Estimating the Number of Species: A Review

J. Bunge; M. Fitzpatrick

Journal of the American Statistical Association, Vol. 88, No. 421. (Mar., 1993), pp. 364-373.

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

Journal of the American Statistical Association is currently published by American Statistical Association.

Estimating the Number of Species: A Review

J. BUNGE 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 (multinomial 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 thickety 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.
>
> —*Condominium,* by John D. MacDonald (1977).

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 Hildebrand 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 Nayak (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 Thorson 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.

Journal of the American Statistical Association
March 1993, Vol. 88, No. 421, Review Paper

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 $\mathbf{n} = [n_1 \cdots n_C]'$, where n_i = the number of sample items from the ith class, $i = 1, \ldots, C$. But the ith 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 $\mathbf{c} = [c_1 \cdots c_n]'$, where c_j = *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_1 is the number of "singletons," c_2 is the number of "twin 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 $n = \sum_{i=1}^{C} n_i = \sum_{j=1}^{n} jc_j$. Finally, $H_=$ 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 ith class, $i = 1, \ldots, C$, $\sum_{i=1}^{C} N_i = N$, and let $M = \max_{1 \le l \le C} N_i$. If we sample n items at random without replacement from this population, then $\mathbf{n}$ has a multiple hypergeometric distribution with probability mass function (pmf) $p_{\mathbf{n}}(\mathbf{n}) = \binom{N}{n}^{-1} \times \prod_{i=1}^{C} \binom{N_l}{n_l}$, $\sum_{i=1}^{C} n_i = n$. The pmf of the *observable* random vector $\mathbf{c}$, $p_{\mathbf{c}}(\mathbf{c})$, is simply $p_{\mathbf{n}}(\mathbf{n})$ summed over all points $\mathbf{n}$ corresponding to $\mathbf{c}$: $p_{\mathbf{c}}(\mathbf{c}) = \sum_S p_{\mathbf{n}}(\mathbf{n})$, 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; Korwar 1988; Shlosser 1981). For this model Goodman (1949) showed that if it is known that $n \ge M$, then there exists a unique unbiased estimator of C, but without this knowledge no unbiased estimator exists. The estimator (when $n \ge M$) is

$$\hat{C}_{\text{GOODMAN1}} = c + \sum_{j=1}^{n} (-1)^{j+1} \frac{(N-n+j-1)!(n-j)!}{(N-n-1)!n!} c_j.$$

Unfortunately, although $\hat{C}_{\text{GOODMAN1}}$, being unique, is a fortiori uniformly minimum variance unbiased (UMVU), in many cases its variance is so large as to render it unusable ($\hat{C}_{\text{GOODMAN1}}$ need not even be positive). This was observed by Goodman himself and by later investigators (Engen 1978; Frank 1978; Knott 1967). Hou and Ozsoyoglu (1988, 1991) found that "when the sampling fraction is lower than some 'threshold,' [$\hat{C}_{\text{GOODMAN1}}$] always gives unreasonable estimates. On the other hand, once the sampling fraction is higher than the 'threshold,' [$\hat{C}_{\text{GOODMAN1}}$] converges very fast" (1991, p. 644). In contrast, Shlosser (1981) took an asymptotic approach (independently of Goodman) in which N, $n \to \infty$ in such a way that $n/N \to q \in (0, 1)$. On this basis, he derived

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

Note that $\hat{C}_{\text{SHLOSSER}} \ge c$. Shlosser 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 Shlosser's simulations were not extensive.

On the Bayesian side, Hill (1968, 1979, 1980) adopted the following prior for $(C, N_1, \ldots, N_C)$: $\Pi(C, N_1, \ldots, N_C) = \Pi_{N_1, \ldots, N_C | C, N}(N_1, \ldots, N_C | C, N) \Pi_{C,N}(C, N) = \binom{N-1}{C-1}^{-1} \Pi_{C,N}(C, N)$, 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

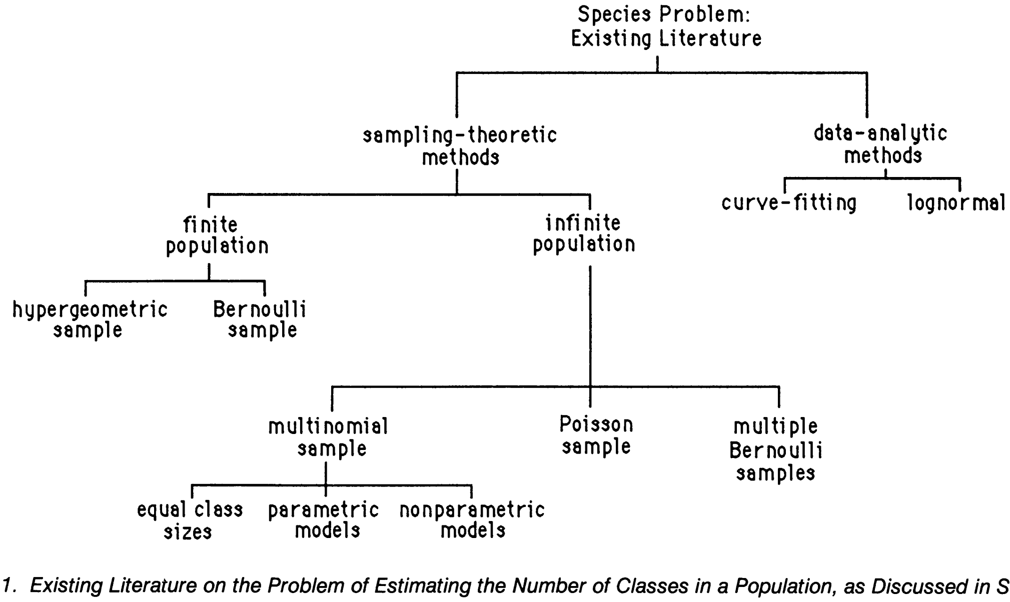

Figure 1. Existing Literature on the Problem of Estimating the Number of Classes in a Population, as Discussed in Section 1.

