Estimating the Number of Species: A Review

J. Bunge; M. Fitzpatrick


Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199303%2988%3A421%3C364%3AETNOSA%3E2.0.CO%3B2-D


Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
Estimating the Number of Species: A Review

J. BUNG and M. FITZPATRICK*

How many kinds are there? Suppose that a population is partitioned into \( C \) classes. In many situations interest focuses not on estimation of the relative sizes of the classes, but on estimation of \( C \) itself. For example, biologists and ecologists may be interested in estimating the number of species in a population of plants or animals, numismatists may be concerned with estimating the number of dies used to produce an ancient coin issue, and linguists may be interested in estimating the size of an author’s vocabulary. In this article we review the problem of statistical estimation of \( C \). Many approaches have been proposed, some purely data-analytic and others based in sampling theory. In the latter case numerous variations have been considered. The population may be finite or infinite. If finite, samples may be taken with replacement (binomial sampling) or without replacement (hypergeometric sampling), or by Bernoulli sampling; if infinite, sampling may be multinomial or Bernoulli, or the sample may be the result of random Poisson contributions of each class. Given a sampling model, one may approach estimation of \( C \) via a parametric or nonparametric formulation; in either case there may be frequentist and Bayesian procedures. We begin by discussing the existing literature on this problem (over 120 references), organizing it by sampling model, population specification, and philosophy of estimation. We find that (a) the problem is quite resistant to statistical solution, essentially because no matter how many classes have been observed, there may still be a large number of very small unobserved classes; (b) many closely related estimation procedures have been developed independently and have not yet been compared; (c) there is not as yet a globally preferable estimator of \( C \), although for some models there is an acceptable estimator (for some not even this is true); and (d) there are promising directions for research to pursue; for example, it appears possible to exploit estimates of the “coverage” of the sample (the total proportion of the population represented by the observed classes) to improve the accuracy of estimators of the number of classes. Finally, we make specific recommendations for future research, regarding parametric estimation, coverage-based estimation, resampling methods, Poisson process representation of sampling models, and frequentist decision theory.

KEY WORDS: Abundance; Bayesian inference; Capture-recapture; Multinomial distribution; Number of classes; Numismatics; Occupancy numbers; Population size; Vocabulary size.

It took her three weeks to determine that she very probably would not live long enough to identify every kind of growth in those fourteen thicket acres. There were century plants, black mangrove, white mangrove, Australian pines, sea grape, punk trees, Brazilian pepper, bay cedar, grape, bayonets, cabbage palm, saw palmetto, wild coffee, greenbriar vines, marsh elder. There were varieties of live oak, a stand of them deep in the middle of the wild place, some of them huge, with low outspread limbs bigger around than her body. In the oaks were air plants in bewildering variety, some of them as big as bushel baskets. There were wild orchids, trailing strands of Spanish moss, strangler fig.


How many kinds are there? Suppose that a population, finite or infinite, is partitioned into \( C \) classes. In many situations interest focuses not on estimation of the relative sizes of the classes, but on estimation of \( C \) itself. For example, biologists and ecologists may be interested in estimating the number of species in a population of plants or animals; apart from their intrinsic interest, these estimates are needed to obtain extinction rates and so are essential in efforts to preserve biodiversity (Mann 1991). Numismatists may be concerned with estimating the number of dies used to produce an ancient coin issue; this information, when combined with an estimate of the number of coins struck per die, yields an estimate of the size of the coin issue and hence information about the ancient monetary economy (Stam 1987). Linguists may be interested in estimating the size of an author’s vocabulary (Efron and Thisted 1976). There are many other applications, including estimating the number of distinct records in a filing system where many records are duplicated (Arnold and Beaver 1988), undiscovered “observational phenomena” in astronomy (Harwit and Hildebrandt 1986), errors in a software system (Bickel and Yahav 1988), executions in South Vietnam (Bickel and Yahav 1985), and connected components in a graph (Frank 1978). The reader can no doubt supply further examples.

Although estimation of the relative class proportions is reasonably well understood when \( C \) is known, estimation of \( C \) itself appears to be quite difficult. In this review we discuss the latter problem. Throughout we focus exclusively on estimation of \( C \); unfortunately, we must leave aside many interesting related topics, such as stochastic abundance models (Engen 1978), measurement of “diversity” (Patil and Taillie 1982), and so on. We also avoid related areas such as capture-recapture problems (Pollock 1991) and estimation of the number of faults in software in continuous time, which, according to Naylor (1989, p. 191), “can be regarded as a continuous analogue of the problem of estimating the number of species in a biological population.” An extended bibliography on all of these subjects, with over 550 references, may be obtained from the authors on request.

This article is organized as follows. In Section 1 we review the existing literature, organized by statistical model (see Fig. 1). In Section 2 we discuss the current state of the art and suggest directions for future research.

1. KNOWN RESULTS

Suppose that we draw a sample of \( n \) items from a population partitioned into \( C \) classes, where \( C \) is unknown (and

