A HYPOTHESIS TEST OF THE COVERAGE OF A SAMPLE FROM AN INFINITE POPULATION

Carlos M. Hernandez and Olga I. Codero-Brana

BU-1325-M March 1996

A HYPOTHESIS TEST OF THE COVERAGE OF A SAMPLE FROM AN INFINITE POPULATION

BY CARLOS M. HERNANDEZ AND OLGA I. CORDERO-BRAÑA *

ABSTRACT: We consider sampling from a population composed of K classes or species in proportions θ_i where K and $\theta_1, \ldots, \theta_K$ are unknown and the population size is infinite. By allowing K not to be restricted to the number of species observed in the sample we obtain unrestricted maximum likelihood estimates of the θ_i's. Using these estimates a test of hypothesis on the coverage of a sample is given. The method does not require any of the usual assumptions, for example equal class sizes, but that all individuals have the same probability of being detected at any moment. Results from a Monte Carlo study suggest the practicality of the method in determining whether a desired coverage had been achieved.

KEY WORDS: coverage of a sample; unrestricted maximum likelihood estimator; singletons; number of classes.

*Carlos M. Hernandez is a graduate student at Cornell University, Biometrics Unit Ithaca, NY 14853-7801 and Olga I. Cordero-Braña is Assistant Professor, Department of Mathematics and Statistics, American University, Washington, DC 20016-8050. This research has been funded partially by a Conacyt Scholarship to Carlos Hernandez and by grant DEB-9253570 (Presidential Faculty Award to Carlos Castillo-Chavez).

1. INTRODUCTION

Consider sampling from a population composed of K classes or "species" in proportions θ_i where K and $\theta_1, \ldots, \theta_K$ are unknown. Two related problems associated with this situation are the estimation of K and the coverage, u, of a sample on the basis of a sample of size n. In the latter, the aim is to make inferences on the proportion that accounts for the species found in the sample. That is, $u = \sum_{i=1}^{K} I_{\{n_i>0\}}\theta_i$ where n_i is the number of individuals from the ith class in the sample, $i = 1, \ldots, K$, and I_A is the indicator function of the event A. However, these problems arise from different objectives. For example, the estimation of the number of species in a biological population provides information on the diversity of the whole population. The coverage, on the other hand, conveys information about the reach of the sample; this is very relevant for survey opinions, where it is desired to make inferences on the proportion of people whose opinions have been recorded. Note that the random vector $\mathbf{n} = (n_1, \ldots, n_K)^T$ is not observable; instead, we observe the vector $\mathbf{c} = (c_1, \ldots, c_n)^T$ where c_j denotes the number of classes represented j times in the sample. For example, c_1 is the number of singletons, c_2 is the number of twin pairs, and so on. Let $c = \sum_{j=1}^{n} c_j$. Under the assumption of equal class sizes, $u = c/K$ so that given a predictor $\hat{u}$ of u, K is estimated by $\hat{K} = c/\hat{u}$.

Bunge and Fitzpatrick (1993) provided a comprehensive review of the literature related to the estimation of the number of species and the coverage. Under the assumption of equal class sizes, Good (1953) and Good and Toulmin (1953) proposed the first estimator of u using the proportion of species in the sample represented by singletons. That is, $\hat{u} = 1 - (c_1/n)$. Robbins (1968) used the proportion of singletons in the sample as an estimator of the proportion occupied by the uncovered species but only after taking an additional sample. The resulting estimator is unbiased. Engen (1975) used a negative binomial model and using a normal approximation to the ratio of two densities derived a confidence interval for u. Starr(1979) considered generalizing Robbins (1968) procedure to an extended search of m individuals and derived a linear estimator for the expected value of the probability of finding a new species at the end of the $n+m$ stage. Esty (1982) proposed an estimator of u by using a Poisson approximation to the number of classes represented at least twice in the population.

