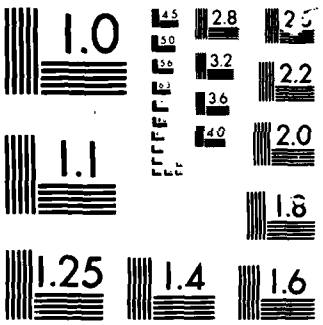
AD-A167 493 NONPARAMETRIC ESTIMATION OF THE PROBABILITY OF DISCOVERING A NEW SPECIES(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER M K CLAYTON ET AL. JAN 86 1/1

UNCLASSIFIED MRC-TSR-2898 DAAG29-80-C-0041 F/G 12/1 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY
CHART

MRC Technical Summary Report #2898

NONPARAMETRIC ESTIMATION OF THE PROBABILITY OF DISCOVERING A NEW SPECIES

Murray K. Clayton and Edward W. Frees

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

January 1986

(Received December 12, 1985)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

NONPARAMETRIC ESTIMATION OF THE PROBABILITY OF DISCOVERING A NEW SPECIES

Murray K. Clayton* and Edward W. Frees*

Technical Summary Report #2898
January 1986

ABSTRACT

A random sample is taken from a population consisting of an unknown number of distinct species. A quantity of interest is the probability of discovering a new species when an additional draw from the population is made. An estimator of this quantity was introduced by Starr (1979). We prove a conjecture of Starr's that the estimator is uniformly minimum variance unbiased and give various asymptotic properties of the estimator. A nonparametric maximum likelihood estimator is introduced which has similar asymptotic properties. A Monte-Carlo study is given which suggests guidelines for choosing an estimator under various circumstances.

AMS (MOS) Subject Classifications: Primary 92A10; Secondary 62G05, 62P10

Keywords and Phrases: U-statistics, nonparametric maximum likelihood estimator

Work Unit Number 4 (Statistics and Probability)

---

*Department of Statistics, 1210 West Dayton Street, University of Wisconsin-Madison, Madison, WI 53706.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

NONPARAMETRIC ESTIMATION OF THE PROBABILITY OF DISCOVERING A NEW SPECIES

Murray K. Clayton* and Edward W. Frees*

§1. Introduction

In many ecological studies, a population is sampled to determine the number of species that exist in the population. This quantity provides a partial description of the population and may be used in the comparison of populations over time or space. Such sampling often takes place sequentially and it is in this context that a related quantity arises: the probability of discovering a new species in a future sample based on sampling which has already taken place. By itself, this probability indirectly leads to information about the number of species in the population; it might also be used in a sequential sampling scheme where the goal is to decide when to stop sampling.

To describe the problem formally, consider a population composed of distinct species and use M_i to represent the i^{th} species, i=1,2,.... We assume that the species have no natural order and that the number of species may be countably infinite. Suppose n independent drawings are made from the population, with replacement if the population is finite, and define $X_j = i$ when the j^{th} draw is from M_i. Let $X_i^n = \sum_{j=1}^{n} I(X_j = i)$ be the number of representatives of the species M_i in n drawings, where I(A) as the indicator function of the set A. The conditional probability of discovering a new species in one additional search is

*Department of Statistics, 1210 West Dayton Street, University of Wisconsin-Madison, Madison, WI 53706.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

$$U_n = \sum_i p_i \, I(X_i^n = 0), \tag{1.1}$$

where $p_i = P(X_1 = i)$. The corresponding unconditional probability of new species discovery is

$$\theta_n = E(U_n) = \sum_i p_i q_i^n, \tag{1.2}$$

where $q_i = 1 - p_i$.

As argued, for example, by Starr (1979), standard statistical procedures for direct estimation of a realization of the random variable U_n are inadequate. An alternative, and closely related, goal is the estimation of the parameter θ_n. Estimation of θ_n has attracted interest in the recent literature; for example, Starr (1979), Chao (1981, correction, 1982), and Banerjee and Sinha (1985) have recently introduced estimators of θ_n. For earlier efforts on this and related problems, see Good (1953, 1965), Good and Toulmin (1956), Goodman (1949), Harris (1959, 1968), Knott (1967) and Robbins (1968). We note that our model is not confined to sampling species from populations; related problems are discussed in Efron and Thisted (1976) and others. The sequential problem mentioned above is discussed in Goodman (1953), Rasmussen and Starr (1979), and Banerjee and Sinha (1985). A Bayesian approach can be found in Hill (1979).

Without additional constraints on the model, it is well known that there is no unbiased estimator of θ_n based on a sample size less than n+1 (cf., Appendix A, Lemma A.1). However, if one additional search is made, Robbins (1968) noted that

$$V_1 = (n + 1)^{-1} \sum_i I(X_i^{n+1} = 1) \tag{1.3}$$

is an unbiased estimator of θ_n. Robbins also argued that V_1 follows U_n in the sense that the expected squared difference is strictly bounded from above by

$(n + 1)^{-1}$. Starr (1979) gave a more general version of the Robbins estimator. Starr supposed that the initial search of size n was extended by m additional stages and defined

$$V_m = \sum_{k=1}^{m} \binom{m-1}{k-1}\binom{n+m}{k}^{-1} \sum_i I(X_i^{n+m} = k). \tag{1.4}$$

The term $\sum_i I(X_i^{n+m} = k)$ is the number of species with k representatives and is a part of the so-called "sampling frequency of frequencies" (cf., Good, 1953). It is important in applications because only the summary statistics $\{\sum_i I(X_i^{n+m} = k)\}_{k=1}^{n+m}$ need to be retained for analysis. Starr showed that V_m is the unique unbiased estimator which is a linear combination of $\{\sum_i I(X_i^{n+m} = k)\}_{k=1}^{n+m}$ and conjectured that it is the minimum variance unbiased estimator (MVUE). This property was discussed by Chao (1981) who proposed an alternative estimator which was further modified by Banerjee and Sinha (1985). Chao's estimator was motivated by Harris's (1968) work in the important special case of equal cell probabilities.

