Authors: Jack C. Yue[1] and Murray K. Clayton[2]

1. Department of Statistics, National Chengchi University, Taipei 11605, Taiwan, R.O.C.

2. Department of Statistics, University of Wisconsin-Madison, Madison, WI. 53706, U.S.A.

Title of the paper: Sequential Sampling in the Search for New Shared Species

Running Title: Sequential Sampling for Shared Species

Number of Words: 6,410

Correspondence to:

Dr. Jack C. Yue

Department of Statistics

National Chengchi University, Taipei 11605

Taiwan, ROC

Tel: 886-2-2938-7695, Fax: 886-2-2939-8024

e-mail: csyue@nccu.edu.tw

## Summary

In microbial sciences, as well as other disciplines, it is often valuable to sample communities in a sequential or group sequential manner, in order to determine their structure or their similarity. We develop sequential sampling procedures to accomplish this by first assuming that one observation is drawn with replacement from each population at a time. Suppose that the sampling is terminated after $n$ pairs of observations and $k$ shared species were discovered, and assume that we receive payoff $h(k)-cn$, where $h(k)$ is non-decreasing and the sampling cost $c$ is non-negative. Similar to Rasmussen and Starr (1979), we show that an optimal stopping rule exists if $h(k+1)-h(k)$ is non-increasing. An analogous result holds for group sequential sampling. This leads to using an estimate of the probability of discovering new shared species as a stopping indicator for comparing two populations with respect to the similarity index. We show by simulation and real examples that this is a feasible approach which can help to reduce the sample size.

Key Words: Optimal stopping; Comparing populations; Similarity index; Discovering new species; Simulation

## 1. Introduction

In ecology, it is often of interest to describe the structure of a community, or to describe or compare the structure of two or more communities. A simple descriptor of a single community is the number of species – its "species richness." Because sampling is usually incomplete, and because of the size and complexity of many communities, this problem is surprisingly challenging – Bunge and Fitzpatrick (1993) refer to a bibliography of over 550 papers on the topic as of 1991. Our primary interest in this paper is the even greater challenge of comparing two communities, but we will borrow ideas from the richness literature in the process. We shall use the terms "community" and "population" interchangeably for the rest of the study.

This work is motivated by applications in microbial ecology, but it has relevance to more general settings. As an example, it is of interest to compare the communities of microbes that exist in two different soils, one representing agricultural land, and one representing undisturbed land (Harris, 2003, Miyoshi et al. 2005). The community, in this context, consists of the microbes present in some volume of soil extracted from the site and subjected to laboratory analyses. In such a setting, there could easily be $10^9$ or more microbes in each extracted volume. In a typical application, a random sample of 100 to 1000 microbes is taken from each volume on a single sampling occasion. This "group sampling" can be done repeatedly to obtain further data. Obtaining the soil volumes is comparatively easy, but sampling microbes from within these is time-consuming, expensive, and nontrivial, because it relies on advanced molecular techniques (Standing et al., 2007). Consequently, there is interest in limiting the number of observations taken. Of course, issues of cost arise in numerous settings beyond the microbial situation.

Whether describing a single community or comparing two communities, we

separate two actions: the process of sampling from communities, and the estimation of richness or similarity once sampling has stopped. Here we focus largely on the issue of sampling, and we begin with a theoretical development based on an earlier sampling approach that addressed the question of species richness for a single community. We adapt this to cover the more realistic group sampling methods outlined above, and then extend it to the problem of comparing two communities.

In an elegant paper, Rasmussen and Starr (1979) supposed that a community is sampled sequentially, one representative at a time (with replacement), and that each observation is classified according to species. Suppose that when sampling stops, there are *n* observations and *k* species discovered. They take a decision-theoretic perspective and assume a payoff equal to $h(k) - cn$, where $h(k+1) - h(k)$ is non-increasing function and *c* is a non-negative cost per observation. An important special case occurs when the reward function *h* is the identity function. In that case, it is optimal to stop sampling when $u(n) \le c$, where *u(n)* is the probability of discovering a new species given that *n* observations have already been taken.

At the time of stopping, a simple estimate of the number of species is the observed number of species, *k*, although more complex estimates could also be used. Just as important, whether sampling proceeds as outlined above, or by some other method, knowing that *u(n)* is large should be an important indicator that the current sample size is likely inadequate. The general utility of an estimate of *u(n)* seems to be largely unappreciated in the ecological literature, however.

Rasmussen and Starr proposed estimating *u(n)* with Turing's estimate *r/n* where *r* is the number of species with a single representative in the sample. This estimate provides a good approximation if the sample size is sufficiently large. Other estimates have also been proposed, such as Chao (1981), Clayton and Frees (1987),

and Lee (1989).

Our ultimate goal is to construct a sampling approach to be used for the *comparison* of communities. Several similarity indices have been proposed in the literature (e.g., Smith et al., 1996; Yue and Clayton, 2005); here we will focus on the one proposed by Yue and Clayton (2005). Given the use of the index, we seek a sequential sampling rule that will indicate how long we should sample to ensure, in some sense, a sufficiently large sample upon which to base an estimate of similarity. Following Rasmussen and Starr, we will take a decision-theoretic approach to this, and we will also describe practical solutions.

Before proceeding, we note that the term "species" can be interpreted quite broadly, and besides its traditional Linnaean use in biology, it has also been used to refer to "operational taxonomic units" in modern metagenomics studies (Schloss and Handelsman, 2008), to words in linguistics (Eforn and Thisted, 1976), to coins in numismatics (Esty, 1984 and 1985), to bugs in computer code (Chao et al., 1993), typographical errors in text (Nayak, 1989; Okello et al., 2005) and, quite recently, in the development of indexes for web search engines; comparable extensions exist for the term "community."

