*Sequential Analysis*, 25: 103–115, 2006

ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474940500452288



# Some Probabilistic Aspects in the Discovery of Species

**A. Goswami and Bikas K. Sinha**
Theoretical Statistics and Mathematics Division, Indian Statistical Institute, Kolkata, India



**Abstract:** From a population containing $m$ different kinds of species in various proportions, units are drawn randomly until all the species are discovered. We show that, irrespective of whether the population is finite or infinite and, in the first case, whether the successive draws are with or without replacement, the number of trials needed for the discovery of all the species is stochastically minimized and hence has the minimum expected value when the species are as equally abundant as possible.

**Keywords:** Coupon collector's problem; Equally abundant; Species discovery; Stochastically smallest.

**Subject Classifications:** Primary 60E15; Secondary 60G40.

## 1. INTRODUCTION: MOTIVATION

Measurement of biodiversity is an important issue in ecological studies and conservation planning. For a general discussion on statistical issues in measurement of diversity, we refer to Gore and Paranjape (2001). Two quantitative aspects of diversity are regarded as central to its measurement. These are *species richness* (number of species of a taxon in a given geographical area) and *species evenness* (differences in relative abundance). The two can be combined in various ways to form diversity indices (see Patil and Taillie, 1982).

The issue of developing a sampling strategy for understanding of species richness and species abundance has been rightly emphasized in the context of measurement of biodiversity. A workable strategy starts with an assumed value of the initial size of species richness and an assumed pattern of species evenness and then recommends an appropriate initial size of the sample for data collection. Gradually, both the aspects are upgraded. Inherent in this practice is the empirical

Received September 23, 2004, Revised March 12, 2005, Accepted March 23, 2005
Recommended by N. Mukhopadhyay
Address correspondence to A. Goswami, Stat-Math Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India; Fax: 91-33-25773071; E-mail: alok@isical.ac.in

observation that the species-evenness distribution allows for a minimum sample size when, for a fixed size of species richness, the abundance rates are all equal (Gore and Paranjape, 1997, 2001).

Our aim in this article is to focus on the distribution of the "effort size" for a given richness size and for arbitrary evenness distribution. In the process we also provide an analytical proof of the empirical observation cited above (see Theorems 3.2 and 3.3).

## 2. MATHEMATICAL FORMULATION OF THE PROBLEM AND THE MAIN RESULT

We consider a population with $m$ different species. The number $m$—that is, the number of species—will be treated as known in this analysis. What are not known are the "abundance rates" of the various species. The "abundance rate vector" for the population is defined to be the vector

$$\underset{\sim}{\mathbf{p}} = (p_1, \ldots, p_m), \tag{2.1}$$

where $p_i$ denotes the abundance rate of the $i$th species, $i = 1, \ldots, m$. Here, by the abundance rate $p_i$ of a particular species labelled $i$ is meant the proportion of that species in the population. This, of course, may also be interpreted as simply the probability that a randomly drawn unit from the population comes from that particular species $i$.

In the case of a finite population, say, of size $N$, each $p_i$ is of the form $\frac{N_i}{N}$, where $N_i$ is a positive integer representing the size of the $i$th species in the population. Thus, here a typical abundance rate vector will look like

$$\underset{\sim}{\mathbf{p}} = \left(\frac{N_1}{N}, \ldots, \frac{N_m}{N}\right), \tag{2.2}$$

where $N_i$, $1 \le i \le m$, are positive integers with $\sum_{i=1}^{m} N_i = N$.

In case of infinite population, however, each $p_i$, $1 \le i \le m$, is allowed to take any value in the interval $(0, 1)$ subject to the condition that $\sum_{i=1}^{m} p_i = 1$.

The problem that we are going to consider can now be described as follows. Suppose that we keep on drawing units randomly from the population until all the $m$ species have been "discovered"—that is, at least one representative unit of each of the $m$ species has appeared in our sample. Let $T$ denote the number of trials needed. Clearly, $T$ is a random variable whose distribution depends on the abundance rate vector $\underset{\sim}{\mathbf{p}}$, and, of course, also on the sampling scheme—specifically, in the case of a finite population, whether the sampling is done with or without replacement. We will consider both the schemes in the finite population case. In the infinite population case, of course, it does not matter. Our main result can then be stated as follows.

*Main Result:* In both the cases of random sampling from an infinite population as well as random sampling with or without replacement from a finite population, the random variable $T$ is stochastically smallest when all the $m$ species are "(almost) equally abundant."