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 Schroeck 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 ith class independently contributes a binomial (N_i, p) random number of items to the sample, and $p_{\mathbf{n}}(\mathbf{n}) = p^n(1-p)^{N-n} \prod_{i=1}^{C} \binom{N_i}{n_i}$, $\sum_{i=1}^{C} n_i = n$, but $p_{\mathbf{c}}(\mathbf{c}) = \sum_S p_{\mathbf{n}}(\mathbf{n})$ again does not have a closed-form expression. Goodman (1949) considered such Bernoulli sampling for *known p* and derived the unbiased estimator

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

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 (θ_1, θ_2) and pmf $p_{N_i}(N_i) = \Gamma(\theta_1 + N_i) \times \theta_2^{\theta_1}(1-\theta_2)^{N_i}/(\Gamma(\theta_1)N_i!)$, denoted $\text{NB}(\theta_1, \theta_2)$. Then $n_1, \ldots, n_C$ are iid $\text{NB}(\theta_1, \theta_2/(\theta_2 + p - \theta_2 p))$ 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* θ_1 (p and θ_2 are implicitly estimated),

$$\hat{C}_{\text{NB}} = \frac{n}{\hat{\mu}},$$

where $\hat{\mu}$ is the root of the equation $n/c = \mu/(1 - (1 + \mu/\theta_1)^{-\theta_1})$. 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 θ_1) of certain estimators derived under $H_=$ (Sec. 1.3.1). He considered estimating θ_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 $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_C]'$, $\sum_{i=1}^{C} \pi_i = 1$. 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, $\mathbf{n}$ has a C-dimensional multinomial distribution, with pmf $p_{\mathbf{n}}(\mathbf{n}) = \binom{n}{n_1, \ldots, n_C} \prod_{i=1}^{C} \pi_i^{n_i}$, $\sum_{i=1}^{C} n_i = n$. But again $p_{\mathbf{c}}(\mathbf{c}) = \sum_S p_{\mathbf{n}}(\mathbf{n})$ 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 Timonina (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

$$\hat{C}_{\text{UMVUE}=} = \frac{S_{c,n+1}}{S_{c,n}},$$

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 π_i's corresponding to the observed classes (i.e., $u = \sum_{i=1}^{C} 1(n_i > 0)\pi_i$, where $1(A)$ is the indicator of the event A). There is a growing literature on coverage; see, for example, Betro and Zielinski (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 $\hat{u}_{\text{GOOD}} = (1 - c_1/n)$. 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 $\hat{C}_{\mathrm{COV}=}$ has high asymptotic efficiency relative to $\hat{C}_{\mathrm{MLE}=}$, 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 $\boldsymbol{\pi}$). 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 $\hat{C}_{\mathrm{MLE}=}$ or $\hat{C}_{\mathrm{COV}=}$ when $H_=$ holds.

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

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 π_i's have a *functional form,* possibly depending on a small number of parameters; that is, $\pi_i = f(i; \boldsymbol{\theta}, C)$, $i = 1, \ldots, C$, where $\boldsymbol{\theta}$ is a parameter vector and f is decreasing in i. Kalinin (1965) studied the "Zipf" model $f_Z(i; \theta, C) = \theta/i$ and the "Mandelbrot" model $f_M(i; \boldsymbol{\theta}, C) = \theta_1/(\theta_2 + i)^{\theta_3}$. He derived families of moment-method estimators (based on $\mathbf{c}$) of $(\boldsymbol{\theta}, C)$ 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 $C \to \infty$, $\pi_{[Cx]} \sim \alpha_C g(x)$, where g is some function and $\alpha_C \to 0$ as $C \to \infty$, so that π_i is approximately proportional to $g(i/C)$. In particular, for $g(x) = 1, x^{-1}$, or $x^{-\theta_3}$ we obtain approximations to $H_=$, f_Z, or f_M (the last with $\theta_2 = 0$). McNeil derived an estimator $\hat{C}_{\mathrm{MCNEIL}}$ (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 π_i's by a probability density function (pdf) that depends on some parameters $\boldsymbol{\theta}$. 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 $\psi(\pi; \theta_1, \theta_2) = \theta_1\sqrt{\theta_2}\exp\{\theta_1 - \pi/\theta_2 - \theta_1^2\theta_2/(4\pi)\}/(2\sqrt{Pi}\pi^{3/2})$. Sichel (1986b,c) showed that if the distribution of the π_i's is approximately given by $\psi(\pi; \theta_1, \theta_2)$ (for some (θ_1, θ_2)), then $C \approx (E_\psi(\pi))^{-1} \doteq 2(\theta_1\theta_2)^{-1}$. He then derived an estimate $(\widehat{\theta_1}, \widehat{\theta_2})$ of (θ_1, θ_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 ψ-mixed Poisson distribution, and hence obtained