In §2 we answer the issues raised by Chao (1981, 1982) and Banerjee and Sinha (1985) by proving Starr's conjecture that V_m is the MVUE. The technique is to use some results of Halmos (1946) on unbiased estimation and show that V_m is a U-statistic. Several other properties of V_m are also immediately available based on the theory of U-statistics and are described in §2. In §3 we introduce a nonparametric maximum likelihood estimator (NPMLE) as an alternative to V_m. Although the NPMLE is biased in finite samples, we show that it has similar large-sample properties. Some heuristic arguments in addition to the simulation results of §4 suggest that the NPMLE may be a desirable alternative to Starr's estimator in certain situations. We close in §5 with some general remarks. Appendices A and B provide details of the proofs of the technical results of §2 and §3, respectively.

§2. Properties of Starr's Estimator

For convenience, we begin by stating some results of Halmos (1946). A direct consequence of these results is the verification of Starr's conjecture. Another consequence is that V_m, defined in (1.4), is a U-statistic. This property has further consequences which we exploit.

To state Halmos's results, define Π^* to be the class of all probability distributions on R, the real line. Let E be a Borel subset of R. Define $\Pi(E)$ to be the class of all $P \in \Pi^*$ that assign probability to some finite subset of E and let Π be some subset of Π^* that contains $\Pi(E)$. For each $P \in \Pi$, let $X_1,\ldots,X_N$ be an i.i.d. random sample. Let $\{i_1,\ldots,i_k\}$ be a subset of size k of $\{1,2,\ldots,N\}$ and let $\sum_c$ be the sum over all $\binom{N}{k}$ distinct combinations of $\{i_1,\ldots,i_k\}$. A linear functional F(P) is said to be homogeneous of degree k if there exists a mapping h from R^k to R such that

$$F(P) = E_P\, h(X_1,\ldots,X_k) = \int \ldots \int h(x_1,\ldots,x_k)\, dP(x_1) \ldots dP(x_k)$$

for all $P \in \Pi$ and if the integer k is minimal.

Lemma 2.1 (Halmos, 1946, Theorems 3 and 5)

Let F(P) be homogeneous of degree k over Π with $F(P) = E_P\, h(X_1,\ldots,X_k)$.

(a) If $f(X_1,\ldots,X_N)$ is a symmetric, unbiased estimate of F(P), then for every point $(x_1,\ldots,x_N)$ with $x_i \in E$, $f(x_1,\ldots,x_N) = \binom{N}{k}^{-1} \sum_c h(X_{i_1},\ldots,X_{i_k})$.

(b) Among all unbiased estimators of F(P), $\binom{N}{k}^{-1} \sum_c h(X_{i_1},\ldots,X_{i_k})$ has minimum variance.

To prove Starr's conjecture, define $E = \{1,2,\ldots,\}$, $N = n+m$ and let Π be the set of all probability distributions defined on E. We shall find the form of $h(\cdot)$ which is appropriate for this application. To motivate the discussion, we note that the indicator of the i^{th} species having one representative

can be expressed by

$$I(X_i^{n+1} = 1) = \sum_{j=1}^{n+1} I(X_j = i) \prod_{\substack{k=1 \\ k\neq j}}^{n+1} I(X_k \neq i). \tag{2.1}$$

We use the kernel function of size n+1 defined by

$$h(X_1,\ldots,X_{n+1}) = (n+1)^{-1} \sum_i \sum_{j=1}^{n+1} I(X_j = i) \prod_{\substack{k=1 \\ k\neq j}}^{n+1} I(X_k \neq i), \tag{2.2}$$

that is, the proportion of species with one representative. It is easy to see that $h(\cdot)$ is symmetric and unbiased for θ_n. The proof that θ_n is homogeneous of degree $k = n+1$ over Π is standard and is given in Appendix A (Lemma A.1). Thus, by Lemma 2.1 we immediately have the following properties.

Property 2.1

The statistic V_m is a U-statistic with kernel $h(\cdot)$ and degree n+1, i.e.,

$$V_m = \binom{n+m}{n+1}^{-1} \sum_c h(X_{i_1},\ldots,X_{i_{n+1}}). \tag{2.3}$$

Property 2.2

Based on a random sample of size n+m, V_m is the MVUE for θ_n over Π.

A consequence of Property 2.2 is that V_m has desirable properties as an estimator of θ_n for any fixed number m additional searches. If the number of additional searches is large, from Property 2.1 and the theory of U-statistics it immediately follows that $V_m \to \theta_n$ with probability one, as $m \to \infty$. Thus, the estimator converges to the parameter of interest. The rate of convergence can further be described by

Property 2.3.