The phrase "(almost) equally abundant" needs an explanation. In the case of infinite population, it simply means that the abundance rates are equal for all the $m$ species, that is, the abundance rate vector is the special vector $\underset{\sim}{\mathbf{p}}_0 = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$.

In the finite-population case, the interpretation is a bit more involved, although the most natural. As before, let us denote the population size by $N$. Since each $p_i$ now has to be necessarily of the form $\frac{N_i}{N}$, where $N_1, \ldots, N_m$ are positive integers with $\sum_{i=1}^{m} N_i = N$, we cannot, in general, demand that the $p_i$ be all equal (equivalently, that the $N_i$ be all equal). This will require that $N$ be a multiple of $m$. Instead of imposing such a restriction on $N$, what we do, in general, is to define $N_0 = \left[\frac{N}{m}\right]$, that is to say that we define $N_0$ to be the unique integer such that $N = mN_0 + k$ with $0 \le k \le m-1$. Since $N \ge m$, such an $N_0$ is automatically positive. We say that the $m$ species are (almost) equally abundant if $N_i = N_0 + 1$ for $1 \le i \le k$ and $N_i = N_0$ for $k+1 \le i \le m$ and the corresponding abundance-rate vector will be also denoted by $\underset{\sim}{\mathbf{p}}_0$. This clearly means that, for given $N$ and $m$, the species are as equally abundant as possible.

Our main result can then be stated as follows: Let $P(T > t \mid \underset{\sim}{\mathbf{p}})$ denote the probability that the number of trials needed to discover all the $m$ species is greater than $t$, given that the abundance rate vector is $\underset{\sim}{\mathbf{p}}$. We will show that, in all the cases, $\underset{\sim}{\mathbf{p}} = \underset{\sim}{\mathbf{p}}_0$ minimizes this probability uniformly over all $t$.

In view of the fact that $E[T \mid \underset{\sim}{\mathbf{p}}] = \sum_{t=0}^{\infty} P(T > t \mid \underset{\sim}{\mathbf{p}})$, it will also follow from our result that $E[T \mid \underset{\sim}{\mathbf{p}}]$ attains a unique minimum when $\underset{\sim}{\mathbf{p}} = \underset{\sim}{\mathbf{p}}_0$, that is, in all the cases, the expected number of trials required to discover all the $m$ species is the smallest when the species are (almost) equally abundant.

## 3. STATEMENTS AND PROOFS OF MAIN RESULTS

We start by making some preliminary observations. The first observation, which is really trivial but will turn out to be useful in proving our main result, is that a permutation of the abundance rate vector does not change the distribution of $T$. In other words, if $\underset{\sim}{\mathbf{p}}$ is an abundance rate vector and $\underset{\sim}{\mathbf{p}}^*$ is just a permutation of (the coordinates of) $\underset{\sim}{\mathbf{p}}$, then the distribution of $T$, given the abundance rate $\underset{\sim}{\mathbf{p}}$, is the same as that given the abundance rate $\underset{\sim}{\mathbf{p}}^*$—in particular, $P(T > t \mid \underset{\sim}{\mathbf{p}}) = P(T > t \mid \underset{\sim}{\mathbf{p}}^*)$, for all $t$. This is because permuting the coordinates of the abundance rate vector simply amounts to relabeling the species and, therefore, does not affect $T$.

Our next observation may not have any direct bearing on our main result, but it is interesting in its own right. Let us consider any two different species, say, the $i$th and the $j$th, that is, $i \ne j$ are any two elements from $\{1, \ldots, m\}$. Let $\underset{\sim}{\mathbf{p}}$ denote the abundance rate vector. Denote by $f(t, i \mid \underset{\sim}{\mathbf{p}})$ the probability of the event that $T = t$ and that the last unit drawn comes from the $i$th species. Let $f(t, j \mid \underset{\sim}{\mathbf{p}})$ be similarly defined with the $j$th species in place of the $i$th. We then have the following result.

**Theorem 3.1.** *If $p_i < p_j$, then, for all $t > m$,*

$$f(t, i \mid \underset{\sim}{\mathbf{p}}) > f(t, j \mid \underset{\sim}{\mathbf{p}}). \tag{3.1}$$

It is easy to see that for $t = m$ (respectively, $t < m$), the two probabilities are equal (respectively, both are equal to 0).

*Proof.* The two cases—namely, the case when the units are drawn with replacement and the case when the population is finite and units are drawn without replacement—will be treated separately. However, our general argument will closely follow those in Rao et al. (2003).