$$\hat{C}_{\mathrm{SICHEL}} = \frac{2}{\widehat{\theta_1}\widehat{\theta_2}}.$$

Sichel also found the asymptotic bias and variance of $\hat{C}_{\mathrm{SICHEL}}$; 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 $\boldsymbol{\pi}$ 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 $\Pi_{C,N}(C, N) = \Pi_C(C)$ be a truncated negative binomial such that $\Pi_C(C) \propto \binom{C+\theta_1-1}{C}\theta_2^C$, $C = 1, \ldots, N$; $\theta_1 \in (0, \infty)$, $\theta_2 \in (0, 1]$. 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 $\Pi_{C,\boldsymbol{\pi}}(C, \boldsymbol{\pi}) = \mathscr{D}(\boldsymbol{\pi}; C, \theta_3)\Pi_C(C)$, where $\mathscr{D}(\boldsymbol{\pi}; C, \theta_3)$ is the C-dimensional symmetric Dirichlet density with parameter $\theta_3 \in (0, \infty)$ and $\Pi_C(C)$ is an arbitrary prior on $\mathbb{N}$. From this they derived the posterior $p_{C|\mathbf{c}}(C|\mathbf{c}) = p_{C|c}(C|c)$; in particular, if $\Pi_C(C)$ is the negative binomial as in Hill's model, then

$$p_{C|\mathbf{c}}(C|\mathbf{c}) = p_{C|c}(C|c)$$

$$\propto \theta_2^C \binom{C+\theta_1-1}{C}\binom{C}{c}\bigg/\binom{\theta_3 C+n-1}{n}.$$

Lewins and Joanes took the mode of $p_{C|c}(C|c)$ to be their estimate of C. They found that "the model appears to be very robust for [θ_1] and [θ_2]," and "variations in [θ_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 $\boldsymbol{\pi} \sim \mathscr{D}(\boldsymbol{\pi}; C, \theta_3)$, where θ_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 θ_3 and C, finding that under some conditions the estimates are sensitive to the choice of θ_3 (when it is given) and that estimation of θ_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 π? No unbiased estimate exists; Engen (1978, p. 28) showed that the bias of any estimator of C based on $\mathbf{c}$ 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,

$$\hat{C}_{\text{CHAO1}} = c + \frac{c_1^2}{2c_2},$$

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

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

which is $\hat{C}_{\text{COV=}}$ plus a bias correction term that depends on an estimate $\hat{\gamma}$ of the coefficient of variation γ of the π_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 ith class in the sample is a Poisson random variable with mean λ_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 $\lambda = \sum_{j=1}^{C} \lambda_j$ and $p_{\mathbf{n}}(\mathbf{n}) = e^{-\lambda} \prod_{i=1}^{C} \lambda_i^{n_i}/n_i!$, but again $p_{\mathbf{c}}(\mathbf{c}) = \sum_S p_{\mathbf{n}}(\mathbf{n})$ does not have a closed-form expression. If we assume in addition that the λ_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 $\widehat{p_0(F)}$ of $p_0(F)$, an estimator of C is

$$\hat{C}_{\text{POISSON}} = \frac{c}{1 - \widehat{p_0(F)}}.$$

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 $\widehat{p_0(F)} = p_0(\hat{\theta}_1, \hat{\theta}_2)$, where $(\hat{\theta}_1, \hat{\theta}_2)$ are the MLE's of (θ_1, θ_2) based on $\mathbf{c}$. The authors analyzed several data sets and concluded that "the fit of the Poisson-inverse Gaussian distribution . . . suggests that it may be an appropriate 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 $\mathbf{c}$) that are robust with respect to F's whose variance is not too large. The simplest is $\widehat{p_0(F)} = \hat{p}_0 = \exp(-2c_2/c_1)$, 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 1987.) Noting that $Cp_0(F)$ 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 $Cp_0(F)$ subject to constraints derived from $\mathbf{c}$. This yields an *estimated lower bound* for C,

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

where c_{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 $[x_{i\nu}]$, where $x_{i\nu} = 1$ (ith class is observed on νth occasion), $i = 1, \ldots, C$, $\nu = 1, \ldots, n$. Only rows with at least one 1 are observable; in fact, c_j = the number of rows with exactly j 1's. Burnham and Overton (1978, 1979) studied such a model in which the $x_{i\nu}$'s are all independent, with $P\{x_{i\nu} = 1\} \equiv \pi_i$ and $\nu = 1, \ldots, n$ (the probabilities are the same on each occasion). Then the contribution of each class to the sample, $n_i = \sum_{\nu=1}^{n} x_{i\nu}$, is a binomial (n, π_i) random variable, $p_{\mathbf{n}}(\mathbf{n}) = \prod_{i=1}^{C} \binom{n}{n_i} \pi_i^{n_i} (1 - \pi_i)^{n-n_i}$, and $p_{\mathbf{c}}(\mathbf{c}) = \sum_S p_{\mathbf{n}}(\mathbf{n})$ again does not have a closed-form expression. Taking the π_i's to be a random sample from some distribution F, they developed a kth order jackknife estimator $\hat{C}_{\text{BO}k}$ of C; for example