Define $\sigma^2 = (n+1)^2\left(\sum_i p_i q_i^{2n-2}(np_i-q_i)^2 - \left(\sum_i p_i q_i^{n-1}(np_i-q_i)\right)^2\right)$. Then,

$$V_m = \theta_n + (n+m)^{-1/2}\,\sigma\, Z + o_p((n+m)^{-1/2}) \qquad (2.4)$$

as $m \to \infty$, where Z is a standard normal random variable.

Remark: The proof of Property 2.3 is standard in the theory of U-statistics (cf., Serfling, 1980, page 192). One only needs to check the calculation of the asymptotic variance which is provided in Appendix A (Lemma A.2). Perhaps the most interesting aspect of Property 2.3 is the fact that in the case of equal species probabilities, it can easily be shown that $\sigma = 0$. Indeed, by another application of U-statistic theory in Appendix A, we have

Property 2.4.

Suppose $p_1 = p_2 = \ldots = p_{1/\mu} = \mu$ for some $\mu > 0$. Then,

$$V_m = \theta_n + (n+m)^{-1}\binom{n+1}{2}\mu(1-\mu)^{n-2}(\mu-2(n+1)^{-1})(\chi^2-1) + o_p((n+m)^{-1})$$

as $m \to \infty$, where χ^2 is a chi-square random variable with 1 degree of freedom.

Thus, the rate at which V_m approaches θ_n in the important special case of equal probabilities is of a different order of magnitude (with respect to weak convergence to a nondegenerate distribution) than the general case. This characteristic is important since a comparison of various alternative estimators in this special case can be misleading when drawing conclusions about their relative performance in the more general set-up of unequal probabilities. In other situations, Starr (1979), Chao (1981), and Banerjee and Sinha (1985) use the equiprobable case as examples of their results. It should also be noted that the equiprobable cells model is unlikely to arise in nature when

sampling for species although it arises naturally in the cataloging problem of, for example, Harris (1959).

§3. An Alternative Nonparametric Estimator

Starr's estimator V_m is attractive computationally, since it is the linear combination of the "frequency of frequencies," and it has desirable theoretical properties since it can be described as a U-statistic. However, because it is derived from summary statistics, there may be some loss of information in a finite number of additional searches, in some sense. For example, if we set m = 1, then from (1.3) we see that V_1 is the sample proportion of species with one representative. Note that this estimator treats species with 0,2,3,...,n+1 representatives equally. Motivated by these heuristic arguments, we introduce the following nonparametric estimator of θ_n based on an initial sample size n and additional search m. Define $\hat{p}_i = (n+m)^{-1} \sum_{j=1}^{n+m} I(X_j = i)$ and $\hat{q}_i = 1 - \hat{p}_i$, $i = 1,2,\ldots$. *The NPMLE of* θ_n *is* defined to be:

$$\hat{\theta}_m = \sum_i \hat{p}_i \hat{q}_i^{\,n}. \tag{3.1}$$

Unlike V_m, $\hat{\theta}_m$ is a biased estimator of θ_n. Since $(n+m)\hat{q}_i$ is a binomial random variable, it is straightforward to explicitly write out the bias as a linear combination of powers of q_i and Stirling numbers of the second kind. Finite sample properties of $\hat{\theta}_m$ are further discussed in §4. Asymptotically (as $m \to \infty$), $\hat{\theta}_m$ behaves similarly to V_m. By the strong law of large numbers, with probability one, $\hat{q}_i \to q_i$, and it is not hard to show that $\hat{\theta}_m \to \theta_n$ with probability, one as $m \to \infty$. We also have the following two asymptotic properties.

Property 3.1

Let σ^2 be as defined in Property 2.3. Then

$$\hat{\theta}_m = \theta_n + (n+m)^{-1/2}\,\sigma\, Z + o_p\left((n+m)^{-1/2}\right)$$

as $m \to \infty$.

Property 3.2

Suppose $p_1 = p_2 = \ldots = p_{1/\mu} = \mu$ for some $\mu > 0$. Then,

$$\hat{\theta}_m = \theta_n + (n+m)^{-1}\binom{n+1}{2}(1-\mu)^{n-2}(\mu-2(n+1)^{-1})\left(\mu(\chi^2-1) + (1-\mu)\right)$$

$$+ o_p\left((n+m)^{-1}\right)$$

as $m \to \infty$.

The proof of Properties 3.1 and 3.2 are in Appendix B. Comparing Properties 2.3 and 3.1, we see that V_m and $\hat{\theta}_m$ are asymptotically equivalent to the first order (i.e., $(n+m)^{-1/2}$). An advantage of the NPMLE $\hat{\theta}_m$ is that, since strongly consistent estimators of q_i and hence σ can be constructed, we have as an immediate corollary of Property 3.1 large sample interval estimates of θ_n. Comparing Properties 2.4 and 3.2, we see that V_m and $\hat{\theta}_m$ are of the same order of magnitude and have same variance in their respective asymptotic distributions. The estimator V_m is slightly superior to $\hat{\theta}_m$ in the sense that the asymptotic distribution of $V_m-\theta_n$ has mean zero unlike that of $\hat{\theta}_m-\theta_n$. We remark that in this special case of equiprobable cells, Chao's (1981) extension of Harris's (1968) estimator is MVUE for fixed m and hence is a strong competitor to V_m and $\hat{\theta}_m$.