Let us consider the "with replacement case" first. If $m = 2$, the result is almost immediate, because in that case, $f(t, 1 \mid \underset{\sim}{\mathbf{p}}) = (p_2)^{t-1} p_1$, while $f(t, 2 \mid \underset{\sim}{\mathbf{p}}) = (p_1)^{t-1} p_2$. Clearly, if $p_1 < p_2$, say, then $(p_2)^{t-1} p_1 > (p_1)^{t-1} p_2$ for all $t > 2$. We now assume $m > 2$ and fix $t > m$. Consider any permutation $\alpha = (i_1, \ldots, i_{m-2})$ of the other $m-2$ indices in $\{1, \ldots, m\}$ except $i$ and $j$ and let $\mathbf{n} = (n_1, \ldots, n_{m-1})$ be any $(m-1)$-tuple of positive integers with $\sum_{k=1}^{m-1} n_k = t - 1$. For $1 \le l \le m-1$, denote $A(i, j, l, \alpha, \mathbf{n})$ to be the event that the $m$ species are "discovered" in the order $i_1, \ldots, i_{l-1}, j, i_{l+1}, \ldots, i_{m-2}, i$ and these discoveries happen exactly at the 1st draw, at the $(1 + n_1)$th draw, ..., at the $\left(1 + \sum_{k=1}^{m-2} n_k\right)$th draw, and finally at the $\left(1 + \sum_{k=1}^{m-1} n_k\right)$th, i.e., the $t$th draw, respectively. The event $A(j, i, l, \alpha, \mathbf{n})$ is defined analogously with just the roles of $i$ and $j$ interchanged. It is then clear that

$$f(t, i \mid \underset{\sim}{\mathbf{p}}) = \sum P[A(i, j, l, \alpha, \mathbf{n}) \mid \underset{\sim}{\mathbf{p}}] \tag{3.2}$$

and

$$f(t, j \mid \underset{\sim}{\mathbf{p}}) = \sum P[A(j, i, l, \alpha, \mathbf{n}) \mid \underset{\sim}{\mathbf{p}}], \tag{3.3}$$

where the sums in both extend over all $\alpha, \mathbf{n}$, and $l$.

We now show that

$$P[A(i, j, l, \alpha, \mathbf{n}) \mid \underset{\sim}{\mathbf{p}}] \ge P[A(j, i, l, \alpha, \mathbf{n}) \mid \underset{\sim}{\mathbf{p}}] \quad \forall \alpha, \mathbf{n}, l, \tag{3.4}$$

with strict inequality holding for some $\alpha, \mathbf{n}, l$.

From the definition of the events $A(i, j, l, \alpha, \mathbf{n})$ and $A(j, i, l, \alpha, \mathbf{n})$, one can easily derive the following formulae:

$$P[A(i, j, l, \alpha, \mathbf{n}) \mid \underset{\sim}{\mathbf{p}}] = \prod_{r=1}^{m} p_r \prod_{k=1}^{l-1} \left(p_{i_1} + \cdots + p_{i_k}\right)^{n_k - 1} \prod_{k=l}^{m-1} \left(p_j + p_{i_1} + \cdots + p_{i_{k-1}}\right)^{n_k - 1}, \tag{3.5}$$

while

$$P[A(j, i, l, \alpha, \mathbf{n}) \mid \underset{\sim}{\mathbf{p}}] = \prod_{r=1}^{m} p_r \prod_{k=1}^{l-1} \left(p_{i_1} + \cdots + p_{i_k}\right)^{n_k - 1} \prod_{k=l}^{m-1} \left(p_i + p_{i_1} + \cdots + p_{i_{k-1}}\right)^{n_k - 1}. \tag{3.6}$$

Now $p_i < p_j$ implies $\left(p_j + p_{i_1} + \cdots + p_{i_{k-1}}\right)^{n_k - 1} \ge \left(p_i + p_{i_1} + \cdots + p_{i_{k-1}}\right)^{n_k - 1}$, whence one gets the inequality in (3.4).

Further, since $t > m$, we must have $n_k > 1$ for some $k$, so that strict inequality holds above whenever $l \le k$. This completes the proof in the "with replacement" case.

