end{aligned}$$

where $\alpha_0 = \alpha(n+m)\Sigma_{k=1}^{m}\binom{m-1}{k-1}/k\binom{n+m}{k}$ is arbitrary (since α is) and $\alpha_1, \cdots, \alpha_{n+m}$ are uniquely determined by the choice of α_0. Thus, $W_m = V_m$ if $EW_m = \theta_n$.

are uniquely determined by the choice of α_0. Thus, $W_m = V_m$ if $EW_m = \theta_n$.

It remains to verify (7). For each $j = 1, \cdots, n+m-1$, consider the column vector $\mathbf{c}_j$ whose kth component, $k = 1, \cdots, n+m-1$, is the kth component of $\mathbf{d}(\mathbf{p})$, where $\mathbf{p}$ is the $k+1$ component vector $\mathbf{p} = (1/(k+1), 1/(k+1), \cdots, 1/(k+1))$, so that $\mathbf{p} \in S$. Form the $(n+m-1) \times (n+m-1)$ matrix C with jth column $\mathbf{c}_j$, so that

$$C = \begin{bmatrix} \left(\frac{1}{2}\right)^{n+m-1} & \left(\frac{1}{2}\right)^{n+m-2} & \cdots\cdots & \left(\frac{1}{2}\right) \\ \left(\frac{2}{3}\right)^{n+m-1} & \left(\frac{2}{3}\right)^{n+m-2} & \cdots\cdots & \left(\frac{2}{3}\right) \\ \vdots & \vdots & & \vdots \\ \left(1-\frac{1}{n+m}\right)^{n+m-1} & \left(1-\frac{1}{n+m}\right)^{n+m-2} & \cdots & \left(1-\frac{1}{n+m}\right) \end{bmatrix}.$$

Then the determinant of C is easily seen to be proportional to the Vandermonde determinant which is nonzero, establishing (7), and completing the proof.

REMARKS.

1. It follows from (5) that for $m \leqslant 0$, there is no choice of $\alpha_0, \cdots, \alpha_{n+m}$ for which $E(W_m) = \theta_n$ identically in $\mathbf{p} \in S$; that is, no linear form in $\{q_k(n+m), k = 1, \cdots, n+m\}$ obtained from a search of size less than $n+1$ has expectation θ_n for every $\mathbf{p} \in S$. This contradicts the assertion (2.09) of Knott

[6] that

$$\Sigma_{i=1}^{n} \frac{(-1)^i q_i(n)}{\binom{n}{i}}$$

is an unbiased estimate of θ_n. In particular, the estimator

$$V_0 = \frac{q_1(n)}{n} \tag{8}$$

of Good [2] has bias $E(V_0 - \theta_n) = \Sigma_i \left(\frac{p_i}{q_i}\right) p_i q_i^n$.

To verify the remark, take $\mathbf{p}$ to be the k component vector $(1/k, \cdots, 1/k)$; then from (5) for $m \leqslant 0$ $E(W_m) = \theta_n$ identically in $\mathbf{p} \in S$ only if *in particular*

$$\alpha_0 + \alpha_{n+m} + \Sigma_{j=1}^{n+m-1} a_j \left(1 - \frac{1}{k}\right)^{n+m-j} = \left(1 - \frac{1}{k}\right)^n \tag{9}$$

for every $k = 2, 3, \cdots$ Clearly there is no set of $n + m + 1$ coefficients for which (9) holds identically in k, proving the remark.

2. We suspect (but have not yet proved) that V_m is the uniformly minimum variance unbiased estimator of θ_n based on a search of size $n + m$.

Next, we shall turn our attention specifically to V_1, the predictor of U_n based on a search of total size $n + 1$. The difficulty with V_1 referred to in the introduction is exhibited as

THEOREM 2. *Suppose that there are k species, that* $p_1 = \cdots = p_k$, *and that n and k become large in such a way that*

$$\frac{n}{k} \to \alpha, \qquad 0 < \alpha < \infty. \tag{10}$$

Then, under the limiting operation defined by (10)

$$\rho(V_1, U_n) \to f_\rho(\alpha) = -\frac{\alpha^2}{\left[(\alpha e^\alpha - \alpha^2 - \alpha)(e^\alpha - \alpha^2 + \alpha - 1)\right]^{1/2}} \tag{11}$$

where ρ *denotes correlation.*

PROOF. For any $\mathbf{p} \in S$

$$EU_n = EV_1 = \Sigma p_i q_i^n$$

$$EU_n^2 = \Sigma p_i^2 q_i^n + \Sigma\Sigma_{i \neq j} p_i q_j (1 - p_i - p_j)^n$$

$$EV_1^2 = \frac{1}{n+1}\left[\Sigma_i p_i q_i^n + n\Sigma\Sigma_{i \neq j} p_i p_j (1 - p_i - p_j)^{n-1}\right]$$

$$EU_n V_1 = \frac{1}{n+1}\left[\Sigma_i p_i q_i^n + \Sigma\Sigma_{i \neq j} p_i p_j (1 - p_i - p_j)^n + n\Sigma\Sigma_{i \neq j} p_i p_j q_i (1 - p_i - p_j)^{n-1}\right].$$