As noted, the rate of convergence of V_m and $\hat{\theta}_m$ is markedly different in the equiprobable case in comparison to the general case. Moreover, in some

sense the equiprobable case is the only one in which this can happen. Specifically, we have the following result.

Property 3.3

Consider σ^2 defined in Property 2.3 and suppose that the number of species exceeds n. Then, $\sigma^2 = 0$ if and only if $p_1 = p_2 = \ldots = p_{1/\mu}$ for some $\mu > 0$.

§4. Small Sample Properties

In this section we investigate the behavior of Starr's estimator, V_m, and the NPMLE, $\hat{\theta}_m$, when m is small via a Monte-Carlo simulation. We look at their bias and mean square error as estimates of θ_n and make some comments regarding modifications of $\hat{\theta}_m$ which have desirable properties. Finally, we investigate modifications of V_m and $\hat{\theta}_m$ suitable for use when $m = 0$. All computations were done on a VAX 11/750 owned and operated by the Department of Statistics at the University of Wisconsin-Madison. The simulations were performed using the National Bureau of Standard's Core Math Library (CMLIB) pseudo-uniform random number generator UNI.

Two classes of distributions were used to construct the probability distribution $\{p_i;\ i \geq 1\}$. These were: (1) equiprobable, with $p_i = \mu$, $1 \leq i \leq 1/\mu$; and (2) truncated geometric, with $p_i = qp^{i-1}/(1-p^c)$, $1 \leq i \leq c$, $0 < p < 1$, $q = 1 - p$. For the equiprobable cells model, values of $\mu = .1, .02, .01$ were used; for the truncated geometric model, values of $p = .1, .5, .9$ and $c = 10, 100$ were used. For each assignment of $\{p_i\}$, θ_n was determined and 1,000 simulations were performed. For each simulation, this involved drawing a sample of size n and a subsequent sample of size m. The pairs $(n,m) = (10,1), (10,10), (50,1), (50,10), (50,50)$ were included. For each sample, $\hat{\theta}_m$ and V_m were computed. Tables 4.1-4.2 show the mean values of $\hat{\theta}_m$ and V_m over the

1,000 samples denoted in the tables as $E\hat{\theta}_m$ and EV_m, respectively. (The rows corresponding to m = 0 will be discussed below.) In addition, the estimated root mean square error of the estimates, denoted by RMSE ($\hat{\theta}_m$) and RMSE (V_m), respectively, are given in Tables 4.1-4.2. Of course, since V_m is unbiased, RMSE (V_m) is also an estimate of the standard error of V_m.

Generally, in the equiprobable case, V_m has lower root mean square error than $\hat{\theta}_m$. Comparing Properties 2.4 and 3.2, we have up to order $(n+m)^{-1}$, $E(V_m-\theta_n)^2 = E(\hat{\theta}_m-\theta_n)^2 \quad 2\mu^2/(3\mu^2-2\mu+1)$. Thus, for μ small, RMSE (V_m) will be approximately $2\mu^2$ times RMSE ($\hat{\theta}_m$). While the differences in RMSE for V_m and $\hat{\theta}_m$ in Table 4.1 are not all of this magnitude, we do see that V_m is a better estimator of $\hat{\theta}_m$ in terms of RMSE.

The situation is reversed to a large extent when the truncated geometric is used for p_i. These results appear in Table 4.2. It is evident in this case, as in Table 4.1, that $\hat{\theta}_m$ tends to underestimate θ_n and that the bias can be considerable. From the results of sections 2 and 3, we expect $\hat{\theta}_m$ and V_m to have the same asymptotic mean square error. From Table 4.2, it appears that, when p is not too large, the mean square error of $\hat{\theta}_m$ is less than V_m, sometimes considerably so. That this can fail when p is large is not surprising since the truncated geometric distribution tends to the equiprobable case when p tends to one. Specifically, $qp^i/(1-p^c) \to 1/c$ for each i as $p \to 1$.

That $\hat{\theta}_m$ dominate V_m in terms of the truncated distribution when p is small can be seen in an example: Let c = 2, m = 1, and n = 2, so $p_1 = q(1-p^2)$ and $p_2 = qp/(1-p^2)$. Then $\theta_2 = p_1p_2$ and it is easy to show that $E\hat{\theta}_1 = 2/3\ p_1p_2$, which represents a considerable bias. For this example it can be shown that $E(\hat{\theta}_1-\theta_2)^2 = (4p_1p_2-9p_1^2p_2^2)/27$ and $E(V_1-\theta_2)^2 = p_1p_2(p_1-p_2)^2$. It follows that $E(V_1-\theta_2)^2 > E(\hat{\theta}_1-\theta_2)^2$ if $p_2 < \frac{1}{2} - \sqrt{77}/66$, or equivalently, if $p < .5799$.