Next, let us consider the case when the population is finite and the units are drawn without replacement. Let us denote the population size by $N$ and the sizes of the different species by $N_1, \ldots, N_m$—these are positive integers adding to $N$. Here $p_i < p_j$ is the same as saying that $N_i < N_j$. As in the "with replacement" situation, the case $m = 2$ is easy to dispose of, because, in this case, $f(t, 1 \mid \underset{\sim}{\mathbf{p}}) = \frac{(N_2)_{t-1} N_1}{(N)_t}$, while $f(t, 2 \mid \underset{\sim}{\mathbf{p}}) = \frac{(N_1)_{t-1} N_2}{(N)_t}$ and if $N_1 < N_2$, say, then $(N_2)_{t-1} N_1 \geq (N_1)_{t-1} N_2$, with strict inequality holding for $t > 2$. Here $(n)_r$, for nonnegative integers $n$ and $r$, denotes the usual factorial power $n!/(n-r)!$. Turning now to the case $m > 2$, it is easy to see from the definitions of $f(t, i \mid \underset{\sim}{\mathbf{p}})$ and $f(t, j \mid \underset{\sim}{\mathbf{p}})$ that for $t > m$,

$$f(t, i \mid \underset{\sim}{\mathbf{p}}) = \sum_{\mathbf{n}} [(N)_t]^{-1} N_i \left[ \binom{N_j}{n_0} \prod_{r \neq i, j} \binom{N_r}{n_r} \right] (t-1)!, \tag{3.7}$$

and

$$f(t, j \mid \underset{\sim}{\mathbf{p}}) = \sum_{\mathbf{n}} [(N)_t]^{-1} N_j \left[ \binom{N_i}{n_0} \prod_{r \neq i, j} \binom{N_r}{n_r} \right] (t-1)!, \tag{3.8}$$

where both the sums are over all $(m-1)$-tuples $\mathbf{n}$ of positive integers $n_0$ and $n_r$, $r \neq i, j$ adding to $t-1$. Since $N_i < N_j$ implies that $N_i \binom{N_j}{n_0} \geq N_j \binom{N_i}{n_0}$, with strict inequality holding whenever $n_0 > 1$, the proof is complete. □

Let us now turn to our main result. The case of infinite population is easier to handle and, therefore, we first proceed with that case. For simplicity, let us denote

$$\Phi(t, m, \underset{\sim}{\mathbf{p}}) = P[T > t \mid m, \underset{\sim}{\mathbf{p}}], \tag{3.9}$$

that is, for an infinite population consisting of $m$ species with abundance rate vector $\underset{\sim}{\mathbf{p}}$, the probability that more than $t$ draws will be required to discover all the species is denoted by $\Phi(t, m, \underset{\sim}{\mathbf{p}})$. We then have the following theorem.

**Theorem 3.2.** *For an infinite population consisting of m different species with* $m \geq 2$,

$$\Phi(t, m, \underset{\sim}{\mathbf{p}}) \geq \Phi(t, m, \underset{\sim}{\mathbf{p}_0}) \quad \forall t \geq m, \tag{3.10}$$

*where* $\underset{\sim}{\mathbf{p}_0} = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$. *Further, strict inequality holds in* (3.10) *unless* $\underset{\sim}{\mathbf{p}} = \underset{\sim}{\mathbf{p}_0}$.

*Proof.* First we consider the case $m = 2$. In this case, a typical abundance rate vector is $\underset{\sim}{\mathbf{p}} = (p_1, p_2)$ and one has

$$\Phi(t, 2, \underset{\sim}{\mathbf{p}}) = (p_1)^t + (p_2)^t. \tag{3.11}$$

The desired result now is an easy consequence of the well-known fact that, subject to the conditions $p_1, p_2 \geq 0$, $p_1 + p_2 = 1$, the right-hand side of the above equation, for any $t \geq 2$, has a unique minimum when $p_1 = p_2 = \frac{1}{2}$.

To prove the result for general $m$, we will use induction on $m$. To be able to do that, that is, to get a formula that will relate $\Phi(\cdot, m, \cdot)$ with $\Phi(\cdot, m-1, \cdot)$, we need

some notations. For any $m$-vector $\underset{\sim}{\mathbf{p}} = (p_1, \ldots, p_m)$ of abundance rates and for any $1 \le i \le m$, let us denote $\underset{\sim}{\mathbf{p}}^{(i)}$ to be the $(m-1)$-vector of abundance rates defined by

$$\underset{\sim}{\mathbf{p}}^{(i)} = \left( \frac{p_1}{1-p_i}, \ldots, \frac{p_{i-1}}{1-p_i}, \frac{p_{i+1}}{1-p_i}, \ldots, \frac{p_m}{1-p_i} \right). \tag{3.12}$$

One can then use an easy conditioning argument to deduce that for any $m \ge 3$, any $\underset{\sim}{\mathbf{p}} = (p_1, \ldots, p_m)$, any $1 \le i \le m$, and any $t \ge m$,

$$\begin{aligned} \Phi(t, m, \underset{\sim}{\mathbf{p}}) = (1-p_i)^t + \sum_{s=t-(m-2)}^{t} \binom{t}{s} (p_i)^s (1-p_i)^{t-s} \\ + \sum_{s=1}^{t-(m-1)} \binom{t}{s} (p_i)^s (1-p_i)^{t-s} \Phi(t-s, m-1, \underset{\sim}{\mathbf{p}}^{(i)}). \end{aligned} \tag{3.13}$$