In the next section we first explore an optimal stopping rule for discovering new shared species of two populations. A Turing type estimate of the probability of discovering new shared species is also discussed and we employ it in a stopping rule. We then extend this work to cover the group sequential setting. Some simulations and real data appear in Section 3, and in Section 4 we make some concluding remarks.

## 2. Optimal Stopping for Discovering Shared Species

Suppose there are two populations and let $\vec{p} = (p_1, p_2, \ldots, p_s)$ and $\vec{q} = (q_1, q_2, \ldots, q_s)$

denote the species proportions of the two populations, where *s* is the number of distinct species in the pooled communities. Initially, we suppose the populations are sampled sequentially, with one observation taken from each population at each sampling occasion. The following argument is similar to that of Rasmussen and Starr (1979).

The random variable

$$v(n) = \sum_{i=1}^{s} p_i q_i \times I(X_i(n) = Y_i(n) = 0) + \sum_{i=1}^{s} \left( p_i \times I(X_i(n) = 0, Y_i(n) > 0) + q_i \times I(X_i(n) > 0, Y_i(n) = 0) \right) \quad (1)$$

is the probability of discovering a new shared species, where $X_i(n)$ and $Y_i(n)$ are the numbers of occurrences for species *i* from *n* observations in each of populations 1 and 2, respectively. In other words, *v(n)* is the conditional probability that a shared species will be discovered in the next, $(n+1)^{st}$ draw from each population.

Suppose the payoff function for discovering a new shared species is of the form $w(n) = h(s_{12}(n)) - cn$, where the reward function *h(k)* is non-decreasing, *h(k+1)–h(k)* is non-increasing, $s_{12}(n)$ is the number of observed shared species in *n* pairs of observations, and $c > 0$ is a sampling cost. Then the expected payoff of taking the $(n+1)^{st}$ pair of observations is

$$E(w(n+1) \mid F_n) = h(s_{12}(n)) \cdot (1 - v(n)) + h(s_{12}(n) + 1) \cdot v(n) - c \cdot (n+1),$$

where $F_n$ is the observed information based on the first *n* pairs of observations. Thus, it is reasonable that sampling will be terminated, i.e., $E(w(n+1) \mid F_n) \le w(n)$, if and only if

$$\left(h(s_{12}(n) + 1) - h(s_{12}(n))\right) \cdot v(n) \le c, \quad (2)$$

and indeed the following result holds.

Theorem 1. If the reward function is such that $h(k+1)-h(k)$ is non-increasing, then it is optimal to stop sampling at time

$$n^* = \inf\{n \geq 0 : [h(s_{12}(n)+1) - h(s_{12}(n))] \cdot v(n) \leq c\}. \quad (3)$$

Proof: The proof is similar to that of Theorem 1 of Rasmussen and Starr (1979), who are interested in the number of new species in one population; their proof is based on the following inequality:

$$\sum_{i=1}^{s} p_i > 1 - c\big(h(s(n)+1) - h(s(n))\big)^{-1},$$

assuming that $p_i$ is a monotone non-increasing sequence. For discovering new shared species, we can show a similar inequality:

$$\sum_{i \in I_1} p_i q_i > 1 - c\big(h(s_{12}(n)+1) - h(s_{12}(n))\big)^{-1},$$

where $I_1$ is a subset of {1, 2, …, *s*} and has *n* members, where $I_1$ represents the list of currently observed species, and where the labels have been arranged such that $p_i q_i$ is a monotone non-increasing sequence. Because the expected sample size needed to discover the shared species $i$ is not larger than $\max\{p_i^{-1}, q_i^{-1}\}$, it is immediate that $E(n^*) < \infty$. The rest of the proof is similar to Rasmussen and Starr's and is omitted.□

Thus far we have an optimal stopping rule for estimating the number of species in a single population (due to Rasmussen and Starr) and an optimal rule for estimating the number of shared species in two populations. However, both of these rules are based on the notion of sampling one observation (resp. one pair of observations) at a time. As noted in Section 1, however, there are situations where group sampling would be desirable. In that light, we note that both Theorem 1 of Rasumssen and Starr (1979) and Theorem 1 above can be extended to the group sequential sampling

case. We shall begin by using the one-population case to show the extension. In the one-population case, Rasmussen and Starr (1979) showed that it is optimal to stop sampling once

$$\big(h(s(n)+1)-h(s(n))\big)\cdot u(n)\le c, \tag{4}$$

where *s(n)* is the number of observed species in *n* observations and *u(n)* is the conditional probability of discovering a new species defined above. If *h* is the identity function, this inequality is equivalent to requiring that the expected number of newly discovered species in a single draw be less than or equal to *c*. Suppose now that a group of *m* observations is drawn at every stage and the sampling cost is $m\cdot c$ per group. Then inequality (4) can be modified to yield the stopping rule: stop taking observations when it first happens that

$$E\big[s(n+m)\,|\,s(n)\big]\le m\cdot c\,. \tag{5}$$

(This rule is optimal within the class of rules based on sampling *m* observations at a time, although it is not optimal among the class of rules based on sampling one observation at a time.)

A simple estimate of the left hand side of equation (5) is not easy to obtain. However, because *u(n)* is non-increasing, we can construct an upper bound, i.e.,

$$E\big[s(n+m)\,|\,s(n)\big]=\sum_{l=1}^{m} l\times P(l\ new\ species\ in\ group\ of\ \ m\ obs.\,|\,s(n))\le\sum_{l=1}^{m} l\times u(n)$$

and thus it is never optimal to take more observations when

$$E\big[s(n+m)\,|\,s(n)\big]\le\sum_{l=1}^{m} l\times u(n)\le m\cdot c\,.$$

This yields the following result.