Table 4.1
Equiprobable Case

μ	n	m	θ_n	EV_n	$RMSEV_m$	$E\hat{\theta}_m$	$RMSE\hat{\theta}_m$	$E\theta^*_m$	$RMSE\hat{\theta}_m$
.1	10	0	.3487	.3923	.1660	.1836	.1717	.3671	.0961
	10	1	.3487	.3488	.1364	.1951	.1603	.3724	.0904
	10	10	.3487	.3471	.0555	.2598	.0555	.3897	.0658
	50	0	.0052	.0057	.0099	.0136	.0953	.0271	.0231
	50	1	.0052	.0058	.0101	.0136	.0092	.0269	.0229
	50	10	.0052	.0053	.0060	.0125	.0080	.0230	.0187
	50	50	.0052	.0050	.0025	.0097	.0050	.0146	.0099
.02	10	0	.8171	.8376	.1577	.3085	.5102	.6169	.2153
	10	1	.8171	.8242	.1469	.3403	.4783	.6496	.1828
	10	10	.8171	.8183	.0782	.5108	.3088	.7661	.0775
	50	0	.3642	.3716	.0695	.1932	.1723	.3863	.0470
	50	1	.3642	.3656	.0670	.1959	.1695	.3880	.0479
	50	10	.3642	.3618	.0530	.2168	.1488	.3975	.0509
	50	50	.3642	.3639	.0260	.2731	.0925	.4097	.0515
.01	10	0	.9044	.9146	.1240	.3280	.5772	.6560	.2556
	10	1	.9044	.9051	.1206	.3614	.5439	.6900	.2224
	10	10	.9044	.9054	.1550	.5534	.3522	.8301	.0869
	50	0	.6050	.6101	.0818	.2650	.3407	.5300	.0870
	50	1	.6050	.6047	.0799	.2699	.3359	.5344	.0829
	50	10	.6050	.6033	.0690	.3078	.2981	.5643	.0591
	50	50	.6050	.6059	.0370	.4077	.1983	.6116	.0308

Table 4.2
Truncated Geometric Distribution

p	c	n	m	θ_n	EV_n	$RMSEV_m$	$E\hat{\theta}_m$	$RMSE\hat{\theta}_m$	$E\hat{\theta}^*_m$	$RMSE\hat{\theta}_m$
.1	10	10	0	.0443	.0494	.0566	.0214	.0293	.0429	.0367
		10	1	.0443	.0457	.0530	.0230	.0283	.0439	.0357
		10	10	.0443	.0460	.0274	.0324	.0204	.0486	.0252
		50	0	.0075	.0078	.0108	.0047	.0051	.0093	.0087
		50	1	.0075	.0079	.0108	.0047	.0050	.0094	.0085
		50	10	.0075	.0071	.0087	.0047	.0051	.0086	.0078
		50	50	.0075	.0072	.0058	.0055	.0044	.0083	.0060
.5	10	10	0	.1302	.1454	.1036	.0708	.0691	.1416	.0717
		10	1	.1302	.1308	.0950	.0745	.0659	.1423	.0686
		10	10	.1302	.1292	.0515	.0968	.0467	.1452	.0512
		50	0	.0273	.0274	.0200	.0158	.0136	.0315	.0148
		50	1	.0273	.0268	.0189	.0160	.0132	.0317	.0141
		50	10	.0273	.0265	.0162	.0173	.0123	.0317	.0136
		50	50	.0273	.0274	.0103	.0211	.0091	.0317	.0109
.9	10	10	0	.3319	.3648	.1567	.1731	.1655	.3463	.0950
		10	1	.3319	.3145	.1386	.1876	.1515	.3582	.0920
		10	10	.3319	.3314	.0604	.2487	.0914	.3730	.0702
		50	0	.0096	.0107	.0131	.0161	.0078	.0322	.0242
		50	1	.0096	.0098	.0130	.0160	.0077	.0317	.0237
		50	10	.0096	.0097	.0083	.0156	.0073	.0286	.0205
		50	50	.0096	.0098	.0042	.0139	.0053	.0208	.0122

Table 4.2 (continued)

p	c	n	m	θ_n	EV_n	$RMSEV_m$	$E\hat{\theta}_m$	$RMSE\hat{\theta}_m$	$E\hat{\theta}^*_m$	$RMSE\hat{\theta}_m$
.1	100	10	0	.0443	.0494	.0566	.0214	.0293	.0428	.0367
		10	1	.0443	.0388	.0498	.0208	.0297	.0396	.0348
		10	10	.0443	.0438	.0270	.0310	.0213	.0465	.0251
		50	0	.0075	.0078	.0108	.0047	.0051	.0093	.0087
		50	1	.0075	.0075	.0106	.0045	.0051	.0089	.0083
		50	10	.0075	.0077	.0093	.0049	.0051	.0090	.0081
		50	50	.0075	.0074	.0057	.0057	.0042	.0086	.0058
.5	100	10	0	.1312	.1471	.1057	.0719	.0691	.1438	.0720
		10	1	.1312	.1305	.0951	.0759	.0656	.1448	.0686
		10	10	.1312	.1318	.0543	.0982	.0476	.1473	.0539
		50	0	.0283	.0290	.0212	.0163	.0141	.0327	.0157
		50	1	.0283	.0277	.0205	.0163	.0142	.0322	.0152
		50	10	.0283	.0284	.0163	.0180	.0126	.0329	.0140
		50	50	.0283	.0291	.0108	.0222	.0093	.0333	.0116
.9	100	10	0	.6095	.0269	.1890	.2505	.3628	.5011	.1511
		10	1	.6095	.6150	.1792	.2771	.3368	.5220	.1319
		10	10	.6095	.6084	.1051	.4011	.2161	.6017	.0861
		50	0	.1855	.1856	.0533	.1030	.0846	.2060	.0428
		50	1	.1855	.1856	.0509	.1059	.0816	.2098	.0435
		50	10	.1855	.1857	.0447	.1155	.0725	.2117	.0437
		50	50	.1855	.1857	.0269	.1410	.0477	.2115	.0367