This is the connecting link between $\Phi(\cdot, m, \cdot)$ and $\Phi(\cdot, m-1, \cdot)$ that allows us to use induction. Assume, therefore, as an induction hypothesis, that the result (3.10) is true for $m-1$. This will imply, in particular, that

$$\Phi(t', m-1, \underset{\sim}{\mathbf{p}}^{(i)}) \ge \Phi\left( t', m-1, \left( \frac{1}{m-1}, \ldots, \frac{1}{m-1} \right) \right) \quad \forall t' \ge m-1, \tag{3.14}$$

with strict inequality holding unless $\underset{\sim}{\mathbf{p}}^{(i)} = \left( \frac{1}{m-1}, \ldots, \frac{1}{m-1} \right)$.

Using (3.14) for each $\Phi(t-s, m-1, \underset{\sim}{\mathbf{p}}^{(i)})$ appearing in the last sum on the right-hand side of (3.13) and then recombining all the terms in the resulting sum, one gets that for every $m \ge 3$, any $\underset{\sim}{\mathbf{p}} = (p_1, \ldots, p_m)$, any $1 \le i \le m$, and any $t \ge m$,

$$\Phi(t, m, \underset{\sim}{\mathbf{p}}) \ge \Phi\left( t, m, \left( \frac{1-p_i}{m-1}, \ldots, \frac{1-p_i}{m-1}, p_i, \frac{1-p_i}{m-1}, \ldots, \frac{1-p_i}{m-1} \right) \right), \tag{3.15}$$

with strict inequality holding unless the $p_j$, $j \ne i$ are all equal.

Now if $\underset{\sim}{\mathbf{p}} \ne \underset{\sim}{\mathbf{p}}_0$, there certainly is an $i$, $1 \le i \le m$, such that the $p_j$, $j \ne i$, are not all equal. Since $\Phi(t, m, \underset{\sim}{\mathbf{p}})$ remains invariant under permutation of coordinates of $\underset{\sim}{\mathbf{p}}$, we can, without loss of generality, take $i = 1$. Denoting $p_1$ by $p$, the above result says that

$$\Phi(t, m, \underset{\sim}{\mathbf{p}}) > \Phi\left( t, m, \left( p, \frac{1-p}{m-1}, \ldots, \frac{1-p}{m-1} \right) \right), \quad \forall t \ge m. \tag{3.16}$$

Note now that the right-hand side of the above inequality is a function of just a single variable $p \in (0, 1)$, for each $t \ge m$. Denoting this function by $g_t(p)$, our proof could be easily completed if we could show that the function $g_t$ on the open interval $(0, 1)$ attains a unique global minimum at $p = \frac{1}{m}$ (for all $t \ge m$)! Writing the explicit expression for $g_t(p)$, it is not difficult to show that $g_t'(\frac{1}{m}) = 0 \; \forall t \ge m$. But, the second derivative $g_t''(p)$ turns out to be not so easily tractable, making it difficult to conclude whether $p = \frac{1}{m}$ is even a local minimum.

A different route is used to complete the proof. Consider the function $f:[0,1]\to[0,1]$ defined as $f(p)=\frac{1-p}{m-1}$. Then, the inequality that we have proved above, using the induction hypothesis, simply says that for any $m\geq 3$, any $\underset{\sim}{\mathbf{p}}$, any $1\leq i\leq m$, and any $t\geq m$,

$$\Phi(t,m,\underset{\sim}{\mathbf{p}})\geq\Phi(t,m,(f(p_i),\ldots,f(p_i),p_i,f(p_i),\ldots,f(p_i))), \tag{3.17}$$

with strict inequality holding unless the $p_j$, $j\neq i$, are all equal.

We now complete the proof of (3.10) as follows. If $\underset{\sim}{\mathbf{p}}\neq\underset{\sim}{\mathbf{p}_0}$, then denoting $p_1=p$ and using the above notation, we have as argued above,

$$\Phi(t,m,\underset{\sim}{\mathbf{p}})>\Phi(t,m,(p,f(p),\ldots,f(p)))\quad\forall t\geq m. \tag{3.18}$$

But applying (3.17) now to the right-hand side of the inequality in (3.18) with $i=2$ yields

$$\begin{aligned}\Phi(t,m,(p,f(p),\ldots,f(p)))&\geq\Phi(t,m,(f(f(p)),f(p),f(f(p)),\ldots,f(f(p))))\\&=\Phi(t,m,(f(p),f(f(p)),\ldots,f(f(p)))),\end{aligned} \tag{3.19}$$

where the last equality follows from the invariance of $\Phi$ under permutations of coordinates of the abundance rate vector.