Here, we propose a simple and practical solution for applied researchers to decide if the desired coverage is reached. The only requirement for this procedure is that at any time all individuals have the same probability of being detected, namely 1/N where N is the unknown population size. Results from a moderate Monte Carlo study are presented.

2. THEORETICAL RESULTS

When assuming sampling from a multinomial distribution with vector of proportion parameters Θ one can get two types of maximum likelihood estimators. By restricting the dimension of Θ to equate the number of classes observed in the sample we obtain the classical maximum likelihood estimator, MLE. By allowing K to vary freely, we get what we call the unrestricted maximum likelihood estimator, UMLE, of Θ. Observe that the likelihood of the sample under the UMLE is always greater than or equal to the likelihood using the classical MLE since the optimization of the former is done over a larger set.

The following result considers the case where the sample is comprised of singletons only. Although a simple result it is essential to the proof of Proposition 2.2. Intuitively, if all individuals in the population are different then the probability that a sample of size n contains singletons only is maximized.

Proposition 2.1 If a sample of size n contains singletons only then the UMLE of Θ has dimension N, the population size, with $\theta_i = 1/N$ for $i = 1, \ldots, N$. In this case the coverage is minimized over all possible values.

Proof. In order to have a sample of size n composed of singletons only it is necessary that every sampled individual be different from all the previous ones. Let ϑ_1 be the population proportion of the first detected species, ϑ_2

that of the second one, and so on. Note that ϑ_j, can be any of the θ_i, $i = 1, \ldots, \mathrm{K}$. The probability of getting n singletons is given by

$$\Pr(n \text{ singletons}) = (1 - \vartheta_1)(1 - \vartheta_1 - \vartheta_2) \cdots (1 - \vartheta_1 - \vartheta_2 - \cdots - \vartheta_{n-1}).$$

This probability is maximized by letting ϑ_j be the smallest possible; that is, by choosing $\vartheta_j = 1/N$ for $j = 1, \ldots, \mathrm{N}$. But this requires K=N, which implies that all individuals of the population must be different. To show that for this particular UMLE the coverage is minimized, note that after taking a sample of size n, with singletons only, the minimum value for the proportion of covered species is n/N and for this population the coverage is effectively n/N.

The next result deals with the case of a mixed sample of singletons and non-singletons species. To find the UMLE in this case it is necessary to have all non-singleton species present in the sample and the remaining species, observed singletons and unobserved ones, being singletons.

Proposition 2.2 Suppose that a sample of size n contains r species, s of them being singletons and the remaining n-s individuals belonging to r-s different types. Let x_i denote the number of individuals of type i in the sample, $i = 1, \ldots, r$-s. Then the UMLE of Θ is given by $\theta_i = x_i/n$ for the r-s

non-singleton species and the remaining individuals are all of different type and have a total proportion of s/n.

Proof. Consider the $r \times 1$ vector $\mathbf{x} = (x_1, \ldots, x_{r\text{-}s}, 1, \ldots, 1)^{\mathrm{T}}$ with x_i being the number of individuals of the ith non-singleton species, $i = 1, \ldots, r\text{-}s$. Note that the s singleton individuals can be considered indistinguishable, for if we substitute any of them by another (imaginary) type, the structure of $\mathbf{x}$ is not altered. This does not hold if we substitute a particular individual from a non-singleton species. Let θ_i denote the population proportion of the ith non-singleton species detected in the sample, $i = 1, \ldots, r\text{-}s$, and let $\phi_1, \ldots, \phi_{\mathrm{K}\text{-}r+s}$ represent the proportions of the remaining K-r+s species. The likelihood function of the sample, $L(\mathbf{x}|\theta, \phi)$, is given by