Theorem 2. If the reward function *h* is the identity function and a group of *m* observations is sampled at every stage, then it is never optimal to continue sampling

beyond time

$$n^* = \inf\{n \geq 0 : u(n) \leq c^*\}, \tag{6}$$

where $c^* = m \cdot c \Big/ \frac{m(m+1)}{2} = \frac{2c}{m+1}$.

Note that the group sampling setting in Theorem 2 can also be extended to the two-population case by replacing the payoff function with $w(n) = h(s_{12}(n)) - cn$, and using the stopping indicator *v(n)*.

Thus far we have developed stopping rules for both individual and group sampling scenarios, and for the goal of estimating the number of species or for estimating the number of shared species. Our primary goal, however, is to use a sampling rule such that, upon stopping, we can compare communities in terms of their similarity, as measured by the index in Yue and Clayton (2005), Jaccard's index, or some other index. Unfortunately, the complexity of such indices means that it is quite difficult to design an associated optimal stopping rule. Such a stopping rule would tell us how to optimally gather a sample such that we could best estimate similarity.

We resort, therefore, to a suboptimal approach, and evaluate it through simulations and applications to real data. Our proposed procedure builds on the preceding work. Specifically, we sample until $v(n) \leq c$, and then estimate similarity. If sampling is done one pair of observations at a time, then this rule results in sampling optimally to estimate the number of shared species, and then uses that sample to estimate similarity. (If sampling proceeds group-wise, then the procedure is suboptimal – within the class of all sampling rules – for estimating the number of shared species). Although a reward function in terms of the similarity index might not be monotone, there is still something to be said about optimality.

Theorem 3. Suppose we have two populations, with species proportions $\vec{p} = (p_1, p_2, \ldots, p_s)$ and $\vec{q} = (q_1, q_2, \ldots, q_s)$, respectively. Then, $\hat{\theta}(n) \to \theta(n)$ in probability, where $\theta = \dfrac{\sum_{i=1}^{s} p_i q_i}{\sum_{i=1}^{s} p_i^2 + \sum_{i=1}^{s} q_i^2 - \sum_{i=1}^{s} p_i q_i}$ and its estimate $\hat{\theta}(n) = \dfrac{\sum_{i=1}^{s} \hat{p}_i(n) \cdot \hat{q}_i(n)}{\sum_{i=1}^{s} \hat{p}_i^2(n) + \sum_{i=1}^{s} \hat{q}_i^2(n) - \sum_{i=1}^{s} \hat{p}_i(n) \cdot \hat{q}_i(n)}$ is obtained by plugging in the nonparametric maximum likelihood estimator $\hat{p}_i(n) = \dfrac{\#of\ \ species\ i\ occurences\ for\ \ pop.1}{n}$ and $\hat{q}_i(n) = \dfrac{\#of\ \ species\ i\ occurences\ for\ \ pop.2}{n}$ (Yue and Clayton, 2005).

Proof: The proof is straightforward and is omitted.

With respect to the payoff function $\hat{\theta}(n) - cn$, suppose we stop group sampling at the point

$$n^* = \inf\{n \geq 0 : v(n) \leq c\}. \tag{7}$$

Unlike (3), the stopping rule in (7) is not necessarily optimal and seemingly only provides an upper bound for stopping the sampling. However, heuristically speaking, if $c$ is small, $v(n) \leq c$ means that there is a very small probability of discovering new shared species. This indicates that the similarity index $\hat{\theta}(n)$ will not change too much and sampling should stop. Therefore, if the sampling cost $c$ is chosen properly, $v(n) \leq c$ may be a possible candidate for judging whether there are enough observations to evaluate the similarity of two populations. We next use simulations to explore the use of $v(n) \leq c$ as a stopping indicator.

## 3. An Adaptive Rule

Parallel to the development of Rasmussen and Starr (1979), we note that the existence of the optimal stopping rule is predicated on knowing *v(n)*, which is in fact unobservable. Also parallel to the work of Rasmussen and Starr, we develop a Turing-type estimate for the probability of discovering shared species given by

$$v'(n) = \sum_{i=1}^{s} \frac{I(X_i(n)=1)}{n} + \sum_{i=1}^{s} \frac{I(Y_i(n)=1)}{n} + \sum_{i=1}^{s} \frac{I(X_i(n)=Y_i(n)=1)}{n} - \sum_{i=1}^{s} \frac{I(X_i(n)=0, Y_i(n)=1)}{n} - \sum_{i=1}^{s} \frac{I(X_i(n)=1, Y_i(n)=0)}{n} \quad (8)$$

(See the Appendix for further discussion of the derivation of this estimator.) Note that the estimate *v'(n)* is always non-negative since the first two terms in (8) are always not smaller than the last two terms. The expected bias in (8) is

$$E(v'(n)-v(n)) = \sum_{i=1}^{s} p_i^2(1-p_i)^{n-1} + \sum_{i=1}^{s} q_i^2(1-q_i)^{n-1} + \sum_{i=1}^{s} p_i^2(1-p_i)^{n-1} q_i^2(1-q_i)^{n-1} - \sum_{i=1}^{s} p_i^2(1-p_i)^{n-1}(1-q_i)^n - \sum_{i=1}^{s} (1-p_i)^n q_i^2(1-q_i)^{n-1} \quad (9)$$

(see Appendix) which converges to 0 as the number of observations *n* goes to infinity.