Let us denote by $f^{(n)}$, $n\geq 0$ the repeated iterates of the function $f$, that is, $f^{(0)}(p)=p$ and $f^{(n)}(p)=f(f^{(n-1)}(p))$, $n\geq 1$. If now $\underset{\sim}{\mathbf{p}}\neq\underset{\sim}{\mathbf{p}_0}$ and if $1\leq i\leq m$ is such that the $p_j$, $j\neq i$, are not all equal, then denoting $p_i=p$, using (3.18) first and then (3.19) repeatedly, we get that, for all $t\geq m$,

$$\begin{aligned}\Phi(t,m,\underset{\sim}{\mathbf{p}})&>\Phi(t,m,(f^{(0)}p,\ f^{(1)}(p),\ldots,f^{(1)}(p)))\\&\geq\Phi(t,m,(f^{(1)}(p),f^{(2)}(p),\ldots,f^{(2)}(p)))\\&\geq\cdots\cdots\\&\geq\Phi(t,m,(f^{(n)}(p),f^{(n+1)}(p),\ldots,f^{(n+1)}(p)))\\&\geq\cdots\cdots.\end{aligned} \tag{3.20}$$

Using now the Lemma 3.1 stated below and the fact that the function $\Phi(t,m,(p_1,\ldots,p_m))$ is continuous in $(p_1,\ldots,p_m)$, it follows from the string of inequalities in (3.20) that for $\underset{\sim}{\mathbf{p}}\neq\underset{\sim}{\mathbf{p}_0}$,

$$\Phi(t,m,\underset{\sim}{\mathbf{p}})>\Phi(t,m,\underset{\sim}{\mathbf{p}_0}),\quad\forall t\geq m, \tag{3.21}$$

which was to be proved. □

**Lemma 3.1.** *The function* $f(p)=\frac{1-p}{m-1}$ *on* $[0,1]$ *into* $[0,1]$ *has a unique fixed point at* $p_0=\frac{1}{m}$. *Moreover,* $p_0$ *is a globally attracting fixed point for* $f$ *in the sense that for any* $p\in[0,1]$, *the sequence* $\{f^{(n)}(p)\}$ *converges to* $p_0$ *as* $n\to\infty$.

*Proof.* That $p_0$ is the unique fixed point of the function $f$ (that is, $f(p) = p$ if and only if $p = \frac{1}{m}$) is immediate from the definition of $f$. For the other part, one simply needs to observe that $|f'(p)| = \frac{1}{m-1} < 1$ for $m \geq 3$ and apply, for example, Theorem 4.48 of Apostol (1974, p. 92). □

**Remark 3.1.** We would like to point out here that the problem of drawing from an infinite population (or drawing with replacement from a finite population) consisting of $m$ species has been discussed in Basu (1958) and also in Feller (1967, pp. 224–225). However, they consider only the special case when the species are all exactly equally abundant. Under this special assumption, they show, among other things, that the probability that, in a sample of size $n$, exactly $r$ distinct species appear is given by $\binom{m}{n} \frac{\Delta^r 0^n}{m^n}$ and that the expected number of trials needed to discover all the $m$ species equals $m \cdot \left[1 + \frac{1}{2} + \cdots + \frac{1}{m}\right]$.

Let us now turn to the case of a finite population, say, with size $N$. In this case, as noted already, the abundance rates are necessarily of the form $p_i = \frac{N_i}{N}$ where $N_i$ represents the size of the $i$th species in the population. Thus, the abundance rate vector is really determined by vector of the actual sizes of the $m$ different species, that is, by the "abundance vector" $\underset{\sim}{\mathbf{N}} = (N_1, \ldots, N_m)$, where $N_i, 1 \leq i \leq m$ are positive integers with $\sum_{i=1}^{m} N_i = N$. Recall also that, with $N_0$ being the unique positive integer such that $N = mN_0 + k$ for some $0 \leq k \leq m-1$, the $m$ species are said to be (almost) equally abundant if and only if exactly $k$ of the $N_i$'s are equal to $N_0 + 1$ and the remaining $m - k$ of the $N_i$'s are equal to $N_0$. In view of the permutation–invariance property of the probabilities that we are interested in, this is, for our purposes, readily seen to be equivalent to demanding that $N_1 = \cdots = N_k = N_0 + 1$ and $N_{k+1} = \cdots = N_m = N_0$. The corresponding abundance vector (again, unique up to permutaion) would be denoted by $\underset{\sim}{\mathbf{N_0}}$. For any abundance vector $\underset{\sim}{\mathbf{N}}$, we define $d(\underset{\sim}{\mathbf{N}}) = \max_{i,j} |N_i - N_j| = \max_i N_i - \min_i N_i$. It is then easy to see that $\underset{\sim}{\mathbf{N_0}}$ is the unique (up to permutaion) abundance vector with $d(\underset{\sim}{\mathbf{N_0}}) \leq 1$.