$$\begin{aligned} L(\mathbf{x}|\theta,\phi) \;=\; & \frac{n!}{x_1!\cdots x_{r\text{-}s}!1!\cdots 1!}\theta_1^{x_1}\cdots\theta_{r\text{-}s}^{x_{r\text{-}s}}\phi_1^1\cdots\phi_s^1\phi_{s+1}^0\cdots\phi_{\mathrm{K}\text{-}r+s}^0 + \\ & \frac{n!}{x_1!\cdots x_{r\text{-}s}!1!\cdots 1!}\theta_1^{x_1}\cdots\theta_{r\text{-}s}^{x_{r\text{-}s}}\phi_1^0\phi_2^1\cdots\phi_{s+1}^1\phi_{s+2}^0\cdots\phi_{\mathrm{K}\text{-}r+s}^0 \\ & +\ldots+ \\ & \frac{n!}{x_1!\cdots x_{r\text{-}s}!}\theta_1^{x_1}\cdots\theta_{r\text{-}s}^{x_{r\text{-}s}}\phi_1^0\cdots\phi_{\mathrm{K}\text{-}r}^0\phi_{\mathrm{K}\text{-}r+1}^1\cdots\phi_{\mathrm{K}\text{-}r+s}^1. \end{aligned} \tag{1}$$

Observe that there are (K-r+s)!/s!(K-r)! terms in the sum and that each term represents a different way of getting s singletons out of the K-r+s species. The set of the K classes in the population can be written as the union of the set of non-singletons in the sample and the set of singletons in

the sample and the unobserved species. Let E be the set of non-singleton species in the sample and let Pr(E) denote its probability. Similarly, let E^c denote the complement of event E which consists of the remaining K-r+s species and let $\Phi \equiv \Pr(E^c) = \sum_{i=1}^{K-r+s} \phi_i$.

Define events A and B as follows:

A = {x_1 individuals of type $1, \ldots, x_{r-s}$ individuals of type r-s given that a sample of size n-s is taken from E}

B = {s singletons given that a random sample of size s is taken from E^c}.

Then the likelihood function (1) can be written as

$$\begin{aligned} L(\mathbf{x}|\theta,\phi) &= \frac{n!}{s!(n-s)!}[\Pr(E)]^{n-s}[\Pr(E^c)]^s \Pr(A)\Pr(B) \\ &= \frac{n!}{s!(n-s)!}(1-\Phi)^{n-s}\Phi^s \Pr(A)\Pr(B) \end{aligned} \tag{2}$$

But,

$$\begin{aligned} \Pr(A) &= \frac{(n-s)!}{x_1!\cdots x_{r-s}!}\left(\frac{\theta_1}{\sum\theta_t}\right)^{x_1}\cdots\left(\frac{\theta_{r-s}}{\sum\theta_t}\right)^{x_{r-s}} \\ &= \frac{(n-s)!}{x_1!\cdots x_{r-s}!}\theta_1^{x_1}\cdots\theta_{r-s}^{x_{r-s}}(1-\Phi)^{-\sum_{i=1}^{r-s}x_i} \\ &= \frac{(n-s)!}{x_1!\cdots x_{r-s}!}\theta_1^{x_1}\cdots\theta_{r-s}^{x_{r-s}}(1-\Phi)^{-(n-s)} \end{aligned} \tag{3}$$

since the probability of event A reduces to a multinomial probability with r-s cells with proportions $\theta_i/\sum\theta_t$, $\sum_{t=1}^{r-s}\theta_t$=1-Φ, and $\sum_{i=1}^{r-s}x_i$ =n-s. Applying

(3) to (2) we obtain

$$L(\mathbf{x}|\theta,\phi) = \frac{n!}{x_1!\cdots x_{r-s}!s!}\theta_1^{x_1}\cdots\theta_{r-s}^{x_{r-s}}\Phi^s \Pr(\mathrm{B}) \tag{4}$$

Observe that Pr(B) depends on the relative proportion of the ϕ_i's with respect to Φ and not on any of the θ_i's. Therefore we can maximize Pr(B) independently of the rest of the expression in (4). But from Proposition 2.1, we know that Pr(B) is maximized if each individual in the group E^c belongs to a different species and the maximum value occurs at ϕ_i=1/(N-n+s). Hence, maximizing (4) is equivalent to maximizing

$$L(\mathbf{x}|\theta,\phi) = \frac{n!}{x_1!\cdots x_{r-s}!s!}\theta_1^{x_1}\cdots\theta_{r-s}^{x_{r-s}}\Phi^s \tag{5}$$

But, $\theta_1 + \cdots + \theta_{r-s} + \Phi = 1$ and $x_1 + \cdots + x_{r-s} + s = n$ so that the likelihood equation (5) represents the likelihood function of a multinomial sample for which the usual MLE of Θ exists namely, $\hat{\theta}_i = x_i/n$ and $\hat{\Phi} = s/n$.