Our goal is to sample sequentially to estimate community similarity, and based on the above work we propose doing so by first using $v'(n) \le c$ as a stopping indicator, and then estimating $\theta$ with $\hat{\theta}$ calculated from the sample obtained upon stopping. To evaluate this approach we use the decision-theoretic ideas cited above. That is, we estimate $E(s(n)) - nc$, the estimated number of species, discounted by the cost of sampling.

There are two additional considerations. First, to use a simulation effectively, we need to know the true parameters being estimated. Moreover, specification of $\theta$ does not provide complete information about community structure, and thus we must

specify that for each community. We take two approaches to this. First, we use geometric distributions to model the distribution of species within each community. That is, we assume that $p_i \propto \alpha^i$ and likewise for $q_i$. In addition, we consider three types of shared species patterns for the geometric distribution (Yue et al., 2001; Yue and Clayton, 2005). For Type 1, the shared species are dominant in both populations, for Type 2 the shared species are dominant in one population but rare in the other, and for Type 3 the shared species are rare in both populations. As an alternative for specifying community structure, we use the structure of communities for which sample data are available. The idea in this is to take existing sample data, pretend that these represent the entire communities, and sample from these in simulations. This will provide additional simulation results, but from populations that arise naturally, and that may not follow simple parametric models. Our focus will be on the third of these, an example from microbial ecology. However, to examine additional possible community structures, we also look at communities of macroorganisms.

To emulate a practical approach for sampling microbial communities, we use a group-sampling approach whereby each community is sampled $m = 10$ members at a time. Group sampling may be less natural for macroorganisms, but again, we rely on those examples more as exemplars of community structure, rather than as communities where our approach would typically be applied. Finally, note that the behavior of $v'(n)$ is erratic for small $n$. For example, the estimated probability $v'(1)$ can be as small as 0, or as large as 3! To stabilize the behavior of $v'(n)$ it is practical to assume that some initial sample has been taken. In the simulations below, this is taken to be between 20 and 1000. The use of an initial sample, and the use of group sampling here is similar to sequential sampling in some clinical trials.

Note that the simulations conducted in this study are based on an Intel-based PC, using the statistical software R, version 2.4.0. All results are from 1,000 simulation replications for each case.

Example 1. Suppose that the species proportions of the two populations follow geometric distributions and $p_i = q_i \propto \alpha^i$ with $\alpha$=0.8. Let the numbers of species in the two populations be 30, and the number of shared species be 15 or 30. The results are each based on 1,000 simulation runs.

[Attach Table 1 here.]

Table 1 lists the similarity indices corresponding to the three types of shared species structures for geometric distributions. The differences in the true similarity values and the observed similarity values are within two times the standard deviations for both Types 2 and 3. When the sample size is large, the differences between the true similarity values and the observed similarity values are about two times the standard deviations in Type 1, regardless of whether $s_{12}$ = 15 or 30. This might suggest that if the shared species are dominant, then the sample required may be larger than in the unbalanced cases. This makes intuitive sense: if the populations are quite different, early sampling might reveal that, while if the populations are very similar, it might require extensive sampling to detect any subtle differences.

Example 2. In addition to considering the same settings as in Example 1, we include a sampling cost and sampling is terminated once $v'(n) \le c$, where *v'(n)* is defined in (8), and c = 0.001. Let the numbers of species in two communities satisfy

$s_1 = s_2 = 30$ or $s_1 = s_2 = 60$. Also, let the number of shared species be $s_{12} = 15$ or 30 if $s_1 = s_2 = 30$, and $s_{12} = 30$ or 60 if $s_1 = s_2 = 60$. For computational simplicity, we set the starting sample size to be 100 and the sampling increment to be 10 observations at a time from each population, in each simulation.

For each simulation, we record the number of observations drawn from each population when the sampling is terminated. Next, we calculated the estimated similarity index $\hat{\theta}$ and the observed 95% confidence interval of $\hat{\theta}$ via the delta method, checking whether the true $\theta$ is in the 95% confidence interval. We can then use the information to check if the sampling rule can provide a reliable estimate for the true $\theta$. Generally speaking, if we take a sample, it is not necessary that the sample information can provide an accurate estimate of the parameter. If the proposed stopping rule can give acceptable estimates, this implies that the proposed rule is not only cost efficient but is also a good practical rule. The summary of these numbers are listed in Table 2, including the averages and their standard errors (inside the parentheses) calculated from 1,000 computer simulation runs.

[Attach Table 2 here.]

Except for the cases with very small similarity index values, the true $\theta$ is covered in the intervals of the average of $\hat{\theta}$ and 2 times the standard error calculated from delta method. This can show us briefly how the estimated similarity index $\hat{\theta}$ behaves using the proposed stopping rule.

The coverage probability of the type 1 cases show promising results, and those probabilities are closer to the nominal 95% value if we increase the number of species

from $s_1 = s_2 = 30$ to $s_1 = s_2 = 60$. It seems that, if empirically the communities are of the Type 1 case, we can use the average estimated similarity index values and their s.e. via the delta method to build confidence intervals for $\theta$. In other words, once the sampling is terminated, the observed $\hat{\theta}$ can provide a fair estimate for the true $\theta$.