We now introduce a notation similar to the infinite-population case. Let $\Phi(t, N, m, \underset{\sim}{\mathbf{N}})$ denote now the probability that for a population with size $N$ consisting of $m$ species with abundance vector $\underset{\sim}{\mathbf{N}}$, more than $t$ draws will be required to discover all the species, that is,

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) = P[T > t \mid N, m, \underset{\sim}{\mathbf{N}}]. \tag{3.22}$$

This probability, of course, depends on whether the successive units are drawn from the population with or without replacement. But we will use the same notation for these probabilities in both the cases. Our main result in the finite-population case is stated in the following theorem.

**Theorem 3.3.** *For a finite population of size $N$ consisting of $m$ different species with $2 \leq m \leq N$,*

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) \geq \Phi(t, N, m, \underset{\sim}{\mathbf{N_0}}), \quad \forall t \geq m, \tag{3.23}$$

*irrespective of whether the units are drawn with or without replacement. Further, strict inequality holds in* (3.23) (*for all t, for which the left-hand side* >0), *unless* $\underset{\sim}{\mathbf{N}} = \underset{\sim}{\mathbf{N}}_0$ (*up to permutations*).

The proof is based on an elementary observation described below. This will play a role very similar to that played by Lemma 3.1 in the proof of Theorem 3.2. Let $m \geq 3$ and $N \geq m$ be positive integers. Let $\underset{\sim}{\mathbf{N}} = (N_1, \ldots, N_m)$ be an $m$-tuple of positive integers with $\sum_{i=1}^{m} N_i = N$. For any $i$ $(1 \leq i \leq m)$, let $N_{0;i}$ be the unique positive integer such that $N - N_i = (m-1)N_{0;i} + l$ with $0 \leq l \leq m-2$. Denote by $\underset{\sim}{\mathbf{N}}^{(i)}$ the $m$-tuple whose $i$th coordinate is the positive integer $N_i$, and, of the other $m-1$ coordinates, the first $l$ are all equal to the positive integer $N_{0;i}+1$, while the remaining are all equal to the positive integer $N_{0;i}$. We then have the following simple yet useful result.

**Lemma 3.2.** (a) *If* $\underset{\sim}{\mathbf{N}}$ *has an even number of coordinates with exactly half of them all equal to some positive integer* $N'$, *say, and the other half all equal to* $N'+2$, *then for some i and j,* $d((\underset{\sim}{\mathbf{N}}^{(i)})^{(j)}) = 0$.

(b) *If* $\underset{\sim}{\mathbf{N}}$ *is not of the special form as considered in* (a) *and if* $d(\underset{\sim}{\mathbf{N}}) > 1$, *then for some i,* $d(\underset{\sim}{\mathbf{N}}^{(i)}) < d(\underset{\sim}{\mathbf{N}})$.

The main content of the lemma that is crucial for our purposes may be understood as follows. Suppose that for any given abundance vector $\underset{\sim}{\mathbf{N}}$, the abundance vectors $\underset{\sim}{\mathbf{N}}^{(i)}$, $1 \leq i \leq m$, as defined above, are thought of as those that are "accessible from $\underset{\sim}{\mathbf{N}}$ in one step." The vectors accessible from the $\underset{\sim}{\mathbf{N}}^{(i)}$ are, in turn, thought of as accessible from $\underset{\sim}{\mathbf{N}}$ "in two steps," and so on. Then what the lemma says is that if $\underset{\sim}{\mathbf{N}}$ is any abundance vector with $d(\underset{\sim}{\mathbf{N}}) > 1$, then there is an abundance vector $\underset{\sim}{\mathbf{N}}^*$, accessible from $\underset{\sim}{\mathbf{N}}$ (in two steps if $\underset{\sim}{\mathbf{N}}$ is of the form described in (a) and in one step in all other cases) such that $d(\underset{\sim}{\mathbf{N}}^*) < d(\underset{\sim}{\mathbf{N}})$. Let us postpone the proof of Lemma 3.2 and proceed first to prove Theorem 3.3, assuming the lemma.