As mentioned earlier, the UMLE maximizes over a larger set therefore the next result is stated without proof.

Corollary 2.1 The likelihood function of $\mathbf{x}$ evaluated at the UMLE is greater or equal to the likelihood function evaluated at the MLE.

Corollary 2.2 If the proportion of undetected species is Π, then the UMLE of Π is given by

$$\hat{\Pi} = s\left(\frac{1}{n} - \frac{1}{N}\right).$$

Proof. The UMLE of Φ is s/n, so by subtracting the proportion occupied by the s singletons detected in the sample the results follows.

With this approach, the use of singletons as statistics to estimate the proportion of undetected species appears very natural. In fact when N$\rightarrow \infty$, the UMLE of $\Phi \rightarrow s/n$ which is precisely Good's estimator (1953). If there are r-s non-singletons in the sample, we can think of sampling from r-s+1 urns where each of the first r-s urns have been sampled at least once and thus completely covered. Taking s individuals from the last urn contributes s/n to the coverage of the sample, hence the proportion of singletons in the sample can be used to estimate and to test hypotheses on the proportion occupied by the last urn. The next section explains the methodology.

3. Hypothesis Test

From Corollary 2.2, with N big, a test on the size of Π can be approximated with a test on Φ. Let the random variable S be the number of singletons in the sample. To test the hypothesis $H_0 : \Phi \geq p_0$ a first approximation is to calculate $\Pr(S \leq s|\Phi = p_0)$ where S is binomial with parameters n and Φ.

Therefore, for a given α and p_0 the decision rule is:

Reject $H_0 : \Phi \geq p_0$ if

$$Pr(S \leq s|\Phi = p_0) = \sum_{i=0}^{s} \frac{n!}{i!(n\text{-}i\ )!}(1-p_0)^{n\text{-}i}{p_0}^i < \alpha. \tag{6}$$

Engen (1978) reported a "suspiciously small" probability when evaluating Eldridge statistics of fully inflected words in American newspaper English (Eldridge, 1911). In that work n=43989 words were sampled from which 2976 were singletons. Assuming N is infinite, the UMLE of the proportion of uncovered species is equal to $2976/n$=0.068. Engen (1978, p. 67) wrote: "The initial (prior) probability that none of the words that did not appear in the sample should be observed is actually as small as $(1-0.068)^{43989} \approx 10^{-914}$. This number may at first glance look suspiciously small, but it is well known in probability theory that quite unlikely events do occur." The observed event is not unlikely at all. Nothing is wrong in the above calculation if we refer to the *particular* set of words that did not come up in the sample. Assuming N is infinite, the UMLE of the proportion of uncovered species is equal to $2976/n$=0.068. To test the hypothesis $H_0 : \Phi \geq p_0$ we use (6). Calculating $\Pr(S \leq 2976|\Phi = p_0)$ gives that for $\alpha = .05$ we reject H_0 for values of p_0 bigger than 0.0697. That is, the coverage is at least 93.03%. For $\alpha = .01$, the coverage is at least 92.95%.

We can improve this approximation using the following argument. Assuming no prior knowledge of the composition of the target population, we can base our decision on having achieved a given coverage when some species are found repeatedly in the sample in a frequency beyond what we would expect if the proportion of undetected species is very large. In a sample of size one nothing can be stated on the proportion occupied by this species until we take a second sample whose outcome is a Bernoulli random variable with probability of success equal to the proportion of the previous species. In general, we can test on the proportion occupied by some particular species, S_i, based on the number of times we find it among the maximum possible n-1. The number of repeated samples of this particular species S_i is then a binomial random variable with parameters n-1 and $p=\theta_i$.