The Types 2 and 3 geometric cases behave somewhat differently. The observed results for the average sample sizes are quite severe, either with very large or zero s.e. Since there are at most 60 species and the values of *v'(n)* depends on the number of species appearing exactly once, *v'(n)* is likely to have a discrete behavior. That is, because the shared species in the Types 2 and 3 cases are rare in at least one community and the starting sample size is 100, the sampling is likely to terminate early unless there are singletons. However, the shared species are rare and on average it requires many observations to observe these shared species. For example, if we increase the starting sample size to 1,000 in the Types 2 and 3 cases of $s_1 = s_2 = 30$, the coverage probabilities are very close to the 95% value. However, taking at least 1,000 observations for a community with 30 species seems excessive.

We next study three real examples with larger numbers of species. The data will be treated as the populations, and the observed species proportions and the similarity values are treated as the real values. We conduct sampling on these pseudo-populations until the proposed stopping rule *v'(n)* is less or equal to the sampling cost. Then, we compare observed $\hat{\theta}$ to the "real" $\theta$ and evaluate if using *v'(n)* is a feasible choice, by checking whether the confidence intervals build at terminating sampling would cover the real $\theta$. Also, we will check the average sample sizes, comparing with the original data size, to see if the proposed sampling rule is cost efficient.

Example 3. The Taiwan Bird data (Yue et al., 2001) contain two communities of wild

birds, with 184 different species and 144,963 observations. These two wild bird communities are at two heavily polluted river estuaries in northwestern Taiwan. The observed similarity index $\theta$ is .8402. We shall treat this value as the true similarity index. Also, since the shared species of bird data are close to the Type 1 case in the previous simulation, we expect that using the optimal stopping rule would have good results.

For practical reasons, the starting sample size is 1,000 for each population, with increments of 10 observations in each simulation. The reason for choosing a larger starting value and the increment of 10 observations is that the wild birds are relatively easy to observe and they often come in groups. Sampling continues unless $v'(n) \leq .005$ or .001. We summarize the value of $\hat{\theta}$, the sample size, and whether the true $\theta$ lies in the 95% confidence interval of $\hat{\theta}$. We also record the theoretical value $v(n)$ and the observed $v'(n)$ when sampling is terminated; the results are listed in Table 3.

[Attach Table 3 here.]

From Table 3, for c = .005 and .001, the average of $\hat{\theta}$ is close to the true $\theta$ and the probability of the 95% confidence interval $\hat{\theta}$ covering the true $\theta$ is also close to its nominal value 0.95. This suggests the feasibility of using the probability of discovering shared species in a stopping rule. In particular, sampling is terminated relatively early and yet still provides a fairly accurate estimate of the true $\theta$. In the case of sampling terminated when $v'(n) \leq 0.005$, the average sample size at stopping is less than 6,000 (i.e., 2 times 2,968), which is about 1/20 of the original sample size – a very appealing outcome. Similar results appear in the case of sampling terminated when $v'(n) \leq 0.001$, where the average sample size is less than 30,000,

which is 5 times that in the case of sampling cost 0.005, and about 1/4 of the original sample size.

Also, the average value of the true *v(n)* is close to the sampling cost. This result supports the idea that the unconditional expectation of *v'(n)* can be used to approximate that of *v(n)*, and the average bias is small. (Note: Approximately, for the sampling cost = 0.005 and 0.001, the standard errors of *v'(n)* and the sample size are in proportion to the cost.)

Example 4. The Panama Crab data (Smith et al., 1996) contain two populations, with 74 different species and 5,831 observations. The crab samples collected are living in two coral communities at two locations in Panama. The observed similarity index $\theta$ is .3591. We shall treat this value as the true similarity index and proceed as in Example 3. This means that the observed species proportions are treated as the true proportions and the sampling is terminated until $v'(n) \le c$. The starting sample size is 100 for each population, with increments of 10 observations in each simulation. The simulation results are listed in Table 4. Again, the shared species of crab data are close to the Type 1 case and the idea of the optimal stopping rule would have some good results.

Again, we can see that $v'(n) \le .005$ or $v'(n) \le .001$ are feasible stopping indicators. In either case, the average sample size needed is about 40% and 100% of the original sample size, for $v'(n) \le .005$ or $v'(n) \le .001$, respectively. Also, the average of $\hat{\theta}$ is close to the true $\theta$ and the probability of the 95% confidence interval $\hat{\theta}$ covering the true $\theta$ is also close to its nominal value 0.95.

[Attach Table 4 here.]

Example 5. The microorganisms data (Table 1 in Miyoshi et al., 2005) contain two populations, with 14 different species and 125 observations. They are from a total of 247 clones of 16S rRNA genes from microorganisms captured by 0.2- and 0.1-m-pore-size filters from sediment. The microorganism samples collected are from in two underground water communities at two locations in Japan. There are two data sets and here we use the one with filter size 0.2. The observed similarity index between communities $\theta$ is .7449. Again, we treat this value as the true similarity index and proceed as in Examples 3 and 4 -- the results are listed in Table 5.

[Attach Table 5 here.]

Because there are not a lot of observations (about 1/50 of those in Example 4), *v'(n)* would behave like a step function. If the sampling cost is chosen to be small, like in previous two examples, the sampling will be terminated only when there are no species appearing exactly once. Since there are about 100 observations, the sampling cost is set to be 0.01 and 0.05. Also, in order to avoid early stopping, the starting sample size is either 20 or 50 for each population, with increments of one observation in each simulation. Unlike in the previous two examples, the coverage probabilities are less than the nominal probability 0.95, except for the case of starting sample size 50 and sampling cost 0.01. The average value of sample size in the case of starting sample 50 and sampling cost 0.01 is about 82, or 164 for two populations, which is larger than the observed value 125. This might suggest that more observations are needed to have a better idea of the similarity level.