$$\hat{C}_{\text{BO1}} = c + \left(\frac{n-1}{n}\right) c_1.$$

Burnham and Overton (1979) computed the mean and variance of $\hat{C}_{\text{BO}k}$ (these depend on F), and described a procedure for choosing optimal k. For the same model Chao (1987) proposed using $\hat{C}_{\text{CHAO1}}$, and Yip (1991c), taking F to be a beta distribution, used Martingale theory to derive an estimating equation for C and associated variance estimators (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, $\hat{C}_{\mathrm{BOk}}$ has the form of an alternating series and can be unstable. Esty (1986b, p. 209) stated that $\hat{C}_{\mathrm{BOk}}$ "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 $\hat{C}_{\mathrm{CHAO1}}$ is preferable to $\hat{C}_{\mathrm{BOk}}$ in some situations but not in others. Yip compared his procedure with both $\hat{C}_{\mathrm{BOk}}$ and $\hat{C}_{\mathrm{CHAO1}}$ 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 given observed values of $\mathbf{c}$ 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) $E(c^{(x)}) = f(x; \boldsymbol{\theta})$, where x is a measure of the "size" of the sample (not necessarily the count of items), $\boldsymbol{\theta}$ is a parameter vector, and f is a given increasing *bounded* function of x; and (b) $\lim_{x\to\infty} f(x; \boldsymbol{\theta}) = C(<\infty)$. If an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained from the data $\mathbf{c}^{(x_1)}, \mathbf{c}^{(x_2)}, \cdots$ (where $\mathbf{c}^{(x_i)}$ is an observed value of $\mathbf{c}$ with sample size x_i), then our estimate of C will be $\lim_{x\to\infty} f(x, \hat{\boldsymbol{\theta}})$. Brainerd (1972) proposed several such f's for estimation of literary vocabulary, the simplest of which is

$$E(c^{(x)}) = f_{\mathrm{BR}}(x; \theta) = \frac{1 - \theta^x}{1 - \theta},$$

$\theta \in (0, 1)$, where x = the number of words in a text. Then $\lim_{x\to\infty} f_{\mathrm{BR}}(x; \theta) = 1/(1 - \theta)$. Brainerd considered "admittedly ad hoc methods of estimation" of θ (Brainerd 1972, p. 517), obtaining $\hat{C}_{\mathrm{BR}} = 1/(1 - \hat{\theta})$ (see also Brainerd 1981, 1982; Najock 1986). Alternatively, De Caprariis, Lindemann, and Collins (1976) proposed the "hyperbolic" model

$$E(c^{(x)}) = f_{\mathrm{DCLH}}(x; \boldsymbol{\theta}) = \frac{\theta_1 x}{1 + \theta_2 x},$$

$\theta_1, \theta_2 > 0$, where x in their application was the mass of a sample of beach sand. Here $\lim_{x\to\infty} f_{\mathrm{DCLH}}(x; \boldsymbol{\theta}) = \theta_1/\theta_2$; applying linear regression to a transformed version of the equation to estimate (θ_1, θ_2), the authors obtained $\hat{C}_{\mathrm{DCLH}} = \widehat{\theta_1}/\widehat{\theta_2}$. (See also de Caprariis 1984; de Caprariis and Lindemann 1978, 1981; de Caprariis, Lindemann, and Haimes 1981; Kalantar 1987; and Tuldava 1977, 1987.) Although de Caprariis and his coauthors were generally satisfied with their curve-fitting and estimates in their biological applications, Brainerd felt that his results in vocabulary estimation were at best "fairly serviceable" (Brainerd 1972, p. 517). In our opinion, it is hard to be very optimistic about the potential of such methods, because if the function $f(x; \boldsymbol{\theta})$ is derived from a sampling model, then sampling theory will give a more efficient estimate of C, and if it is not, then its form seems difficult to justify.

1.6.2 Lognormal Fit. Preston (1948) found that the graph of $(\log_2 j, c_j)$ often resembles a Gaussian curve; this curve is truncated on the left, because the leftmost available point is $(0, c_1)$. An estimate of C can be obtained by fitting a Gaussian curve to the observed graph, extrapolating the curve to the left, and integrating it over $(-\infty, +\infty)$. This method was applied to ecological datasets by Slocomb, Stauffer, and Dickson (1977), who found it unsatisfactory. Palmer (1990) studied the method in a botanical experiment with *known* C and found the resulting estimator to be little better than c itself. (See also Nee, Harvey, and May 1991.)

2. DISCUSSION AND RECOMMENDATIONS

2.1 The State of the Art

The researcher faced with the need to estimate the number of classes in a population may justifiably feel somewhat bewildered at this point. There are many models and procedures, each claiming some justification and optimality properties, but little comparative information available about them. We are not particularly enthusiastic about cutting this Gordian knot. Given extensive knowledge about a particular situation (i.e., an *accurate* model), one can attempt to "fine-tune" a suitable procedure, possibly incorporating information such as an upper bound for C, a postulated distribution of the class proportions, and so on,—although the best procedure for a given model need not have been derived from that model, and the best available procedure still may not give usable results. In any case, it is rare that a sampling model obtains exactly; ideally, one would like to have an estimator of C, based solely on $\mathbf{c}$, that is robust across various sampling plans and population structures. Two authors have addressed such issues directly, in the numismatic context.

Stam (1987) performed a detailed analysis of some datasets involving Roman coins. He tested the fit of the data to various models, including multinomial sampling under $H_=$ and the "invariant" Bernoulli/negative binomial model of Section 1.2, and decided in favor of the latter (at least for his dataset). He proposed new estimators of the negative binomial parameters, 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 θ_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 $H_=$ 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 θ_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 $\hat{C}_{\text{CHAO2}}$, 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 θ must be estimated jointly with C, and authors have typically been dissatisfied with the results. But a substantial literature deals with estimation of θ 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 superpopulation model like that of Section 1.2, in which $N_1, \ldots, N_C$ are iid random variables with distribution F_θ. If, under Bernoulli sampling, $n_1, \ldots, n_C$ (the observed numbers of representatives of each class) are iid random variables with distribution $F_{\theta'}$, (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 MLE's of θ 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 MLE'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 $H_=$ holds or not), and coverage-based estimators of C under $H_=$ 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 $\hat{u}_{\text{GOOD}}$ 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 $H_=$.