According to Proposition 2.2, the singletons belong to a set of species with total proportion Φ, with the characteristic that each species is represented by a single individual. This implies that their contribution to the coverage of the sample is negligible, 1/N, and thus the coverage achieved is due mainly to the non-singleton species. Therefore, we can test on the proportion occupied by the non-singleton species found using the total number of "repeated" samples from this group. Each repetition corresponds to a "success" and thus "failures" corresponds to every singleton found.

To establish the independence and constant probability required in the assumptions for a binomial distribution we analyze the frequency of failures. If w non-singletons out of r species are found, the maximum number of failures possible is n-w. The decision on the coverage achieved by the particular set of species chosen, the non-singletons, can be based on the number of failures, singletons, that have a constant probability Φ of being detected at any one of n-w trials. Since w=r-s, it follows that the distribution of S, the number of singletons in the sample, is binomial with parameters n-r+s and p=Φ. The decision rule is then:

Reject $H_0 : \Phi \geq p_0$ if

$$Pr(S \leq s|\Phi = p_0) = \sum_{i=0}^{s} \frac{(n\text{-}r\text{+}s)!}{i!(n\text{-}r\text{+}s\text{-}i)!}(1-p_0)^{n\text{-}r\text{+}s\text{-}i}{p_0}^i < \alpha. \qquad (7)$$

Applying the above rule on Eldridge's data we have r=6001 so that we assume S has a binomial distribution with parameters n=40964 and p=Φ. For α = .05 we reject H_0 for values of p_0 bigger than 0.0748 providing a coverage of at least 92.52%. For α = .01, the coverage is at least 92.43%. With such a large sample no improvement, compared with the calculation using rule (6), is found. However, as suggested by the results from a moderate Monte Carlo study to be presented in the next section, when small sample sizes are considered the effectiveness of the method is noticeable.

4. MONTE CARLO SIMULATIONS

This section reports the results of a Monte Carlo study designed to test the hypothesis $H_0 : \Phi \geq p_0$ using the rejection rules (6) and (7). Six different populations were used with 15, 30, 50, 100, 200, and 500 classes. Figure 1 shows the distributions of classes for these populations. Observe that population 6 has 500 equally likely classes. The simulations performed on every population for every combination of α and p_0 were as follows. An initial sample of size 5 was taken. For fixed α and p_0, H_0 was tested using the rejection rule (7), and an additional randomly chosen individual is added until H_0 is rejected. This rule is referred as "Repeated Observations". As a comparison, the rejection rule (6) is also evaluated. This method is referred as "Total". During the simulations it is always possible to record the exact sample size at which the desired coverage was achieved. Thus, for each simulation, three sample sizes were recorded: the sample size when the intended coverage was achieved and the sample sizes when H_0 was rejected according to each method. Also, for every simulation, the true coverage is evaluated and it was noted whether the intended coverage was reached or not.