## 4. Conclusion and Discussions

Some authors (e.g., Esty, 1986; Bunge and Fitzpatrick, 1993, I.J.Good, 1991, personal communication, cited in Bunge and Fitzpatrick) have argued that the presence of rare species in communities means that we can never be assured of characterizing them fully. If, indeed, we accept the notion that it is impossible to guarantee perfect characterization of populations, then instead we should seek to characterize them "well enough" – that is, to balance the goal of learning as much as possible about the populations while recognizing the finite limits of resources available to do so. Rasmussen and Starr and the above development provide one way to do this, which leads itself to a method for sampling "enough."

In this paper we have extended a simple decision theoretic approach of Rasmussen and Starr (1979), to the examination of population similarity. If the reward of discovering new shared species $h(k)$ is such that $h(k+1) - h(k)$ is non-increasing, then an optimal stopping rule for estimating the number of shared species can be formulated based on the probability of discovering new shared species in the two populations. Simulation studies support the possibility of adapting an empirical estimate of this probability $v'(n)$ into a stopping rule, and combining that stopping rule with an estimate of population similarity. The examples based on real data show promising results.

However, according to our simulation, if the probability of observing species occurring exactly once is small (like the Type 2 and Type 3 cases), then the sampling is likely to stop too early. In that case, $v'(n)$ gives us little information. This is similar to using a Turing estimate to estimate the number of species in a population, where the estimate relies solely on the number of observed species and the species appear exactly once. Because the Turing type estimate is highly discrete in the cases

where it is not easy to observe the species appearing exactly once, the sample size could have large variance. We can see from Table 2 that all of the geometric distribution cases show very large variances. We expect that this is the case in general, especially given a small number of species.

Therefore, in order to have an acceptable result for applying the optimal stopping rule, we recommend checking the observed species structures first. If the shared species follow a Type 1 structure, the observed similarity index would be a feasible tool to measure the community similarity and we can enjoy the advantage of sequential approach by reducing the sample sizes. Fortunately, many data sets we encounter are of the Type 1 structure. However, if the shared species are close to the Type 2 or 3, many observations may be required before getting a good estimate of the similarity index value, and a larger starting sample size is suggested.

Simplicity is the advantage of using the Turing estimate, but it also suffers from its discreteness. Of course, the Turing estimate can be extended to a function of species appearing fewer than 10 times (Chao et al., 2000). We suggest using *v'(n)* with care if the number of observed species is small and recommend taking more initial observations in this case.

There is another interesting result in our simulation that, although the observed similarity index of two identical populations is close to the true value 1, the 95% confidence interval fails to provide good coverage of true $\theta$. For the number of species being 30, even a sample of 3,000 cannot guarantee a 95% confidence interval covering $\theta = 1$. Note that the calculation of confidence intervals is based on delta method, for obtaining the standard errors and this approximation might be inadequate for this sample size. This suggests that it may require a much larger sample or alternative approach to determine whether two populations are identically distributed.

This paper has focused on sequential stopping rules for estimating the number of shared species in two populations, and for estimating similarity of two populations. The proposed stopping rule is quite simple: stop sampling when the (empirical) estimate of discovering a new shared species is sufficiently small. In the introduction, we noted that this probability is itself an interesting quantity. We conclude by calculating $v'(n)$ for each of the three real examples provided in the preceding section. Because our estimator requires equal sample sizes from each population, we approximate $v'(n)$ by conservatively using the minimum number of observations taken per population. Respectively, then, we find $v'(n) = .0005$ for the Taiwan Bird data; .0072 for the Crab data; and .033 for the Microbial data. Using a value of $c = .01$ would suggest that the former two data sets are adequate in size, and that some further sampling of the microbial data would be in order. Although it might not always be possible to sample sequentially in the manner outlined in this paper, calculating $v'(n)$ might still be a useful tool for evaluating the adequacy of a given sample size.

## APPENDIX: Turing type estimate *v'(n)*

The probability of discovering new shared species after *n* observations, from (1), is

$$v(n) = \sum_{i=1}^{s} p_i q_i \times I(X_i(n) = Y_i(n) = 0)$$
$$+ \sum_{i=1}^{s} \left( p_i \times I(X_i(n) = 0, Y_i(n) > 0) + q_i \times I(X_i(n) > 0, Y_i(n) = 0) \right)$$

and its expectation is

$$E(v(n)) = \sum_{i=1}^{s} \left( p_i (1-p_i)^n q_i (1-q_i)^n + p_i (1-p_i)^n [1-(1-q_i)^n] + q_i (1-q_i)^n [1-(1-p_i)^n] \right)$$
$$= \sum_{i=1}^{s} \left( p_i (1-p_i)^n + q_i (1-q_i)^n + (1-p_i)^{n+1} (1-q_i)^{n+1} - (1-q_i)^n (1-p_i)^n \right)$$
$$= \sum_{i=1}^{s} p_i (1-p_i)^n + \sum_{i=1}^{s} q_i (1-q_i)^n + \sum_{i=1}^{s} (1-q_i)^n (1-p_i)^n (p_i q_i - p_i - q_i) \qquad (a)$$
$$= \sum_{i=1}^{s} p_i (1-p_i)^n + \sum_{i=1}^{s} q_i (1-q_i)^n + \sum_{i=1}^{s} \left( (1-p_i)^{n+1} (1-q_i)^{n+1} - (1-q_i)^n (1-p_i)^n \right) \quad (b)$$
$$\cong E(u_1(n)) + E(u_2(n)) + \sum_{i=1}^{s} E\left( I(X_i(n+1) = Y_i(n+1) = 0) - I(X_i(n) = Y_i(n) = 0) \right)$$