* J. Bunge is Assistant Professor and M. Fitzpatrick is Research Assistant, Department of Economic and Social Statistics, New York State School of Industrial and Labor Relations, Cornell University, Ithaca, NY 14853-3901. The authors extend sincere thanks to the over 100 researchers who responded to requests for information and offer apologies to those researchers whom they were unable to contact. The authors also thank the New York State School of Industrial and Labor Relations of Cornell University for research funding, Linda Lowry for thorough database searches, Kathy B. Thomson for literature retrieval, Martin Wells for reading the manuscript and helpful suggestions, the editor for advice and encouragement, and an anonymous referee for detailed suggestions that helped us to improve the article's readability and organization.
hence the identities of the classes are not all known a priori). We assume that the classes can be identified once observed, so that members of the same class can be matched together in the sample. The outcome of this sampling theoretically can be represented by the random vector \( \mathbf{n} = [n_1, \ldots, n_C] \), where \( n_i \) is the number of sample items from the \( i \)th class, \( i = 1, \ldots, C \). But the \( i \)th class appears in the sample if and only if \( n_i > 0 \), and it is not known from the sample which \( n_i \)'s are zero. In short, \( \mathbf{n} \) is not observable. Instead, the observable random vector is \( \mathbf{c} = [c_1, \ldots, c_C] \), where \( c_j \) is the number of classes represented \( j \) times in the sample; that is, \( c_j = \#\{ n_i : n_i = j \} \), \( j = 1, \ldots, n \). Thus \( c_j \) is the number of “singletons,” \( c_2 \) is the number of “twins pairs,” and so on, in the sample. Good (1953) called the \( c_j \)'s “frequencies of frequencies.” The problem is to estimate \( C \) based solely on \( \mathbf{c} \). Figure 1 depicts one way of organizing the existing literature on this problem. In this section we review the literature following the tree diagram; for each topic we discuss first frequentist, then Bayesian approaches (if any). Herein \( c \) will denote the total number of classes in the sample, so that \( c = \sum_{j=1}^{n} c_j \); note that \( n = \sum_{j=1}^{c} n_j = \sum_{j=1}^{c} c_j \). Finally, \( H_c \) will signify the hypothesis of equal class sizes; that is, the assumption that all \( C \) classes are the same size.

### 1.1 Finite Population, Hypergeometric Sample

Suppose the population is finite with known size \( N \). This is realistic, for example, in sampling a database for duplicate records. Let \( N_i \) denote the number of units in the \( i \)th class, \( i = 1, \ldots, C \), \( \sum_{i=1}^{C} N_i = N \), and let \( M = \max_{1 \leq i \leq C} N_i \). If we sample \( n \) items without replacement from this population, then \( \mathbf{n} \) has a hypergeometric distribution with probability mass function (pmf) \( p_n(\mathbf{n}) = \binom{N}{\mathbf{n}}^{-1} \times \prod_{i=1}^{C} \binom{N_i}{n_i} \), \( \sum_{\mathbf{n} \leq N} n_i = n \). The pmf of the observable random vector \( \mathbf{c} \), \( p_c(\mathbf{c}) \), is simply \( p_n(\mathbf{n}) \) summed over all possible \( \mathbf{n} \) corresponding to \( \mathbf{c} : p_c(\mathbf{c}) = \sum_{\mathbf{n} \in S} p_n(\mathbf{n}) \), where \( S = \{ \mathbf{n} : \#\{ n_i = j \} = c_j, j = 1, \ldots, n \} ; \) it does not have a closed-form expression in general (Chapman 1951; Knott 1988; Shlomos 1981). For this model, Goodman (1949) showed that if it is known that \( n \geq M \), then there exists a unique unbiased estimator of \( C \), but without this knowledge no unbiased estimator exists. The estimator (when \( n \geq M \)) is

\[
\hat{C}_{\text{GOODMAN}} = c + \sum_{j=1}^{n} \frac{(-1)^{j+1}(N-n+j-1)!(n-j)!}{(N-n-j)!n!} c_j
\]

Unfortunately, although \( \hat{C}_{\text{GOODMAN}} \), being unique, is a fortiiori uniformly minimum variance unbiased (UMVU), in many cases its variance is so large as to render it unusable (\( \hat{C}_{\text{GOODMAN}} \) need not even be positive). This was observed by Goodman himself and by later investigators (Engen 1978; Frank 1978; Knott 1967). Hou and O'zsoyoglu (1988, 1991) found that “when the sampling fraction is lower than some ‘threshold,’ \( \hat{C}_{\text{GOODMAN}} \) always gives unreasonable estimates. On the other hand, once the sampling fraction is higher than the ‘threshold,’ \( \hat{C}_{\text{GOODMAN}} \) converges very fast” (1991, p. 644). In contrast, Shlomos (1981) took an asymptotic approach (independently of Goodman) in which \( N \to \infty \) in such a way that \( n/N \to q \in (0, 1) \). On this basis, he derived

\[
\hat{C}_{\text{SHLOMOS}} = c + c_1 \left( \sum_{i=1}^{n} iq(1-q)^{i-1} c_i \right)^{-1} \sum_{j=1}^{n} (1-q)^{j} c_j
\]

Note that \( \hat{C}_{\text{SHLOMOS}} \geq c \). Shlomos did not calculate the bias or variance of this estimator, but it performed reasonably well in his simulations for sampling fractions as low as 10%. Thus a biased estimator derived from asymptotic considerations apparently outperforms the UMVUE in this case, although there has been no formal comparison, and Shlomos’s simulations were not extensive.

On the Bayesian side, Hill (1968, 1979, 1980) adopted the following prior for \( (C, N_1, \ldots, N_C) \): \( \Pi(C, N_1, \ldots, N_C) = \Pi_{N_1} \cdots \Pi_{C,N} \), \( \Pi_{C,N}(C, N) = (\varnothing_{C,N}^{-1})^{-1} \Pi_{C,N}(C, N) \), where \( \Pi_{C,N}(C, N) \) is an arbitrary distribution on \( \mathbb{N} \times \mathbb{N} \). This prior is “noninformative” in that every partition of the \( N \) items into \( C \) classes is equally likely. From this Hill computed the bivariate posterior \( p_{C,N|\mathbf{c}}(C, N|\mathbf{c}) \), which turns out to depend on the data only through \( c \); that is, \( p_{C,N|\mathbf{c}}(C, N|\mathbf{c}) = p_{C,N|c}(C, N|c) \). Hill (1979) was...
primarily interested in \( p_{C,N|c}(C, N|c) \) itself. He considered estimation of \( C \) via the posterior mode, but concluded that this “should not be confused with the very different use made of such characteristics in Bayesian decision theory with a specified loss function” (Hill 1979, p. 672). (See also Marchand and Schroek 1982.)