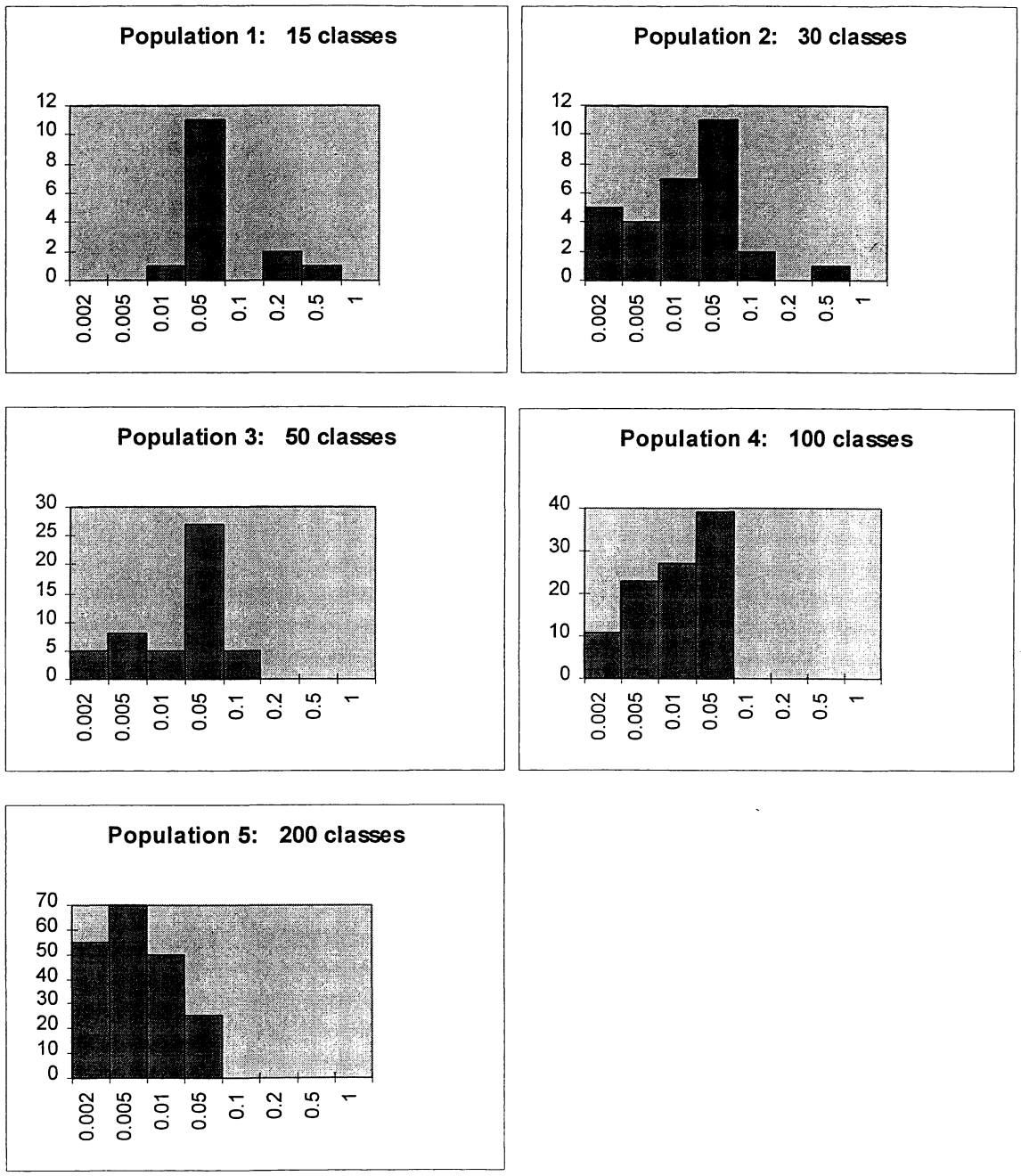
We considered α values of 0.05 and 0.01, and p_0 values of 0.05, 0.1, 0.2 and 0.3 for intended coverages of 95, 90, 80 and 70 percent respectively. This sampling procedure was repeated 10000 times on each population. Tables

1-4 summarize the results from these simulations. Given in the tables are average and standard deviation values of the coverage and the sample size at which the coverage was achieved. Also given is the proportion of the 10000 repetitions for which the desired coverage was achieved. This proportion is referred as "proportion of success"

Evaluating the efficacy of a given stopping rule is a difficult task. Relevant measures are the coverage achieved, the frequency in which an intended coverage is achieved, frequency of success, and the sample size required. The task is especially difficult since it is possible to obtain a high proportion of successes with an insensitive stopping rule. As a rule of thumb, a good procedure should have a proportion of success slightly larger than 1-α, small sample size and a mean coverage at least as the desired one.