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 ith 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 $\lambda_1 t, \ldots, \lambda_C t$. 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 also could be extended to accommodate randomly generated λ_i's. Such an investigation could explain, for example, the relationship between $\hat{C}_{\text{POISSON}}$ and coverage-based estimators like $\hat{C}_{\mathrm{COV}\simeq}$. Recently, M. T. Chao (1992) derived approximate moments of the number of classes observed in such a $\mathbb{P}$ during $[0, t)$, given certain conditions on the λ_i's, and Nayak and Christman (in press) gave majorization results to quantify estimation bias in this model as the λ_i'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

Arnold, B. C., and Beaver, R. J. (1988), "Estimation of the Number of Classes in a Population," *Biometrical Journal,* 30, 413–424.

Atkinson, A. C., and Yeh, L. (1982), "Inference for Sichel's Compound Poisson Distribution," *Journal of the American Statistical Association,* 77, 153–158.

Betro, B., and Zielinski, R. (1987), "A Monte Carlo Study of a Bayesian Decision Rule Concerning the Number of Different Values of a Discrete Random Variable," *Communications in Statistics, Part B—Simulation and Computation,* 16, 925–938.

Bickel, P. J., and Yahav, J. A. (1985), "On Estimating the Number of Unseen Species: How Many Executions Were There?," Technical Report No. 43, University of California, Berkeley, Dept. of Statistics.

——— (1986), "On Estimating the Total Probability of the Unobserved Outcomes of an Experiment," in *Adaptive Statistical Procedures and Related Topics: Proceedings of a Symposium in Honor of Herbert Robbins* (Vol. 8), ed. J. Van Ryzin, Hayward, CA: Institute of Mathematical Statistics, pp. 332–337.

——— (1988), "On Estimating the Number of Unseen Species and System Reliability," in *Statistical Decision Theory and Related Topics IV* (Vol. 2), eds. S. S. Gupta and J. O. Berger, New York: Springer-Verlag, pp. 265–271.

Boender, C. G. E., and Rinnooy Kan, A. H. G. (1987), "A Multinomial Bayesian Approach to the Estimation of Population and Vocabulary Size," *Biometrika,* 74, 849–856.

Brainerd, B. (1972), "On the Relation Between Types and Tokens in Literary Text," *Journal of Applied Probability,* 9, 507–518.

——— (1981), "Some Elaborations Upon Gani's Model for The Type-Token Relationship," *Journal of Applied Probability,* 18, 452–460.

——— (1982), "On the Relation Between the Type-Token and Species-Area Problems," *Journal of Applied Probability,* 19, 785–793.

Brown, I. D. (1955), "Some Notes on the Coinage of Elizabeth I With Special Reference to Her Hammered Silver," *British Journal of Numismatics,* 28, 568–603.

Burnham, K. P., and Overton, W. S. (1978), "Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals," *Biometrika,* 65, 625–633.

——— (1979), "Robust Estimation of Population Size When Capture Probabilities Vary Among Animals," *Ecology,* 60, 927–936.

Burrell, Q. (1989), "On the Growth of Bibliographies With Time: An Exercise in Bibliometric Prediction," *Journal of Documentation,* 45, 302–317.

Burrell, Q., and Fenton, M. (1993), "Yes, The GIGP Really Does Work—And Is Workable!", *Journal of the American Society for Information Science,* 44.

Castledine, B. J. (1981), "A Bayesian Analysis of Multiple-Recapture Sampling for a Closed Population," *Biometrika,* 67, 197–210.

Chao, A. (1981), "On Estimating the Probability of Discovering a New Species," *The Annals of Statistics,* 9, 1339–1342.

——— (1984), "Nonparametric Estimation of the Number of Classes in a Population," *Scandinavian Journal of Statistics, Theory and Applications,* 11, 265–270.

——— (1987), "Estimating the Population Size for Capture-Recapture Data With Unequal Catchability," *Biometrics,* 43, 783–791.

Chao, A., and Lee, S.-M. (1990), "Estimating the Number of Unseen Species With Frequency Counts," *Chinese Journal of Mathematics,* 18, 335–351.

——— (1992), "Estimating the Number of Classes Via Sample Coverage," *Journal of the American Statistical Association,* 87, 210–217.

Chao, M.-T. (1992), "From Animal Trapping to Type-Token," *Statistica Sinica,* 2, 189–201.

Chapman, D. G. (1951), "Some Properties of the Hypergeometric Distribution With Applications to Zoological Sample Censuses," in *University of California Publications in Statistics* (Vol. 1), eds. M. Loeve, G. M. Kuznets, E. L. Lehmann, and J. Neyman, Berkeley and Los Angeles: University of California Press, pp. 131–160.

Charalambides, C. A., and Singh, J. (1988), "A Review of the Stirling Numbers, Their Generalizations and Statistical Applications," *Communications in Statistics, Part A—Theory and Methods,* 17, 2533–2595.

Chen, W.-C. (1980), "On the Weak Form of Zipf's Law," *Journal of Applied Probability,* 17, 611–622.

——— (1981a), "Limit Theorems for General Size Distributions," *Journal of Applied Probability,* 18, 139–147.

——— (1981b), "Some Local Limit Theorems in the Symmetric Dirichlet-Multinomial Urn Models," *Annals of the Institute of Statistical Mathematics,* 33, 405–415.

Clayton, M. K., and Frees, E. W. (1987), "Nonparametric Estimation of the Probability of Discovering a New Species," *Journal of the American Statistical Association,* 82, 305–311.