It should be noted that the last approximation is due to the fact that the upper bound of the summation is different. Also, although equation (b) is useful for interpretation, it is easier to find the estimate via equation (a) since the parameters of the form $p_i(1-p_i)^n$ can be estimated via Turing's estimate but parameters of the form $(1-p_i)^n$ cannot. Therefore, a natural estimate of *v(n)* would be

$$v'(n) = \sum_{i=1}^{s} \frac{I(X_i(n) = 1)}{n} + \sum_{i=1}^{s} \frac{I(Y_i(n) = 1)}{n} + \sum_{i=1}^{s} \frac{I(X_i(n) = Y_i(n) = 1)}{n}$$
$$- \sum_{i=1}^{s} \frac{I(X_i(n) = 0, Y_i(n) = 1)}{n} - \sum_{i=1}^{s} \frac{I(X_i(n) = 1, Y_i(n) = 0)}{n}$$

The unconditional expectation of *v'(n)* is

$$E(v'(n)) = \sum_{i=1}^{s} p_i (1-p_i)^{n-1} + \sum_{i=1}^{s} q_i (1-q_i)^{n-1} + \sum_{i=1}^{s} p_i (1-p_i)^{n-1} q_i (1-q_i)^{n-1}$$
$$- \sum_{i=1}^{s} p_i (1-p_i)^{n-1} (1-q_i)^n - \sum_{i=1}^{s} (1-p_i)^n q_i (1-q_i)^{n-1}$$

which gives the bias approximately as

$$E(v'(n)-v(n)) \cong \sum_{i=1}^{s} p_i^2(1-p_i)^{n-1} + \sum_{i=1}^{s} q_i^2(1-q_i)^{n-1} + \sum_{i=1}^{s} p_i^2(1-p_i)^{n-1} q_i^2(1-q_i)^{n-1}$$
$$- \sum_{i=1}^{s} p_i^2(1-p_i)^{n-1}(1-q_i)^n - \sum_{i=1}^{s} (1-p_i)^n q_i^2(1-q_i)^{n-1}.$$

Similar to Turing's estimate, the bias converges to 0 as the number of observations increases and it converges at an exponential rate.

## REFERENCES


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

Chao, A., Hwang, W., Chen, Y., and Kuo, C. (2000). Estimating the Number of Shared Species in Two Communities. Statistica Sinica 10: 227-246.

Chao, A., Ma, M., and Yang, M.C.K. (1993). Stopping rule and estimation for recapture debugging with unequal detection rates. *Biometrika* 80: 193-201.

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

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.

Efron, B. and Thisted, R. (1976). Estimating the Number of Unseen Species: How Many Words did Shakespeare Know? Biometrika 63: 435-447.

Esty, W.W. (1984). Confidence Intervals for an Occupancy Problem Estimator Used by Numismatists. Mathematical Scientists 9: 111-115.

Esty, W.W. (1985). Estimation of the Number of Classes in a Population and the Coverage of a Sample. Mathematical Scientists 10: 41-50.

Esty, W.W. (1986). Te Efficiency of Good's Nonparametric Coverage Estimator. The Annals of Statistics 14; 1257-1260.

Feller, W. (1968). An Introduction to Probability Theory and its Applications, vol. 1, 3rd ed. New York: Wiley.

Frank, O. (1978). Estimation of the Number of Connected Components in a Graph by using a Sampled Subgraph. Scandinavian Journal of Statistics 5: 177-188.

Harris, J.A. (2003). Measurements of the soil microbial community for estimating the success of restoration. European Journal of Soil Science 54: 801-808.

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, Department of Statistics.

Miyoshi, T., Iwatsuki, T., and Naganuma, T. (2005). Phylogenetic Characterization of 16S rRNA Gene Clones from Deep-Froundwater Microorganisms that pass through 0.2-Micrometer-Pore-Size Filters. Applied and Environmental Microbiology 71(2): 1084-1088.

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

Okello, J.B.A., Wittemyer, G., Rasmussen, H.B., Douglas-Hamilton, I., Nyakaana, S., Arctander, P., and Siegismund, H.R. (2005). Noninvasive Genotyping and Mendelian Analysis of Microsatellites in African Savannah Elephants. Journal of Heredity 96(6): 679-687.

Rasmussen, S.L. and Starr, N. (1979). Optimal and Adaptive Stopping in the Search for New Species. Journal of the American Statistical Association 74:661-667.

Smith, W., Solow, A.R., and Preston, P.E. (1996). An Estimator of Species Overlap Using a Modified Beta-binomial Model. Biometrics 52: 1472-1477.

Standing, D., Baggs, E.M., Wattenbach, M., Smith, P., and Killham, K. (2007). Meeting the challenge of scaling up processes in the plant–soil–microbe system. Biology and fertility of soils 44: 245-257.

Starr, N. (1979). Linear Estimation of the Probability of Discovering a new Species. Annals of Statistics 7:644-652.

Yue, J.C., Clayton, M.K., and Lin, F. (2001). A Nonparametric Estimator of Species Overlap. Biometrics 57:743-749.

Yue, J.C. and Clayton, M.K. (2005). An Overlap Measure based on Species Proportions. Comm. Statist. Theory Methods 34:2123-2131.