1.2 Finite Population, Bernoulli Sample

Suppose now that the \( N \) items of the population enter the sample independently, each with probability \( p \); numismatists, for instance, use this as a model for the appearance of coins in a “hoard.” Then the total sample size is a binomial \((N, p)\) random variable, the \( i \)th class independently contributes a binomial \((N_i, p)\) random number of items to the sample, and \( p_C(n) = p^n(1 - p)^{N-n} \prod_{i=1}^{C} \binom{N_i}{n_i} \), \( \sum_{i=1}^{C} n_i = n \), but \( p_C(c) = \sum_{n=0}^{n} p_C(n) \) again does not have a closed-form expression. Goodman (1949) considered such Bernoulli sampling for known \( p \) and derived the unbiased estimator

\[
\hat{C}_{\text{GOODMAN2}} = c + \sum_{j=1}^{n} (-1)^{j+1} \left( \frac{1}{p} \right)^j c_j.
\]

The undesirable properties of \( \hat{C}_{\text{GOODMAN1}} \) are shared by \( \hat{C}_{\text{GOODMAN2}} \), as was noted again by Frank (1978), who gave its variance and a variance estimator. For Bernoulli sampling under \( H_+ \) with small \( p \) and large \( N \), Harwit (1981, app. A) proposed an approximate moment method estimator (see also Harwit and Hildebrand 1986); but, like \( \hat{C}_{\text{GOODMAN2}} \), Harwit’s estimator may be less than \( c \).

A more useful proposal is due to Esty (1985), who considered Bernoulli sampling in a “superpopulation” model where \( N_1, \ldots, N_C \) are iid negative binomial random variables with parameters \((\theta_1, \theta_2)\) and pmf \( p_{N|c}(N) = \Gamma(\theta_1 + N) / \Gamma(\theta_1) N ! \Gamma(\theta_2) (1 - \theta_2)^N / (\theta_1 \theta_2) \), denoted NB\((\theta_1, \theta_2)\). Then \( n_1, \ldots, n_C \) are iid NB\((\theta_1, \theta_2 / (\theta_1 + \theta_2))\) random variables; this is a special case of an invariant abundance distribution (Wani and Lo 1983a,b) (see Sec. 2). On this basis Esty derived the following estimator, for known \( \theta_1 \) (and \( \theta_2 \) are implicitly estimated),

\[
\hat{C}_{\text{NB}} = \frac{n}{\hat{c}},
\]

where \( \hat{\mu} \) is the root of the equation \( n / \mu = \mu / (1 - (1 + \mu / \theta_1)^{-\theta_1}) \). But Esty (1986b, p. 198) concluded from simulation results that \( \hat{C}_{\text{NB}} \) “is not recommended even for the populations for which it was derived”; as alternatives he suggested modifications (using the known value of \( \theta_1 \)) of certain estimators derived under \( H_+ \) (Sec. 1.3.1). He considered estimating \( \theta_1 \) via a further equation, but concluded that “my computer simulations show that the simultaneous estimation of \( \theta_1 \) must be judged unacceptable” (Esty 1985, p. 46). On the other hand, Wani and his coauthors studied parameter estimation in such invariant abundance models, but did not directly address estimation of \( C \). We discuss possible cross-fertilization of these ideas in Section 2.

1.3 Infinite Population, Multinomial Sample

Suppose now that we sample \( n \) items at random from an infinite population partitioned into \( C \) classes in proportions \( \pi = [\pi_1, \ldots, \pi_C], \sum_{i=1}^{C} \pi_i = 1 \). This is typically used in applications as an approximation to a finite population model, such as hypergeometric sampling (Sec. 1.1) with \( \pi_i = N_i / N \) and \( n \ll N \). In the infinite population case, \( n \) has a \( C \)-dimensional multinomial distribution, with pmf \( p_{n|c}(n) = \prod_{i=1}^{C} \pi_i^{n_i} \sum_{n} p_{n|c}(n) \) does not in general have a closed-form expression (Emigh 1983; Korwar 1988).

1.3.1 Infinite Population, Multinomial Sample, Equal Class Sizes. Although \( H_+ : \pi_1 = \cdots = \pi_C = C^{-1} \) is rarely realistic in applications, it is the most tractable case and has attracted the most attention. Questions about multinomial sampling under \( H_+ \) form part of the classical “occupancy” and “coupon collector’s problems” and have a vast literature (Holst 1986). It is difficult to give an authoritative account of estimating \( C \) in this case, but we will venture the following.

The MLE and UMVUE. The maximum likelihood estimator (MLE) of \( C \) in this case, \( \hat{C}_{\text{MLE}} \), is approximately equal to the solution \( C^* \) of the equation

\[
c = C^*(1 - e^{-n/C^*})
\]

(Good 1950, p. 73). Lewontin and Prout gave the variance of \( \hat{C}_{\text{MLE}} \) and discussed associated confidence intervals, pointing out that “any departure from equal representation of classes in the population makes \( \hat{C}_{\text{MLE}} \) an underestimate of \( |C| \” (Lewontin and Prout 1956, p. 213, our emphasis). Alternative expressions and approximations for \( \hat{C}_{\text{MLE}} \), along with associated confidence intervals, have been given by Darroch (1958), Driml and Ullrich (1967), Johnson and Kotz (1977), McNeil (1973), Holst (1981), Ivchenko and Timofina (1983), Esty (1986b), and Arnold and Beaver (1988). Darroch (1958) also showed that if it is known that \( n \geq C \) (he thought this unlikely), then the UMVUE of \( C \) is

\[
\hat{C}_{\text{UMVUE}} = \frac{S_{i,\text{LM}}}{S_{i,\text{LM}}} = \frac{S_{i,\text{LM}}}{S_{i,\text{LM}}},
\]

where \( S_{i,j} \) is the Stirling number of the second kind (Charalambides and Singh 1988; Harris 1968; Johnson and Kotz 1977). Holst (1981) showed that asymptotically \( \hat{C}_{\text{MLE}} \approx \hat{C}_{\text{UMVUE}} \), so the latter seems to offer no advantage.