Cohen, A., and Sackrowitz, H. B. (1990), "Admissibility of Estimators of the Probability of Unobserved Outcomes," *Annals of the Institute of Statistical Mathematics,* 42, 623–636.

Cox, D. R., and Isham, V. (1980), *Point Processes,* London: Chapman and Hall.

Darroch, J. N. (1958), "The Multiple-Recapture Census, I. Estimation of a Closed Population," *Biometrika,* 45, 343–359.

Darroch, J. N., and Ratcliff, D. (1980), "A Note on Capture-Recapture Estimation," *Biometrics,* 36, 149–153.

de Caprariis, P. (1984), "Estimating Species Diversity: Comparison of Two Algorithms," *Mathematical Geology,* 16, 237–248.

de Caprariis, P., and Lindemann, R. H. (1978), "Species Richness in Patchy Environments," *Mathematical Geology,* 10, 73–90.

——— (1981), "Maximum Diversities From Cumulative Species Curve," *Lethaia,* 14, 134.

de Caprariis, P., Lindemann, R. H., and Collins, C. M. (1976), "A Method for Determining Optimum Sample Size in Species Diversity Studies," *Mathematical Geology,* 8, 575–581.

de Caprariis, P., Lindemann, R., and Haimes, R. (1981), "A Relationship Between Sample Size and Accuracy of Species Richness Predictions," *Mathematical Geology,* 13, 351–355.

Doss, H., and Sethuraman, J. (1989), "The Price of Bias Reduction When There is No Unbiased Estimate," *The Annals of Statistics,* 17, 440–442.

Driml, M., and Ullrich, M. (1967), "Maximum Likelihood Estimate of the Number of Types," *Acta Technica Csav,* 3, 300–303.

Efron, B., and Thisted, R. (1975), "Estimating the Number of Unseen Species (How Many Words Did Shakespeare Know?)," Technical Report No. 9, Stanford University, Division of Biostatistics.

——— (1976), "Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?," *Biometrika,* 63, 435–447.

Emigh, T. H. (1983), "On the Number of Observed Classes From a Multinomial Distribution," *Biometrics,* 39, 485–491.

Engen, S. (1975), "The Coverage of a Random Sample From a Biological Community," *Biometrics,* 31, 201–208.

——— (1978), *Stochastic Abundance Models,* London: Chapman and Hall.

Esty, W. W. (1982), "Confidence Intervals for the Coverage of Low Coverage Samples," *The Annals of Statistics,* 10, 190–196.

——— (1983), "A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample," *The Annals of Statistics,* 11, 905–912.

——— (1984), "Confidence Intervals for an Occupancy Problem Estimator Used by Numismatists," *Mathematical Scientist,* 9, 111–115.

——— (1985), "Estimation of the Number of Classes in a Population and the Coverage of a Sample," *Mathematical Scientist,* 10, 41–50.

——— (1986a), "The Efficiency of Good's Nonparametric Coverage Estimator," *The Annals of Statistics,* 14, 1257–1260.

——— (1986b), "Estimation of the Size of a Coinage: A Survey and Comparison of Methods," *Numismatic Chronicle,* 146, 185–215.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943), "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population," *Journal of Animal Ecology,* 12, 42–58.

Frank, O. (1978), "Estimation of the Number of Connected Components in a Graph by Using a Sampled Subgraph," *Scandinavian Journal of Statistics, Theory and Applications,* 5, 177–188.

Good, I. J. (1950), *Probability and the Weighing of Evidence,* London: Charles Griffin.

——— (1953), "The Population Frequencies of Species and the Estimation of Population Parameters," *Biometrika,* 40, 237–264.

Good, I. J., and Toulmin, G. H. (1956), "The Number of New Species, and the Increase in Population Coverage, When a Sample is Increased," *Biometrika,* 43, 45–63.

Goodman, L. A. (1949), "On the Estimation of the Number of Classes in a Population," *Annals of Mathematical Statistics,* 20, 572–579.

Hall, P. (1991), "Bahadur Representation for Uniform Resampling and Importance Resampling with Applications to Asymptotic Relative Efficiency," *The Annals of Statistics,* 19, 1062–1072.

Harris, B. (1959), "Determining Bounds on Integrals with Applications to Cataloging Problems," *Annals of Mathematical Statistics,* 30, 521–548.

——— (1968), "Statistical Inference in the Classical Occupancy Problem Unbiased Estimation of the Number of Classes," *Journal of the American Statistical Association,* 63, 837–847.

Harwit, M. (1981), *Cosmic Discovery: The Search, Scope, and Heritage of Astronomy,* New York: Basic Books.

Harwit, M., and Hildebrand, R. (1986), "How Many More Discoveries in the Universe?," *Nature,* 320, 724–726.

Hill, B. M. (1968), "Posterior Distribution of Percentiles: Bayes's Theorem for Sampling From a Population," *Journal of the American Statistical Association,* 63, 677–691.

——— (1979), "Posterior Moments of the Number of Species in a Finite Population and the Posterior Probability of Finding A New Species," *Journal of the American Statistical Association,* 74, 668–673.

——— (1980), "Invariance and Robustness of the Posterior Distribution of Characteristics of a Finite Population, With Reference to Contingency Tables and the Sampling of Species," in *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (Vol. 1), ed. A. Zellner, Amsterdam: North-Holland, pp. 383–395.

Holst, L. (1981), "Some Asymptotic Results for Incomplete Multinomial or Poisson Samples," *Scandinavian Journal of Statistics,* 8, 243–246.