Table 1. Similarity indices of Geometric distribution
(Numbers inside the parentheses are the standard deviations.)

| n | $s_{12} = 15$ | | |
|---|---|---|---|
| | Type 1 | Type 2 | Type 3 |
| 50 | .7630(.1274) | .0036(.0051) | .00049(.00104) |
| 100 | .8610(.0598) | .0039(.0037) | .00060(.00076) |
| 200 | .9250(.0317) | .0040(.0024) | .00058(.00047) |
| 500 | .9665(.0141) | .0040(.0016) | .00060(.00026) |
| 1000 | .9818(.0073) | .0041(.0011) | .00061(.00018) |
| 2000 | .9922(.0026) | .0042(.0007) | .00062(.00010) |
| True | .9975 | .0042 | .00062 |

| n | $s_{12} = 30$ | | |
|---|---|---|---|
| | Type 1 | Type 2 | Type 3 |
| 50 | .7633(.0928) | .0079(.0077) | -- |
| 100 | .8644(.0558) | .0080(.0053) | -- |
| 200 | .9267(.0336) | .0082(.0038) | -- |
| 500 | .9700(.0136) | .0083(.0023) | -- |
| 1000 | .9843(.0072) | .0083(.0016) | -- |
| 2000 | .9921(.0038) | .0084(.0013) | -- |
| True | 1 | .0084 | -- |

Note: When $s_{12} = 30$, Type 3 is identical to Type 1. The results are based on 1,000 simulation runs.

Table 2. Using sampling cost for Geometric distributions

| | | | $\theta$ | $\hat{\theta}$ | Coverage of $\theta$ | Sample Size |
|---|---|---|---|---|---|---|
| $s_1 = s_2 = 30$ | $s_{12}$ =15 | Type 1 | .9975 | 0.9279 (.0534) | 0.891 (.0099) | 271.4 (142.5) |
| | | Type 2 | .0042 | 0.0028 (.0026) | 0.586 (.0156) | 1,189 (1,112) |
| | | Type 3 | .00062 | 0.00021 (.00052) | 0.238 (.0135) | 496.1 (1037) |
| | $s_{12}$ =30 | Type 1 | 1 | 0.9718 (.0559) | 0.928 (.0082) | 1,977 (1331) |
| | | Type 2 | .0084 | 0.0073 (.0031) | 0.814 (.0123) | 2,668 (1,332) |
| | | Type 3 | -- | -- | -- | -- |
| $s_1 = s_2 = 60$ | $s_{12}$ =30 | Type 1 | .999997 | 0.9828 (.0286) | 0.939 (.0076) | 2105 (1243) |
| | | Type 2 | .00062 | 0.00013 (.00047) | 0.120 (.0103) | 373.2 (847.1) |
| | | Type 3 | $7.6\times10^{-7}$ | 0 (0.0000) | 0 (0.0000) | 100 (0.0000) |
| | $s_{12}$ =60 | Type 1 | 1 | 0.9858 (.0229) | 0.946 (.0071) | 3002 (2360) |
| | | Type 2 | .00002 | 0 (0.0000) | 0 (0.0000) | 100 (0.0000) |
| | | Type 3 | -- | -- | -- | -- |

Note: For the cases of $s_1 = s_2 = s_{12}$, the type 3 case is identical to the type 1 case. If the average sample size is 100, this indicates that the sampling stops right away. The results are based on 1,000 simulation runs.

Table 3. Using Sampling cost on Taiwan Bird Data

| Sampling Cost | $\hat{\theta}$ | Coverage of $\theta$ | *v(n)* | *v'(n)* | Sample Size |
|---|---|---|---|---|---|
| 0.005 | 0.8348 (.0163) | 0.9550 (.0066) | 0.0058 (.0030) | 0.0048 (.0002) | 2,968 (964.6) |
| 0.001 | 0.8394 (.0074) | 0.9520 (.0068) | 0.0011 (.0006) | 0.0010 (.00004) | 14,222 (4,014) |

Note: The true $\theta$ is 0.8402 and there are 184 species from 144,963 observations altogether. The results are based on 1,000 simulation runs.

Table 4. Using Sampling cost on Panama Crab Data

| Sampling Cost | $\hat{\theta}$ | Coverage of $\theta$ | *v(n)* | *v'(n)* | Sample Size |
|---|---|---|---|---|---|
| 0.005 | 0.3554 (.0247) | 0.9580 (.0063) | 0.0046 (.0024) | 0.0042 (.0010) | 1,203 (266.9) |
| 0.001 | 0.3577 (.0163) | 0.9470 (.0071) | 0.0017 (.0018) | 0.0009 (.00015) | 3,031 (1,038) |

Note: The true $\theta$ is 0.3591 and there are 74 species from 5,831 observations altogether. The results are based on 1,000 simulation runs.

Table 5. Using Sampling cost on Microorganism Data

(a) Starting = 20

| Sampling Cost | $\hat{\theta}$ | Coverage of $\theta$ | $v(n)$ | $v'(n)$ | Sample Size |
|---|---|---|---|---|---|
| 0.05 | 0.6923 (.1478) | 0.916 (.0088) | 0.0438 (.0340) | 0.0269 (.0224) | 24.44 (7.844) |
| 0.01 | 0.7122 (.1247) | 0.910 (.0090) | 0.0336 (.0300) | 0.0018 (.0040) | 50.95 (37.11) |

(b) Starting = 50

| Sampling Cost | $\hat{\theta}$ | Coverage of $\theta$ | $v(n)$ | $v'(n)$ | Sample Size |
|---|---|---|---|---|---|
| 0.05 | 0.7238 (.1059) | 0.912 (.0090) | 0.0228 (.0200) | 0.0207 (.0154) | 52.14 (3.981) |
| 0.01 | 0.7313 (.0897) | 0.941 (.0066) | 0.0163 (.0202) | 0.0037 (.0047) | 82.05 (27.38) |

Note: The true $\theta$ is 0.7499 and there are 14 species from 125 observations. The results are based on 1,000 simulation runs.