Coverage. The coverage, \( u \), of a sample is the (random) sum of the \( \pi_i \)’s corresponding to the observed classes (i.e., \( u = \sum_{i=1}^{C} \pi_i n_i > 0 \), where \( I(A) \) is the indicator of the event \( A \)). There is a growing literature on coverage; see, for example, Betro and Ziliński (1987), Bickel and Yahav (1986), Chao (1981), Clayton and Frees (1987), Cohen and Sackrowitz (1990), Engen (1975), Esty (1983, 1986a), Lee (1989), Lo (1992), Robbins (1968), Starr (1979), Tiwari and Tripathi (1989), Trybula (1958), and Yatracos (1991). Under \( H_+ \), \( u = c/C \); so given an “estimator” (predictor) \( \hat{u} \) of \( u \), an estimate of \( C \) is \( c/\hat{u} \). The first such \( \hat{u} \) was proposed by Good (1953) and by Good and Toulmin (1956), who studied \( \hat{u}_{\text{GOOD}} = (1 - c/n) \). This in fact can be used without assuming \( H_+ \) (Esty 1983). (See Sec. 1.3.3.) Under \( H_+ \), then, one obtains

\[
\hat{C}_{\text{COV}} = \frac{c}{\hat{u}_{\text{GOOD}}},
\]
This was studied by Darroch and Ratcliff (1980), who derived it from Robbins' (1968) work (independently of Good). Darroch and Ratcliff found that the easy-to-calculate \( \hat{C}_{\text{COV}} \) has high asymptotic efficiency relative to \( \hat{C}_{\text{MLE}} \), and hence may be preferable in applications. Esty (1982) also derived a related coverage-based estimator for use with low-coverage samples. We return to the subject of coverage in Section 2.2.

Other Approaches. Emigh (1983) proposed a rather complicated stepwise estimation procedure based on a hybrid of \( H \) and the nonparametric model (no restrictions on \( \pi \)). Esty (1984) studied the properties of an estimator proposed by Brown (1955) that uses the total number of pairs in the sample (this is larger than \( c_2 \)). Neither procedure is preferable to \( \hat{C}_{\text{MLE}} \) or \( \hat{C}_{\text{COV}} \); when \( H \) holds.

On the Bayesian side, Marchand and Schroek (1982) derived the posterior distribution of \( C \) given an arbitrary prior \( \Pi_C(C) \). For the improper uniform prior \( \Pi_C(C) = 1 \), the authors computed the posterior mean and variance (these have rather complicated expressions), and Schroek (1981) tabulated the results. Esty (1986b, p. 202) considered this method but concluded that “it cannot be expected to improve upon \( \hat{C}_{\text{MLE}} \).”

1.3.2 Infinite Population, Multinomial Sample, Parametric Models. In applications the class sizes are rarely believed to be equal; indeed, the difficulty is often precisely that some classes are much smaller than others. Two types of parametric models have been introduced to account for this. In the first type, one assumes that the \( \pi_i \)'s have a functional form, possibly depending on a small number of parameters; that is, \( \pi_i = f(i; \theta, C) \), \( i = 1, \ldots, C \), where \( \theta \) is a parameter vector and \( f \) is decreasing in \( i \). Kalinin (1965) studied the “Zipf” model \( f_Z(i; \theta, C) = \theta^i / i \) and the “Mandelbrot” model \( f_M(i; \theta, C) = \theta_i / (\theta_i + \theta) \). He derived families of moment-method estimators (based on \( e \) of \( \theta, C \)) in \( f_Z \) and \( f_M \), but the estimators are complicated and he did not calculate their variances. Subsequently, McNeil (1973), independently of Kalinin, considered asymptotic approximations to models of this type, as follows. Assume that as \( C \to \infty \), \( \pi_i(C) \to \alpha g(x) \), where \( g \) is some function and \( \alpha \) is finite, so that \( \pi_i \) is approximately proportional to \( g(i/C) \). In particular, for \( g(x) = 1, x^{-1} \), or \( x^{-k} \), we obtain approximations to \( H_{\text{MLE}}, f_Z, \) or \( f_M \) (the last with \( \alpha \) = 0). McNeil derived an estimator \( \hat{C}_{\text{MCNEIL}} \) (which depends on the given \( g \)) and its asymptotic distribution; again the procedures are complicated, and their small-sample behavior is not well understood.