Setting $p_i = 1/k$, $i = 1, \cdots, k$ and (carefully) taking the limit defined by (10) as n and k tend to infinity yield

$$(n+1)\mathrm{Cov}(V_1, U_n) \to f_c(\alpha) = -\alpha^2 e^{-2\alpha}$$

$$(n+1)\mathrm{Var}(U_n) \to f_u(\alpha) = \alpha e^{-\alpha} - \alpha^2 e^{-2\alpha} - \alpha e^{-2\alpha}$$

$$(n+1)\mathrm{Var}(V_1) \to f_v(\alpha) = e^{-\alpha} - \alpha^2 e^{-2\alpha} - e^{-2\alpha} + \alpha e^{-2\alpha}.$$

The result is immediate.

REMARKS.

1. The limit $f_\rho(\alpha)$ of the correlation functions is increasing in α, tends to -1 as $\alpha \to 0$ and to 0 as $\alpha \to \infty$. The values of f_ρ are given in Table 1 for various α.

TABLE 1
Values of $f_\rho(\alpha)$, defined by (11), *as a function of α.*

α	0.1	0.2	0.3	0.4	0.5
$f_\rho(\alpha)$	−0.9954	−0.9900	−0.9835	−0.9759	−0.9671
α	0.6	0.7	0.8	0.9	1.0
$f_\rho(\alpha)$	−0.9569	−0.9452	−0.9319	−0.9169	−0.9001
α	1.5	2.0	3.0	5.0	10.0
$f_\rho(\alpha)$	−0.7896	−0.6444	−0.3582	−0.0830	−0.0014

2. The limit f_u of the variance of U_n is maximized at the α value which solves $e^\alpha(1-\alpha) = 1 - 2\alpha^2$, and the limit f_v of the variance of V_1 at the α value which solves $3 - 4\alpha + 2\alpha^2 = e^\alpha$. Thus, the maximum values of f_u and f_v, achieved at about $\alpha = 1.97$ and 0.46 respectively, are approximately 0.16 and 0.33; f_u and f_v tend to zero as $\alpha \to$ zero or infinity.
3. From (12) it follows that

$$(n+1)E(v_1 - u_0)^2 \to e^{-\alpha}(1+\alpha) - e^{-2\alpha},$$

agreeing with [7]. The limiting quantity has a maximum value of about 0.61.
4. We do not know whether Robbins-type predictors of U_n may be positively correlated with U_n for some choice of $m > 2$. However, for $m = 2$, $\lim \rho(V_2, U_n) = \lim \rho(V_1, U_n)$ and $\lim \mathrm{Var}(V_2) = \lim \mathrm{Var}(V_1)$.
5. Concerning the predictor V_0 of U_n defined by (8), we have also that $\lim \rho(V_0, U_n) = \lim \rho(V_1 U_n)$ and $\lim \mathrm{Var}(V_0) = \lim \mathrm{Var}(V_1)$. (See [1] for details.)
6. After completing the paper, we discovered that our expression (12) could have been deduced from expression (14) of reference [8]. Our method is more direct.

Acknowledgment. I would like to thank the referees for a number of suggestions which were incorporated in the final version. I am also indebted to my colleagues Michael Woodroofe and Bruce Hill for very stimulating and useful discussions.

REFERENCES

[1] GOOD, I. J. (1953). On the population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264.

[2] GOOD, I. J. (1965). The estimation of probabilities. Research Monograph No. 30, M.I.T. Press, Cambridge.

[3] GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase of population coverage, when a sample is increased. *Biometrika* **43** 45–63.

[4] HARRIS, B. (1959). Determining bounds on integrals with applications to cataloguing problems. *Ann. Math. Statist.* **30** 521–548.

[5] KNOTT, MARTIN. (1967). Models for cataloguing problems. *Ann. Math. Statist.* **38** 1255–1260.

[6] RASMUSSEN, S. and STARR, N. (1977). Optimal and adaptive stopping in the search for new species. Technical Report No. 73, Depart. Statist., Univ. Michigan. (Revised April 1977).

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

[8] SEVAST'YANOV, B. A. and CHISTYAKOV, N. P. (1964). Asymptotic normality in the classical ball problem. *Theor. Probability Appl.* **9** 198–211.

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
1447 MASON HALL
419 S. STATE ST.
ANN ARBOR, MICHIGAN 48109