While $\hat{\theta}_m$ may be an attractive estimator in the truncated geometric case in terms of its mean square error, it has already been noted that its bias can be considerable. In fact,

$$E(\hat{\theta}_m) = \theta_n + (n+m)^{-1}\{\binom{n}{2}\theta_{n-1} - \binom{n+1}{2}\theta_n\} + o((n+m)^{-1}).$$

This suggests that the quantity $\hat{\theta}_m + (n+m)^{-1}(\binom{n+1}{2}\hat{\theta}_m - \binom{n}{2}\sum_i \hat{p}_i\hat{q}_i^{n-1})$ would be a better estimator of θ_n than $\hat{\theta}_m$ alone. For the size of the samples discussed here, $\sum_i \hat{p}_i\hat{q}_i^{n-1}$ tends to underestimate θ_{n-1} too severely and a better estimator can be obtained by replacing $\sum_i \hat{p}_i\hat{q}_i^{n-1}$ by $\hat{\theta}_m$, leading to the estimator

$$\theta_m^* = \hat{\theta}_m\,(1 + n/(n+m)).$$

Values of $E(\theta_m^*)$ and RMSE (θ_m^*) are given in Tables 4.1-4.2. Generally, θ_m^* has good bias properties and compares favorably with V_m in terms of RMSE, even for the equiprobable case.

It should be noted that $\hat{\theta}_m$ and V_m are, in some sense, "retrodictors." That is, they predict, on the basis of n+m observations, what would be observed for the last m observations. In Starr (1979), an argument is given that this is not a vacuous exercise; V_m can be used effectively to predict, on the basis of an initial sample size n and a subsequent sample of size m, what will occur in a large future sample of size M. This argument applies equally well to the NPMLE $\hat{\theta}_m$. However, it can be argued that the principle interest of estimators such as $\hat{\theta}_m$ and V_m is in their properties as true predictors. For example, Rasmussen and Starr (1979) used the estimator $V_o = n^{-1}\sum_i^n I(X_i = 1)$ to consider a rule for sequentially sampling a population. Similarly, the estimators $\hat{\theta}$ and θ^* could also be used in such a capacity. We leave the examination of such sequential rules to a future paper and consider here only the properties of V_o, $\hat{\theta}_o$ and θ_o^* as estimates of θ_n. Simulation results appear

in Tables 4.1-4.2. In terms of mean square error, again we see that, in the equiprobable case, V_o dominates $\hat{\theta}_o$ and that θ_o^* compares favorably with V_o. In the truncated geometric case, both $\hat{\theta}_o$ and θ_o^* dominate V_o except when p is near one in which case V_o tends to be a better estimator than $\hat{\theta}_o$.

§5. Summary and Discussion

This paper has focused on nonparametric estimators of θ_n, the probability of discovering a new species. We have shown V_m to be a minimum variance unbiased estimator with a high rate of convergence in the equiprobable case. The nonparametric maximum likelihood estimator, $\hat{\theta}_m$, has similar asymptotic properties. In small samples, V_m is a better estimator than $\hat{\theta}_m$ in the equiprobable cell case with respect to mean square error; this is reversed for truncated geometric distributions when p is not large. An estimator with somewhat less bias than $\hat{\theta}_m$ is θ_m^*, defined in (4.1); it compares favorably with V_m in terms of mean square error.

Besides the theoretical interest in $\hat{\theta}_m$ as an estimator which competes well with V_m in the truncated geometric case, we argue that this has practical implications. For example, data collected by Andrews (1985) of the species abundance of epiphytic fungi on apple leaves fit a truncated geometric distribution quite well with p = .77. Arguments are given by Pielou (1977) that a geometric distribution, or more generally, a negative binomial distribution is appropriate in some situations for modeling species distributions. It remains to be seen how V_m and $\hat{\theta}_m$ compare over a wider class of distributions.

Appendix A. Proof of §2 Results

Lemma A.1

The parameter θ_n is homogeneous over Π and is of degree $n + 1$.

Proof:

Sufficient for the proof is to show that $\sum_i q_i^k$ is homogeneous of degree k. Since $E_p\left(\sum_i \prod_{j=1}^k I(X_j \neq i)\right) = \sum_i q_i^k$ for all $P \in \Pi$, we have that $\sum_i q_i^k$ is homogeneous of degree $\leq k$. We now suppose that $\sum_i q_i^k$ is homogeneous of degree h and show that $h \geq k$. Thus, assume there exists $\phi(x_1,\ldots,x_h)$ so that

$$\sum_i q_i^k = E_p\left(\phi(X_1,\ldots,X_h)\right) \tag{A.1}$$

for all $P \in \Pi$. Suppose Π_1 is a subset of Π so that

$$P_q(1) = q \text{ and } P_q(2) = 1 - q$$

and $\Pi_1 = \{P_q \in \Pi,\ 0 < q < 1\}$. With the choice of P_q the left-hand side of (A.1) is a polynomial in q of degree k while the right-hand side is a polynomial in q of degree, say, $h_1 \leq h$. Since these polynomials are must be of the same degree, we have $k = h_1 \leq h$. ∎

Define $h_{1n}(X_1) = E(h(X_1,\ldots,X_n)|X_1) - \theta_n$. The proof of Property 2.3 is complete with $\sigma^2 = (n+1)^2\,\mathrm{Var}(h_{1n}(X_1))$ and the following