In the second type of parametric model, one approximates the histogram of the \( \pi_i \)'s by a probability density function (pdf) that depends on some parameters \( \theta \). Sichel (1986a, b, c, 1992a, b, in press) developed such a model in detail, using the generalized inverse Gaussian (GIG) pdf (Atkinson and Yeh 1982; Sichel 1975, 1982; Stein, Zucchini, and Juritz 1987; Willmot 1987, 1988a). This has three parameters, but Sichel found that typically one could be fixed in his applications, yielding the inverse Gaussian pdf \( \psi(\pi; \theta_1, \theta_2) = \theta_1 \sqrt{2 \pi} \exp \left( \frac{-\pi - \pi \theta_2 - \theta_1^2 \theta_2 / (4 \pi)}{2 \theta_1^2 \theta_2} \right) \). Sichel (1986b, c) showed that if the distribution of the \( \pi_i \)'s is approximately given by \( \psi(\pi; \theta_1, \theta_2) \) (for some \( \theta_1, \theta_2 \)), then

\[ C \approx (E_0(\pi))^{-1} = 2(\theta_1, \theta_2)^{-1} \]

He then derived an estimate \( \hat{C}_{\text{SICHEL}}(\theta_1, \theta_2) \) of \( \theta_1, \theta_2 \) based on \( (c, c_1) \), using the fact that the probability that an arbitrary class will occur \( j \) times in a sample of size \( n \) is approximately given by a \( \psi \)-mixed Poisson distribution, and hence obtained

\[ \hat{C}_{\text{SICHEL}} = \frac{2}{\hat{\theta}_1 \hat{\theta}_2}, \]

Sichel also found the asymptotic bias and variance of \( \hat{C}_{\text{SICHEL}} \); he also studied its small-sample behavior via simulation, finding that “some strong biases appear in . . . estimation of \( [C] \). It appears that, for this particular purpose, sample sizes should be larger than 1,500 . . . ” (Sichel 1986b, p. 947). Recently, Sichel (1992a, b, c) studied a similar application of the more flexible GIG distribution, estimating its parameters by a minimum chi-squared fitting procedure. Burrell and Fenton (1993) gave a maximum likelihood estimation procedure for the parameters of the zero-truncated GIG-mixed Poisson distribution.

Of the two types of parametric models just described, the second seems clearly preferable for applications; it is less restrictive, since it does not seek to specify \( \pi \) exactly, and its estimation theory is better developed and more promising, as will be seen in Section 2.2.

There have been substantial parametric Bayesian efforts in this situation. First, in Hill's model of Section 1.1, let \( N \) be fixed and let \( \Pi_{\text{COV}}(C, N) = \Pi_C(C) \) be a truncated negative binomial such that \( \Pi_C(C) \propto (C^{-1})^{\theta_1} C^{\theta_2 + 1} \) \( \theta_1 \in (0, \infty), \theta_2 \in (0, 1] \). Hill (1979) obtained a posterior distribution for the infinite population case by allowing \( N \to \infty \) in this model. Lewins and Joanes (1984) extended Hill's results by proceeding directly from an infinite population assumption; they adopted the prior \( \Pi_{\text{COV}}(C, \pi) = \mathcal{D}(\pi; C, \theta_3) \Pi_C(C) \), where \( \mathcal{D}(\pi; C, \theta_3) \) is the \( C \)-dimensional symmetric Dirichlet density with parameter \( \theta_3 \in (0, \infty) \) and \( \Pi_C(C) \) is an arbitrary prior on \( N \). From this they derived the posterior \( p_{\text{COV}}(C | c) = p_{\text{COV}}(C | c) \); in particular, if \( \Pi_C(C) \) is the negative binomial as in Hill's model, then

\[ p_{\text{COV}}(C | c) = \frac{C + \theta_1 - 1}{c + \theta_3 n - 1} \]

Lewins and Joanes took the mode of \( p_{\text{COV}}(C | c) \) to be their estimate of \( C \). They found that “the model appears to be very robust for \( \theta_1 \) and \( \theta_3 \),” and “variations in \( \theta_3 \) clearly have the largest effect on the model,” and they discussed various methods of dealing with the unknown parameters (Lewins and Joanes 1984, p. 327). (See also Boender and Rinnooy Kan 1987; Zielinski 1981.) Finally, Keener, Rothman, and Starr (1987) took an “empirical Bayes” approach; assuming that \( \pi \sim \mathcal{D}(\pi; C, \theta_3) \), where \( \theta_3 \) is known, they computed \( p(c; C, \theta_3) \), the unconditional distribution of \( c \). From this they derived the MLE and other estimators of \( C \). They also considered joint estimation of \( \theta_3 \) and \( C \), finding that under some conditions the estimates are sensitive to the choice of \( \theta_3 \) (when it is given) and that estimation of \( \theta_3 \) may give undesirable results, although under other conditions the procedures seemed to perform well (see also Chen 1980, 1981a, b).
1.3.3 Infinite Population, Multinomial Sample, Nonparametric Model. Can $C$ be estimated with no assumptions about $\pi$? No unbiased estimate exists; Engen (1978, p. 28) showed that the bias of any estimator of $C$ based on $c$ is unbounded over the set of possible populations. As Harris (1959, p. 538) put it, “there is no way for the experimenter to establish the non-existence of an arbitrarily large number of classes each with negligible probability.” Nevertheless, Chao (1984), following Harris (1959), used estimates of the moments $E(c_i)$ to obtain a nonparametric estimated lower bound for $C$,

$$\hat{C}_{\text{CHAO1}} = c + \frac{c^2}{2c_2},$$

and also discussed associated bootstrap confidence intervals. Subsequently, Chao and Lee (1992) used the idea of coverage to derive the nonparametric estimator

$$\hat{C}_{\text{CHAO2}} = \frac{c}{\mu_{\text{GOOD}}} + \frac{n(1 - \mu_{\text{GOOD}})}{\mu_{\text{GOOD}}} \hat{\gamma}^2,$$

which is $\hat{C}_{\text{COV}} - \mu_{\text{GOOD}} +$ plus a bias correction term that depends on an estimate $\hat{\gamma}$ of the coefficient of variation $\gamma$ of the $\pi_i$’s. They also gave a corresponding variance estimator and carried out a simulation study, finding that “the proposed estimators for non-equiprobable cases are generally biased downward due to the underestimation of $[\gamma]$” (Chao and Lee 1992, p. 216). Despite this bias, there is some empirical and theoretical evidence that $\hat{C}_{\text{CHAO2}}$ is preferable to $\hat{C}_{\text{CHAO1}}$ for applications, although the two procedures have not been formally compared. We discuss coverage-based estimation further in Section 2.2.