Examination of the tables shows the mean coverage of the 10000 simulations for each population is always greater then the intended one with both methods. However, the frequency of success is always closer to 1-α for the "Repeated" method than for "Total". Also, larger sample sizes are required for the "Repeated" procedure than for "Total", but this is due to the fact that the latter method stops early, hence it has a smaller proportion of successes.

Fig. 1 Class Distribution



* Horizontal labels are upper limits for each interval.

Table 1. Method "Repeated observations"

Alpha = 0.05

	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5	Pop. 6
0.7	0.863 ± 0.078 32.76 ± 7.3 0.963	0.847 ± 0.077 43.54 ± 9.8 0.952	0.838 ± 0.063 77.36 ± 11.1 0.967	0.827 ± 0.046 150.33 ± 15.5 0.99	0.812 ± 0.034 276.41 ± 21.8 0.998	0.811 ± 0.022 834.35 ± 26.7 1
0.8	0.911 ± 0.062 46.03 ± 8.9 0.961	0.898 ± 0.053 63.89 ± 13.2 0.951	0.894 ± 0.045 106.94 ± 14.6 0.959	0.884 ± 0.033 203.91 ± 19.1 0.986	0.874 ±0.025 381.66 ± 26.9 0.995	0.876 ± 0.018 1045.52 ± 28.7 0.9999
0.9	0.961 ± 0.037 71.65 ± 15.8 0.9487	0.949 ± 0.030 116.3 ± 23.5 0.95	0.948 ± 0.025 171.66 ± 22.5 0.961	0.942 ± 0.020 322.36 ± 29.6 0.970	0.936 ± 0.015 605.66 ± 42.2 0.9834	0.938 ± 0.012 1396.95 ± 36.6 0.9978
0.95	0.983 ± 0.021 111.36 ± 20.3 0.9421	0.973 ± 0.017 198.87 ± 39.4 0.9258	0.975 ± 0.015 264.63 ± 38.1 0.9329	0.971 ± 0.012 485.38 ± 47.9 0.9405	0.967 ± 0.009 912.64 ± 66.9 0.9618	0.969 ± 0.008 1748.80 ± 50.7 0.985

Mean coverage ± std
Mean sample size ± std
Proportion of success

Table 2. Method "Total"

Alpha = 0.05

	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5	Pop. 6
0.7	0.824 ± 0.1 26.27 ± 7.5 0.910	0.808 ± 0.097 34.56 ± 9.3 0.877	0.782 ± 0.081 60.15 ± 11.1 0.860	0.766 ± 0.06 116.29 ± 15.2 0.864	0.749 ± 0.044 213 ± 21.5 0.869	0.729 ± 0.028 653.92 ± 28.6 0.8513
0.8	0.89 ± 0.071 39.09 ± 9.0 0.911	0.877 ± 0.065 53.28 ± 12.9 0.903	0.863 ± 0.058 87.77 ± 14.5 0.875	0.846 ± 0.043 165.59 ± 18.7 0.865	0.834 ± 0.032 308.79 ± 26.7 0.857	0.821 ± 0.022 864.29 ± 30.1 0.8462
0.9	0.952 ± 0.042 65.73 ± 12.2 0.9084	0.941 ± 0.035 102.86 ± 23.4 0.913	0.936 ± 0.030 149.69 ± 22.1 0.895	0.927 ± 0.025 278.95 ± 28.9 0.871	0.919 ± 0.018 522.49 ± 41.2 0.8556	0.913 ± 0.015 1228.3 ± 36.5 0.8334
0.95	0.979 ± 0.026 103.53 ± 20.6 0.9097	0.971 ± 0.019 183.8 ± 40.1 0.8991	0.970 ± 0.017 240.33 ± 36.5 0.8863	0.965 ± 0.014 438.98 ± 47.3 0.86	0.961 ± 0.010 823.43 ± 65.7 0.8577	0.958 ± 0.010 1596.2 ± 49.0 0.8315