——— (1986), "On Birthday, Collectors', Occupancy and Other Classical Urn Problems," *International Statistical Review,* 54, 15–27.

Hou, W., and Ozsoyoglu, G. (1991), "Statistical Estimators for Aggregate Relational Algebra Queries," *ACM Transactions on Database Systems,* 16, 600–654.

Hou, W., Ozsoyoglu, G., and Taneja, B. K. (1988), "Statistical Estimators for Relational Algebra Expressions," in *Proceedings of the ACM Symposium on Principles of Database Systems,* pp. 276–287.

Ivchenko, G. I., and Timonina, E. E. (1983), "Estimating the Size of a Finite Population," *Theory of Probability and Its Applications,* 27, 403–406.

Johnson, N. L., and Kotz, S. (1977), *Urn Models and Their Application,* New York: John Wiley.

Kalantar, A. H. (1987), "Using the Hyperbolic Model to Estimate Species Richness," *Mathematical Geology,* 19, 151–154.

Kalinin, V. M. (1965), "Functionals Related to the Poisson Distribution, and the Statistical Structure of a Text," *Proceedings of the Steklov Institute of Mathematics,* 79, 6–19.

Keener, R., Rothman, E., and Starr, N. (1987), "Distributions on Partitions," *The Annals of Statistics,* 15, 1466–1481.

Knott, M. (1967), "Models for Cataloguing Problems," *Annals of Mathematical Statistics,* 38, 1255–1260.

Korwar, R. M. (1988), "On the Observed Number of Classes From Multivariate Power Series and Hypergeometric Distributions," *Sankhya: The Indian Journal of Statistics,* 50, 39–59.

Kovar, J. G., Rao, J. N. K., and Wu, C. F. J. (1988), "Bootstrap and Other Methods to Measure Errors in Survey Estimates," *Canadian Journal of Statistics,* 16, 25–45.

Lee, J. (1989), "On Asymptotics for the NPMLE of the Probability of Discovering a New Species and an Adaptive Stopping Rule in Two-Stage Searches," unpublished Ph.D. dissertation, University of Wisconsin-Madison, Dept. of Statistics.

Lewins, W. A., and Joanes, D. N. (1984), "Bayesian Estimation of the Number of Species," *Biometrics,* 40, 323–328.

Lewontin, R. C., and Prout, T. (1956), "Estimation of the Number of Different Classes in a Population," *Biometrics,* 12, 211–223.

Liu, R. C., and Brown, L. D. (in press), "Non-Existence of Good Unbiased Estimators in Singular Problems," *The Annals of Statistics.*

Lloyd, C. J., and Yip, P. (1991), "A Unification of Inference From Capture-Recapture Studies Through Martingale Estimating Functions," in *Estimating Functions,* ed. V. P. Godambe, Oxford, UK: Clarendon Press, pp. 65–88.

Lo, H., and Wani, J. K. (1983), "Maximum Likelihood Estimation of the Parameters of the Invariant Abundance Distributions," *Biometrics,* 39, 977–986.

Lo, S. (1992), "From Species Problem to a General Coverage Problem Via a New Interpretation," *The Annals of Statistics,* 20, 1094–1109.

MacDonald, J. D. (1977), *Condominium,* Philadelphia: J. B. Lippincott.

Mann, Charles C. (1991), "Extinction: Are Ecologists Crying Wolf?," *Science,* 253, 709–824.

Marchand, J. P., and Schroeck, F. E. (1982), "On the Estimation of the Number of Equally Likely Classes in a Population," *Communications in Statistics, Part A—Theory and Methods,* 11, 1139–1146.

McNeil, D. (1973), "Estimating an Author's Vocabulary," *Journal of the American Statistical Association,* 68, 92–96.

Najock, D. (1986), "Bootstrap Experiments for the Evaluation of Expected Values and Variances of Vocabulary Sizes," in *Methodes Quantitatives et Informatiques dans L'etude des Textes,* Geneva: Slatkine-Champion, pp. 658–670.

Nayak, T. K. (1989), "A Note on Estimating the Number of Errors in a System by Recapture Sampling," *Statistics & Probability Letters,* 7, 191–194.

Nayak, T. K., and Christman, M. C. (in press), "Effect of Unequal Catchability on Estimates of the Number of Classes in a Population," *Scandinavian Journal of Statistics.*

Nee, S., Harvey, P. H., and May, R. M. (1991), "Lifting the Veil on Abundance Patterns," *Proceedings of the Royal Society of London,* Ser. B, 243, 161–163.

Ord, J. K., and Whitmore, G. A. (1986), "The Poisson-Inverse Gaussian Distribution as a Model for Species Abundance," *Communications in Statistics, Part A—Theory and Methods,* 15, 853–871.

Palmer, M. W. (1990), "The Estimation of Species Richness by Extrapolation," *Ecology,* 71, 1195–1198.

Patil, G. P., and Taillie, C. (1982), "Diversity as a Concept and its Measurement" (with comments), *Journal of the American Statistical Association,* 77, 548–567.

Pollock, K. H. (1991), "Modeling Capture, Recapture, and Removal Statistics for Estimation of Demographic Parameters for Fish and Wildlife Populations: Past, Present, and Future," *Journal of the American Statistical Association,* 86, 225–238.

Preston, F. W. (1948), "The Commonness, and Rarity, of Species," *Ecology,* 29, 254–83.

Rao, J. N. K., and Wu, C. F. J. (1988), "Resampling Inference With Complex Survey Data," *Journal of the American Statistical Association,* 83, 231–241.