1.4 Infinite Population, Poisson Sample

Now suppose that the number of representatives of the $i$th class in the sample is a Poisson random variable with mean $\lambda_i, i = 1, \ldots, C$, and these variables are independent. This model was proposed by Fisher (Fisher, Corbet, and Williams 1943), in connection with a study involving species of Lepidoptera (see also Engen 1978, p. 10). In this case the total sample size is a Poisson random variable with mean $\sum_{j=1}^n \lambda_j$ and $p_0(n) = e^{-\lambda} \prod_{j=1}^n \frac{\lambda_j^n}{n!}$, but again $p_0(c) = \sum_S p_S(n)$ does not have a closed-form expression. We assume in addition that the $\lambda_i$’s are themselves a random sample from some distribution $F$, then $E(c) = C(1 - p_0(F))$, where $p_0(F)$ is the probability that an $F$-mixed Poisson random variable is equal to 0. Thus, given an estimate $\hat{p}_0(F)$ of $p_0(F)$, an estimator of $C$ is

$$\hat{C}_{\text{POISSON}} = \frac{c}{1 - \hat{p}_0(F)}.$$

Ord and Whitmore (1986) (independently of Sichel’s 1986 papers) studied such a model in which $F$ is the inverse Gaussian distribution $\psi(\theta_1, \theta_2)$. They derived $\hat{C}_{\text{POISSON}}$ for this case, using $p_0(F) = p_0(\hat{\theta}_1, \hat{\theta}_2)$, where $(\hat{\theta}_1, \hat{\theta}_2)$ are the MLE’s of $(\theta_1, \theta_2)$ based on $c$. The authors analyzed several data sets and concluded that “the fit of the Poisson-inverse Gaussian distribution . . . suggests that it may be an appro-

priate model for species abundance data under specific conditions but will not always be the clear choice” (Ord and Whitmore 1986, p. 865). Alternatively, Zelterman (1988) (independently of Sichel and Ord and Whitmore) gave a family of estimators of $p_0(F)$ (based on $c$) that are robust with respect to $F$’s whose variance is not too large. The simplest is $\hat{p}_0(F) = \hat{p}_0 = \exp(-2c_2/c_1)$, and he based $\hat{C}_{\text{POISSON}}$ on $\hat{p}_0$. Zelterman discussed asymptotics, robustness, and efficiency of his estimators, concluding that “while it is not advocated that these . . . estimators replace standard techniques . . . it is clear that they can be useful under certain circumstances” (Zelterman 1988, p. 235). Zelterman’s version of $\hat{C}_{\text{POISSON}}$ has not yet been formally compared to Ord and Whitmore’s version. (See also Burrell 1989.)

Efron and Thisted (1975, 1976) took a “nonparametric empirical Bayes” approach to this model. (See also Thisted and Efron 1978.) Noting that $C_0(F)$ is the expected number of unobserved classes, Efron and Thisted defined a linear program that finds a value $C^*$ and a distribution $F^*$ that minimize $C_0(F)$ subject to constraints derived from $c$. This yields an estimated lower bound for $C$,

$$\hat{C}_{\text{ET}} = \hat{c}_{\text{adj}} + C^* p_0(F^*),$$

where $\hat{c}_{\text{adj}}$ is an adjusted version of $c$. Efron and Thisted found that, although this bound is “reasonably conservative, . . . without a parametric model the data give very little information” about $C$ (Efron and Thisted 1976, p. 446). This procedure has not been pursued further, possibly due to its complicated nature.

1.5 Infinite Population, Multiple Bernoulli Sample

Suppose that an infinite population (partitioned into $C$ classes) is observed on each of $n$ “occasions” (or by each of $n$ “observers”), and on each occasion each class either is or is not observed. This model originated with capture-recapture experiments but has been used for estimation of $C$. The sample can be represented by the $C \times n$ matrix $[x_{ni}]$, where $x_{ni} = 1$ (i $\text{th}$ class is observed on $i$ $\text{th}$ occasion), $i = 1, \ldots, C, n = 1, \ldots, n$. Only rows with at least one 1 are observable; in fact, $c_j$ is the number of rows with exactly $j$ 1’s. Burnham and Overton (1978, 1979) studied such a model in which the $x_{ni}$’s are all independent, with $P\{x_{ni} = 1\} = \pi_i$ and $n = 1, \ldots, n$ (the probabilities are the same on each occasion). Then the contribution of each class to the sample, $n_i = \sum_{i=1}^n x_{ni}$, is a binomial $(n, \pi_i)$ random variable, $p_0(n_i) = \prod_{i=1}^n \left(\frac{\pi_i^n}{n!}\right)(1 - \pi_i)^{n - n_i}$, and $p_0(c) = \sum_S p_S(n)$ again does not have a closed-form expression. Taking the $\pi_i$’s to be a random sample from some distribution $F$, they developed a $k$th order jackknife estimator $\hat{C}_{\text{BOK}}$ of $C$; for example

$$\hat{C}_{\text{BO1}} = c + \left(\frac{n - 1}{n}\right)c_1.$$
(see also Castledine 1981; Lloyd and Yip 1991; Yip 1989, 1991a,b).