Lemma A.2.

$$\mathrm{Var}(h_{1n}(X_1)) = \sum_i (p_i - (n+1)^{-1})^2\, p_i q_i^{2n-2} - (\theta_n - (1+n^{-1})^{-1}\theta_{n-1})^2.$$

Proof:

Use (2.2) to get

$$E\left(h(X_1,\ldots,X_{n+1})|X_1\right) = (n+1)^{-1} \sum_i q_i^{n-1}\{np_i(X_1 \neq i) + q_i I(X_1 = i)\}.$$

Thus, by rearranging terms

$$h_{1n}(X_1) = (n+1)^{-1} \sum_i q_i^{n-1}(p_i - (n+1)^{-1})(p_i - I(X_1 = i)).$$

Hence,

$$E\, h_{1n}(X_1)^2 = (n+1)^{-2} \{\sum_i p_i q_i^{2n-1}(p_i - (n+1)^{-1})^2$$

$$- \sum_{i \neq j} p_i q_i\, q_i^{n-1} q_j^{n-1}(p_i - (n+1)^{-1})(p_j - (n+1)^{-1})\}$$

which gives the result upon a rearrangement of terms. ∎

To prove Property 2.4, we need to examine the properties of the following projection of h,

$$h_{2n}(X_1,X_2) = E(h(X_1,\ldots,X_{n+1})|X_1,X_2) - h_{1n}(X_1) - h_{1n}(X_2) - \theta_n$$

$$= (1-\mu)^{n-2}(\mu - 2(n+1)^{-1})(I(X_1=X_2) - \mu). \qquad (A.2)$$

To see (A.2), first note that it is easy to check that $\theta_n = (1 - \mu)^n$ and that $h_{1n}(X_1) = h_{1n}(X_2) = 0$. Now, use (2.2) to get

$$E(h(X_1,\ldots,X_{n+1})|X_1,X_2)$$

$$= (n+1)^{-1}(1-\mu)^{n-2} \sum_i \{(n-1)\mu\, I(X_1 \neq i)\, I(X_2 \neq i)$$

$$+ (1-\mu)(I(X_1 = i)\, I(X_2 \neq i) + I(X_1 \neq i)\, I(X_2 = i))\}$$

$$= (1-\mu)^{n-2}\{1 - 2\mu n/(n+1) + (\mu - 2(n+1)^{-1})\, I(X_1 = X_2)\}$$

after some algebra. Subtracting θ_n yields (A.2). The proof of Property 2.4 is now an application of a result independently due to Gregory (1977) and Serfling (see Serfling, 1980, page 192).

Proof of Property 2.4:

Let $K = (1-\mu)^{n-2}(\mu - 2(n+1)^{-1})$ so that $h_{2n}(X_1,X_2) = K(I(X_1=X_2) - \mu)$. It is immediate that $\mathrm{Var}(h_{2n}(X_1,X_2)) = K^2\mu(1-\mu) > 0$. Now, let g be an arbitrary, measurable function such that $E(g(X))^2 < \infty$ and let x,λ be real constants. The forms of $g(\cdot)$ and λ satisfying

$$\lambda g(x) = E\{h_{2n}(x,X)g(X)\}$$

$$= K\mu\{\textstyle\sum_i I(x=i)g(i) - Eg(X)\} = K\mu\{g(x) - Eg(X)\}$$

are of two types. If $E\, g(X) \neq 0$, then

$$g(x) = E\, g(X)/(1 - \lambda/K\, \mu)$$

is a constant ($\neq 0$) and thus $\lambda = 0$. If $E\, g(X) = \mu \sum_i g(i) = 0$, then $\lambda = K\mu$. Thus, for example, by Serfling (1980, page 194), we have the result. ∎

Appendix B. Proof of §3 Results

Proof of Property 3.1:

Define $G(x) = x(1-x)^n$ and note that $\theta_n = \sum_i G(p_i)$ and that $\hat{\theta}_m = \sum_i G(\hat{p}_i)$. By a Taylor-series expansion,

$$\hat{\theta}_m = \theta_n + \sum_i (\hat{p}_i - p_i) G'(p_i) + O(\sum_i (\hat{p}_i - p_i)^2),$$

since $G'(x)$ is bounded for $0 < x < 1$. Now, since

$$(n+m)^{1/2} \, E \sum_i (\hat{p}_i - p_i)^2 = (n+m)^{-1/2} \sum_i p_i q_i$$

$$\leq (n+m)^{-1/2} \to 0, \tag{B.1}$$

we have that $(n+m)^{1/2} \sum_i (\hat{p}_i - p_i)^2 \to 0$ in probability. By Fubini's Theorem, we have that

$$\sum_i (\hat{p}_i - p_i) \, G'(p_i) = (n+m)^{-1} \sum_{j=1}^{n+m} \{\sum_i G'(p_i)[I(X_j = i) - p_i]\}.$$

This, the central limit theorem and Slutsky's theorem give the result. ∎

Proof of Property 3.2:

By a Taylor-series expansion,

$$\hat{\theta}_m = \theta_n + G''(\mu)/2 \sum_i (\hat{p}_i - \mu)^2 + O(\sum_i (\hat{p}_i - \mu)^3),$$

since $G''(x)$ is bounded for $0 < x < 1$ and $\sum_i (\hat{p}_i - \mu) = 0$. Similarly to (B.1), we have that $(n+m) \sum_i (\hat{p}_i - \mu)^3 \to 0$ in probability. Thus

$$(n+m)(\hat{\theta}_m - \theta_n) = (n+m) \, G''(\mu)/2 \sum_i (\hat{p}_i^2 - \mu^2) + o_p(1). \tag{B.2}$$

Now,

$$\sum_i (\hat{p}_i^2 - \mu^2) = \sum_i \left((n+m)^{-1} \sum_{j=1}^{n+m} I(X_j=i)\right)^2 - \mu$$

$$= (n+m)^{-1} + 2(n+m)^{-2} \sum_{j<k} I(X_j=X_k) - \mu$$

$$= (n+m)^{-1} (1-\mu) + \left(1 - (n+m)^{-1}\right) U, \qquad (B.3)$$

where $U = \binom{n+m}{2}^{-1} \sum_{j<k} I(X_j=X_k) - \mu$ is a U-statistic. As in the proof of Property 2.4, $E(U|X_1) = 0$ and

$$\binom{n+m}{2} E(U|X_1,X_2) = \left(I(X_1=X_2) - \mu\right).$$

Thus, by the same argument as in the proof of Property 2.4 (with K=1), we have

$$(n+m)\, U \to_D \mu(\chi^2-1).$$

This, (B.2), (B.3) and Slutsky's Theorem yields the result. ∎

Proof of Property 3.3

We need only show that $\sigma^2 = 0$ implies $p_i = p_j$ for each i,j. To do this we construct the random variable

$$X = (n+1)(np_i - q_i)\, q_i^{n-1} \qquad \text{with probability } p_i,\ i=1,2,\ldots$$

Now, it is easy to see that $\mathrm{Var}(X) = \sigma^2$ and thus, $\sigma^2 = 0$ means that $(n+1)(np_i - q_i)q_i^{n-1} = (n+1)^2(p_i - (n+1)^{-1})q_i^{n-1}$ must be some constant C for $i=1,2,\ldots$ Since the number of species exceeds n, we have $p_i < (n+1)^{-1}$ for some i and C must be nonpositive. The question of whether different p_i may satisfy $(n+1)^2\left(p_i - (n+1)^{-1}\right)q_i^{n-1} = C$ is equivalent to finding the number of roots of

$$h(x) = (n/(n+1) - x)x^{n-1} - C \qquad 0 < x < 1.$$

Now, $h'(x) = nx^{n-2}((n-1)/(n+1)-x)$ is positive for $0 < x < (n-1)/(n+1)$ and is negative for $(n-1)/(n+1) < x < 1$. Further, $h(0) = -C$ and $h(1) = -(n+1)^{-1} - C$. Thus, for $-(n+1)^{-1} < C < 0$ there is exactly one root and no roots for $C < -(n+1)^{-1}$. ∎

References

Andrews, J. (1985). Personal communication.

Banerjee, P. K. and Sinha, B. K. (1985). Optimal and adaptive strategies in discovering new species. Sequent. Anal. 4, 111-122.

Chao, A. (1981). On estimating the probability of discovering a new species. Ann. Statist. 9, 1339-1342.

Chao, A. (1982). Correction to "On estimating the probability of discovering a new species." Ann. Statist. 10, 1311.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species. Biometrika 63, 435-447.

Good, I. (1953). On the population frequencies of species and the estimation of population parameters. Biometrika 40, 237-264.

Good, I. (1965). The Estimation of Probabilities. Research Monograph 30, M.I.T. Press, Cambridge.

Good, I. and Toulmin, G. (1956). The number of new species and the increase of population coverage, when a sample is increased. Biometrika 43, 45-63.

Goodman, L. (1949). On the estimation of the number of classes in a population. Ann. Math. Statist. 20, 572-579.

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

Gregory, G. (1977). Large sample theory for U-statistics and tests of fit. Ann. Statist. 5, 110-123.

Halmos, P. (1946). The theory of unbiased estimation. Ann. Math. Statist. 17, 34-43.

Harris, B. (1959). Determining bounds on integrals with applications to cataloging problems. Ann. Math. Statist. 30, 521-548.

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

Hill, B. (1979). Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. J. Amer. Statist. Assoc. 74, 668-673.

Knott, M. (1967). Models for cataloging problems. Ann. Math. Statist. 38, 1255-1260.

Pielou, E. C. (1977). Mathematical Ecology. Wiley, New York.

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

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

Serfling, R. (1980). Approximation Theorems of Mathematical Statistics. Wiley, New York.

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

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2898	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* NONPARAMETRIC ESTIMATION OF THE PROBABILITY OF DISCOVERING A NEW SPECIES		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Murray K. Clayton and Edward W. Frees		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53705		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE January 1986 13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

U-statistics
nonparametric maximum likelihood estimator

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

A random sample is taken from a population consisting of an unknown number of distinct species. A quantity of interest is the probability of discovering a new species when an additional draw from the population is made. An estimator of this quantity was introduced by Starr (1979). We prove a conjecture of Starr's that the estimator is uniformly minimum variance unbiased and give various asymptotic properties of the estimator. A nonparametric maximum likelihood estimator is introduced which has similar asymptotic properties. A Monte-Carlo study is given which suggests guidelines for choosing an estimator under various circumstances.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

END

FILMED

6-86

DTIC