Robbins, H. E. (1968), "Estimating the Total Probability of the Unobserved Outcomes of an Experiment," *Annals of Mathematical Statistics,* 39, 256–257.

Schroeck, F. E. (1981), "Tabulated Results of the Estimation of the Number of Dies of a Coin and the Analysis of a Hoard of Copper Falus of Taimur Shah," *Numismatic Circular,* 89, 37–38.

Shlosser, A. (1981), "On Estimation of the Size of the Dictionary of a Long Text on the Basis of a Sample," *Engineering Cybernetics,* 19, 97–102.

Sichel, H. S. (1975), "On a Distribution Law for Word Frequencies," *Journal of the American Statistical Association,* 70, 542–547.

——— (1982), "Asymptotic Efficiencies of Three Methods of Estimation for the Inverse Gaussian-Poisson Distribution," *Biometrika,* 69, 467–472.

——— (1986a), "The GIGP Distribution Model with Applications to Physics Literature," *Czechoslovak Journal of Physics,* Ser. B, 36, 133–137.

——— (1986b), "Parameter Estimation for a Word Frequency Distribution

Based on Occupancy Theory," *Communications in Statistics, Part A—Theory and Methods,* 15, 935–949.

——— (1986c), "Word Frequency Distributions and Type-Token Characteristics," *Mathematical Scientist,* 11, 45–72.

——— (1992a), "Anatomy of the Generalized Inverse Gaussian-Poisson Distribution with Special Applications to Bibliometric Studies," *Information Processing and Management,* 28, 5–17.

——— (1992b), "Note on a Strongly Unimodal Bibliometric Size Frequency Distribution," *Journal of the American Society for Information Science,* 43, 299–303.

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

Slocomb, J., Stauffer, B., and Dickson, K. L. (1977), "On Fitting the Truncated Lognormal Distribution to Species-Abundance Data Using Maximum Likelihood Estimation," *Ecology,* 58, 693–696.

Stam, A. J. (1987), "Statistical Problem in Ancient Numismatics," *Statistica Neerlandica,* 41, 151–173.

Starr, N. (1979), "Linear Estimation of the Probability of Discovering a New Species," *The Annals of Statistics,* 7, 644–652.

Stein, G. Z., Zucchini, W., and Juritz, J. M. (1987), "Parameter Estimation for the Sichel Distribution and Its Multivariate Extension," *Journal of the American Statistical Association,* 82, 938–944.

Thisted, R., and Efron, B. (1987), "Did Shakespeare Write a Newly-Discovered Poem?," *Biometrika,* 74, 445–455.

Tiwari, R. C., and Tripathi, R. C. (1989), "Nonparametric Bayes Estimation of the Probability of Discovering a New Species," *Communications in Statistics, Part A—Theory and Methods,* 18, 877–895.

Trybula, S. (1958), "The Estimation of Frequency in a Population of Elements Belonging to Classes not Represented in the Sample," *Zastosowania Matematyki* (in Polish), 4, 244–248.

Tuldava, J. (1977), "Quantitative Relations Between the Size of Text and the Size of Vocabulary," *SMIL Quarterly,* 28–35.

——— (1987), *Problems and Methods of the Quantitative-Systemic Investigation of Vocabulary* (in Russian), Tallin: Valgus.

Wani, J. K., and Lo, H. P. (1983a), "A Characterization of Invariant Power-Series Abundance Distributions," *Canadian Journal of Statistics,* 11, 317–323.

——— (1983b), "The Structure of Invariant Abundance Distributions," *Journal of the Indian Statistical Association,* 21, 1–7.

——— (1986), "Selecting a Power-Series Distribution for Goodness of Fit," Canadian Journal of Statistics, 14, 347–353.

Wani, J. K., and Watterson, G. A. (1982), "Sampling Theory for Species Abundances in Certain Biological Populations," *Canadian Journal of Statistics,* 10, 207–211.

Willmot, G. E. (1987), "The Poisson-Inverse Gaussian Distribution as an Alternative to the Negative Binomial," *Scandinavian Actuarial Journal,* 113–127.

——— (1988a), "Parameter Orthogonality for a Family of Discrete Distributions," *Journal of the American Statistical Association,* 83, 517–521.

——— (1988b), "Sundt and Jewell's Family of Discrete Distributions," *Astin Bulletin,* 18, 19–29.

Wu, C. F. J. (1991), "Balanced Repeated Replications Based on Mixed Orthogonal Arrays," *Biometrika,* 78, 181–188.

Yatracos, Y. G. (1991), "On the Species and Related Problems," *Statistics & Probability Letters,* 12, 209–212.

Yip, P. (1989), "An Inference Procedure for a Capture and Recapture Experiment With Time-Dependent Capture Probabilities," *Biometrics,* 45, 471–479.

——— (1991a), "Estimating Population Size From a Capture-Recapture Experiment With Known Removals," *Theoretical Population Biology,* 40, 1–13.

——— (1991b), "A Martingale Estimating Equation for a Capture-Recapture Experiment in Discrete Time," *Biometrics,* 47, 1081–1088.

——— (1991c), "A Method of Inference for a Capture-Recapture Experiment in Discrete Time with Variable Capture Probabilities," *Communications in Statistics—Stochastic Models,* 7, 343–362.

Zelterman, D. (1988), "Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments," *Journal of Statistical Planning and Inference,* 18, 225–237.

Zielinski, R. (1981), "A Statistical Estimate of the Structure of Multi-Extremal Problems," *Mathematical Programming,* 21, 348–356.