The picture concerning estimation in this case is unclear. Doss and Sethuraman (1989) showed that when no unbiased estimator exists, attempts to reduce the bias of a biased estimator (e.g., by jackknifing) necessarily result in increased variance; indeed, \( \hat{C}_{\text{BOK}} \) has the form of an alternating series and can be unstable. Esty (1986b, p. 209) stated that \( \hat{C}_{\text{BOK}} \) "relies heavily on a very inaccurate polynomial approximation to the function \( 1/x \)" and "should not be used." On the other hand, Chao's (1987) simulations showed that \( \hat{C}_{\text{CHAO1}} \) is preferable to \( \hat{C}_{\text{BOK}} \) in some situations but not in others. Yip compared his procedure with both \( \hat{C}_{\text{BOK}} \) and \( \hat{C}_{\text{CHAO1}} \) via simulation, concluding that "the main advantages of the martingale method are in computational simplicity, conceptual simplicity and flexibility" (Yip 1991c, p. 357). But Yip's procedure is not fully developed; there are unsolved problems regarding, for example, estimation of the parameters of the beta distribution. Alternative versions and extensions of this model have been studied by Bickel and Yahav (1985, 1988) and Chao and Lee (1990), but the results cannot yet be termed conclusive.

### 1.6 Data-Analytic Methods

#### 1.6.1 Extrapolation of Curves

For many of the models discussed previously, one can in principle derive a "coupon collector's," "type-token," or "species-area" curve; that is, the graph of the expected number of observed classes as a function of the sample size \( n \), denoted \( E(c^{(n)}) \). For example, in the unrestricted multinomial model,

\[
E(c^{(n)}) = C - \sum_{i=1}^{C} (1 - \pi_i)^n.
\]

Note that as \( n \) increases, \( E(c^{(n)}) \) approaches \( C \). This suggests that, given the form of such a curve and a given observed values of \( c \) for several sample sizes, we might be able to estimate \( C \) by extrapolation, without reference to a sampling-theoretic model. In other words, suppose that we assume only that: (a) \( E(c^{(x)}) = f(x; \theta) \), where \( x \) is a measure of the "size" of the sample (not necessarily the count of items), \( \theta \) is a parameter vector, and \( f \) is a given increasing bounded function of \( x \); and (b) \( \lim_{x \to \infty} f(x; \theta) = C(< \infty) \). If an estimate \( \hat{\theta} \) of \( \theta \) can be obtained from the data \( c^{(x_1)}, c^{(x_2)}, \ldots \) (where \( c^{(x)} \) is an observed value of \( c \) with sample size \( x_j \)), then our estimate of \( C \) will be \( \lim_{x \to \infty} f(x; \hat{\theta}) \). Brainerd (1972) proposed several such \( f \)'s for estimation of literary vocabulary, the simplest of which is

\[
E(c^{(x)}) = f_{\text{BR}}(x; \theta) = \frac{1 - \theta^x}{1 - \theta},
\]

\( \theta \in (0, 1) \), where \( x \) is the number of words in a text. Then \( \lim_{x \to \infty} f_{\text{BR}}(x; \theta) = 1/(1 - \theta) \). Brainerd considered "admittedly ad hoc methods of estimation" of \( \theta \) (Brainerd 1972, p. 517), obtaining \( \hat{C}_{\text{BR}} = 1/(1 - \hat{\theta}) \) (see also Brainerd 1981, 1982; Najock 1986). Alternatively, De Caprariis, Linde-
rameters, finally concluding that "as to the best method of estimating \([C]\), we provisionally keep our preference for the ML estimate under [the invariant model], but intend to do new research on the . . . large variation of [the estimate of \(\theta_1\)] and its influence" on the estimate of \(C\) (Stam 1987, p. 170).

Esty (1986b) carried out an extensive comparative simulation study of \(\hat{C}_{\text{MLE}}\), \(\hat{C}_{\text{COV}}\), Brown's "pairs" estimator (Sec. 1.3.1), modifications of these for a negative binomial population (Sec. 1.2) with \(\theta_1 = 2\), and \(\hat{C}_{\text{NB}}\) for the same population. He used Bernoulli sampling and also nonrandom sampling from various populations, finding that \(\hat{C}_{\text{COV}}\) "is less sensitive to nonrandomness and is to be preferred" to \(\hat{C}_{\text{MLE}}\) (Esty 1986b, p. 198). He recommended reporting (essentially) \(\hat{C}_{\text{COV}}\) while mentioning \(H_\ast\) and, if desired, reporting also a modification of \(\hat{C}_{\text{COV}}\), adjusted for the negative binomial population with a known (or postulated) value of \(\theta_1\).

For our part, we believe that all of the estimators discussed in Section 1 have drawbacks that raise reasonable doubts about experimental conclusions in most cases. Nevertheless, in the absence of precise information regarding the sampling plan and population structure, our provisional choice of estimator would be \(\hat{C}_{\text{CHAOS}}\), which hopefully will preserve the desirable properties of \(\hat{C}_{\text{COV}}\) (such as robustness against deviation from the sampling plan) while adjusting upwards to account for inhomogeneity in class sizes.

2.2 Recommendations for Future Research

Researchers have not been particularly sanguine about the prospects for estimating the number of classes under minimal assumptions. For example, Esty (1986b, p. 198) stated that "a good estimator should be able to estimate \([C]\) when the distribution is unknown. Unfortunately, this is impossible . . . A single . . . estimate cannot be proper for all the various possible . . . distributions." In fact, Good stated that "I don't believe it is usually possible to estimate the number of unseen species . . . but only an approximate lower bound to that number. This is because there is nearly always a good chance that there are a very large number of extremely rare species" (I. J. Good, personal communication, May 13, 1991). Undeterred by these pessimistic assessments, however, we devote the remainder of this article to speculative recommendations for future research directions. We are responsible for all opinions expressed.