Mean coverage ± std
Mean sample size ± std
Proportion of success

Table 3. Method "Repeated Observations"

Alpha = 0.01

	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5	Pop. 6
0.7	0.895 ± 0.067 39.74 ± 7.2 0.987	0.877 ± 0.061 52.27 ± 9.8 0.987	0.864 ± 0.049 87.63 ± 10.7 0.996	0.843 ± 0.041 162.61 ± 15.5 0.997	0.826 ± 0.032 294.7 ± 21.8 1.00	0.82 ± 0.022 858.79 ± 26.2 1.00
0.8	0.937 ± 0.045 56.03 ± 8.5 0.998	0.918 ± 0.038 76.39 ± 13.5 0.988	0.91 ± 0.037 119.85 ± 14.2 0.992	0.896 ± 0.031 220.78 ± 19.7 0.995	0.884 ± 0.023 405.01 ± 27.7 0.9983	0.882 ± 0.017 1072.58 ± 28.9 1.00
0.9	0.972 ± 0.027 69.84 ± 30.7 0.9848	0.96 ± 0.02 139.76 ± 23.6 0.993	0.958 ± 0.022 192.92 ± 24.2 0.982	0.95 ± 0.018 350.3 ± 31.2 0.99	0.942 ± 0.014 642.82 ± 43.2 0.9933	0.942 ± 0.012 1429.5 ± 37.2 0.9989
0.95	0.989 ± 0.014 136.62 ± 22.9 0.9822	0.98 ± 0.012 240.08 ± 42.6 0.9777	0.98 ± 0.012 302.26 ± 41.4 0.9752	0.975 ± 0.01 529.63 ± 50.3 0.977	0.971 ± 0.008 972.54 ± 71.3 0.9833	0.9722 ± 0.008 1793.5 ± 51.7 0.9938

Mean coverage ± std
Mean sample size ± std
Proportion of success

Table 4. Method "Total"

Alpha = 0.01

	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5	Pop. 6
0.7	0.869 ± 0.075 33.02 ± 7.3 0.966	0.847 ± 0.076 42.85 ± 9.8 0.952	0.82 ± 0.062 69.54 ± 10.5 0.958	0.789 ± 0.053 127.21 ± 15.2 0.944	0.767 ± 0.040 228.75 ± 21.4 0.947	0.741 ± 0.027 676.46 ± 28.2 0.9352
0.8	0.924 ± 0.052 48.69 ± 8.6 0.983	0.9 ± 0.047 64.24 ± 13.1 0.965	0.883 ± 0.049 99.67 ± 14.2 0.941	0.863 ± 0.038 180.99 ± 19.1 0.946	0.848 ± 0.029 330.46 ± 27.0 0.9393	0.830 ± 0.022 889.36 ± 30.0 0.921
0.9	0.961 ± 0.032 77.47 ± 11.1 0.9661	0.954 ± 0.024 125.04 ± 23.4 0.976	0.948 ± 0.027 169.72 ± 23.3 0.946	0.937 ± 0.021 305.37 ± 30.2 0.946	0.927 ± 0.017 557.26 ± 42.3 0.9331	0.919 ± 0.014 1260.3 ± 38.2 0.9087
0.95	0.987 ± 0.016 128.38 ± 22.7 0.9737	0.979 ± 0.013 224.4 ± 42.2 0.965	0.976 ± 0.014 275.95 ± 40.3 0.9515	0.970 ± 0.012 481.94 ± 49.4 0.9351	0.965 ± 0.019 880.87 ± 69.7 0.9317	0.962 ± 0.009 1637.4 ± 49.9 0.9005

Mean coverage ± std
Mean sample size ± std
Proportion of success

REFERENCES

Bunge, J., and Fitzpatrick, M. (1993), "Estimating the Number of Species: A Review," *Journal of the American Statistical Association* 88, 364–373.

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

Engen, S. (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.

Good, I. J. (1950), "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.

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

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