Parametric Models. The trouble with the parametric models discussed in Section 1 is that the nuisance parameter \(\theta\) must be estimated jointly with \(C\), and authors have typically been dissatisfied with the results. But a substantial literature deals with estimation of \(\theta\) per se, and the results have not yet been exploited for estimation of \(C\). Wani and his coauthors (Wani and Lo 1983a,b, 1986; Wani and Watterson 1982) have studied a superposition model like that of Section 1.2, in which \(N_1, \ldots , N_C\) are iid random variables with distribution \(F_\ast\). If, under Bernoulli sampling, \(n_1, \ldots , n_C\) (the observed numbers of representatives of each class) are iid random variables with distribution \(F_\ast\), (the same form but different parameter), then \(F\) is called an invariant abundance distribution; the negative binomial model of Section 1.2 is a special case. (See also Willmot 1988b, sec. 4.) Lo and Wani (1983) derived ML's of \(\theta\) for such distributions and studied their properties in detail, but their results have not yet been incorporated directly into estimators of \(C\). Also, in regard to Sichel's GIG mixture model (Sec. 1.3.2), Stein et al. (1987) and Willmot (1988a) derived an orthogonal reparameterization for the closely related GIG-mixed Poisson distribution, which has an advantage in that the ML's of the new parameters are asymptotically uncorrelated; again, the results have not yet been used in estimation of \(C\).

Coverage. In general, estimators (predictors) of the sample coverage are better understood and better behaved than are estimators of \(C\) (whether \(H_\ast\) holds or not), and coverage-based estimators of \(C\) under \(H_\ast\) exploit this fact. Chao's attempt (Sec. 1.3.3) to account for unequal class sizes by adjusting \(\hat{C}_{\text{COV}}\) based on a further function of \(u_{\text{GOOD}}\) represents a logical next step. But there are other procedures for estimating the coverage (Sec. 1.3.2); for example, Lee (1989) proposed the nonparametric MLE of the expected coverage. Furthermore, Chao's method is not the only possible way to incorporate the coverage into the estimate of \(C\). Thus there is ample scope for development of coverage-based estimation without assuming \(H_\ast\).

Resampling Methods. Although bias reduction via the jackknife may not be appropriate in this problem (Sec. 1.5), variance estimation via the bootstrap is not excluded a priori. The only author to mention such a procedure has been Chao (1984) (see Sec. 1.3). Clearly, estimation of \(C\) is a nonstandard problem with respect to the bootstrap; however, there have been recent advances in resampling methods for complex sampling plans (Hall 1991; Kovar, Rao, and Wu 1988; Rao and Wu 1988; Wu 1991). It may be possible to use these to devise general bootstrap confidence intervals for \(C\); the question has yet to be explored.

Unification of Sampling Theory. The reader may have observed that the various sampling models discussed—hypergeometric, Bernoulli, multinomial, Poisson, multiple Bernoulli—are all approximations of each other. This can be clarified by representing "species sampling" as the superposition \(\mathbb{P}\) of \(C\) independent homogeneous Poisson processes \(\mathbb{P}_1, \ldots , \mathbb{P}_C\) on \(\mathbb{R}^+\), with rates \(\lambda_1, \ldots , \lambda_C\). The interpretation is that the \(i\)th class contributes items to the sample according to \(\mathbb{P}_i\). If \(\mathbb{P}\) is observed during \([0, t]\), then we recover Poisson sampling (Sec. 1.4) with parameters \(\lambda_1t, \ldots , \lambda_Ct\). If we condition on the number of events (the sample size) in \([0, t]\), then we recover multinomial sampling (Sec. 1.3) with \(\pi_i = \lambda_i / \sum_{j=1}^C \lambda_j\). These in turn can be regarded as approximations to the Bernoulli and hypergeometric models. Other authors have used this idea in some form (Efron and Thisted 1976; McNeil 1973), but a unified sampling theory based on it has not yet been produced. It seems that such a theory would be quite useful, because a great deal is known about superpositions of Poisson processes, particularly in terms of asymptotics (see, for example, Cox and Isham 1980, sec. 4.5). The model could also be extended to accommodate randomly generated \(\lambda_i\)'s. Such an investigation could explain, for example, the relationship between \(\hat{C}_{\text{POISSON}}\) and coverage-
based estimators like $C_{\text{COV}}$. Recently, M. T. Chao (1992) derived approximate moments of the number of classes observed in such a $P$ during $[0,1]$, given certain conditions on the $\lambda$'s, and Nayak and Christman (in press) gave majorization results to quantify estimation bias in this model as the $\lambda$'s become more unequal. But the general sampling problem remains open.

**Frequentist Decision Theory.** Our impression is that most authors have approached the problem of estimation of $C$ from a relatively applied point of view, deriving estimators on the basis of a particular formulation of the problem and then studying their properties. As a consequence, decision-theoretic issues such as sufficiency, ancillarity, and admissibility, although not entirely neglected, have not played an important role. It seems that understanding the problem would be enhanced by a study of such issues, both within individual models and comparatively across models. Cohen and Sackrowitz (1990) initiated this type of investigation in the coverage problem, with promising and somewhat surprising results. To take just one example concerning estimation of $C$, we know that no unbiased estimator exists in the unrestricted multinomial model. Liu and Brown (in press) have given decision-theoretic results concerning the deep structure of such “singular” problems.

There are certainly many interesting open questions for Bayesian research, but we will disqualify ourselves from making such recommendations, due to lack of expertise. To conclude, although entirely satisfactory results have not yet been obtained, there is much scope for accomplishment in this problem, and we encourage the reader to contribute his or her effort to its solution.

[Received November 1991. Revised August 1992.]

**REFERENCES**


——— (in press), "Modelling Species-Abundance Frequencies and Species-Individual Functions with the Generalized Inverse Gaussian-Poisson Distribution Law," *Ecology*.