*Proof of Theorem* 3.3. Just like Theorem 3.2, the proof here also will be through induction on $m$. So, we first consider the case $m = 2$.

With $\underset{\sim}{\mathbf{N}} = (N_1, N_2)$, where $1 \leq N_2 \leq N_1 \leq N-1$ and $N_1 + N_2 = N$, we have

$$\Phi(t, N, 2, \underset{\sim}{\mathbf{N}}) = \frac{\binom{N_1}{t} + \binom{N_2}{t}}{\binom{N}{t}} \quad \forall t \geq 2, \tag{3.24}$$

in case the successive draws are without replacement, while if the draws are with replacement, then

$$\Phi(t, N, 2, \underset{\sim}{\mathbf{N}}) = \left(\frac{N_1}{N}\right)^t + \left(\frac{N_2}{N}\right)^t \quad \forall t \geq 2. \tag{3.25}$$

It is now easy to see that if $\underset{\sim}{\mathbf{N}} \neq \underset{\sim}{\mathbf{N}}_0$, that is, $N_1 \geq N_2 + 2$, then

$$\Phi(t, N, 2, \underset{\sim}{\mathbf{N}}) > \Phi(t, N, 2, \underset{\sim}{\mathbf{N}}^*), \tag{3.26}$$

in both the cases, where $\underset{\sim}{\mathbf{N}}^* = (N_1 - 1, N_2 + 1)$. Indeed, in the case of sampling without replacement, inequality (3.26) follows from the fact that $\binom{N_1-1}{t-1} > \binom{N_2}{t-1}$ for all $2 \le t \le N_1$, if $N_1 - 1 > N_2$, while in the case of sampling with replacement, it is a consequence of the fact that, for $t \ge 2$, the function $x \mapsto x^t + (1-x)^t$, $x \in (0,1)$ is strictly increasing for $x \in \left(\frac{1}{2}, 1\right)$ and strictly decreasing for $x \in \left(0, \frac{1}{2}\right)$. By repeated use of (3.26), one can easily deduce that the desired result (3.23) is true for $m = 2$.

To use induction now, we need a notation. For any $m \ge 3$, $N \ge m$ and any $\underset{\sim}{\mathbf{N}} = (N_1, \ldots, N_m)$, we denote, $\mathbf{N}_{(\mathbf{i})} = (N_1, \ldots, N_{i-1}, N_{i+1}, \ldots, N_m)$, for $1 \le i \le m$. It is then easy to see that, for all $t \ge m$ and any $1 \le i \le m$, one has

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) = \frac{\binom{N-N_i}{t}}{\binom{N}{t}} + \sum_{s=t-(m-2)}^{t} \frac{\binom{N_i}{s}\binom{N-N_i}{t-s}}{\binom{N}{t}} + \sum_{s=1}^{t-(m-1)} \frac{\binom{N_i}{s}\binom{N-N_i}{t-s}}{\binom{N}{t}} \Phi(t-s, N-N_i, m-1, \mathbf{N}_{(\mathbf{i})}), \tag{3.27}$$

in case of sampling without replacement, and

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) = \left(\frac{N-N_i}{N}\right)^t + \sum_{s=t-(m-2)}^{t} \binom{t}{s}\left(\frac{N_i}{N}\right)^s\left(\frac{N-N_i}{N}\right)^{t-s} + \sum_{s=1}^{t-(m-1)} \binom{t}{s}\left(\frac{N_i}{N}\right)^s\left(\frac{N-N_i}{N}\right)^{t-s} \Phi(t-s, N-N_i, m-1, \mathbf{N}_{(\mathbf{i})}), \tag{3.28}$$

in case of sampling with replacement.

Using the induction hypothesis that the result (3.23) is true for $m-1$ and recalling the notation introduced just before Lemma 3.2, one obtains from the above that for all $m \ge 3$, $N \ge m$, $\underset{\sim}{\mathbf{N}} = (N_1, \ldots, N_m)$, and, for all $1 \le i \le m$,

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) \ge \Phi(t, N, m, \underset{\sim}{\mathbf{N}}^{(i)}), \quad \forall t \ge m, \tag{3.29}$$

with strict inequality holding unless $d(\mathbf{N}_{(\mathbf{i})}) \le 1$.

Now if $\underset{\sim}{\mathbf{N}} \ne \underset{\sim}{\mathbf{N}}_{\mathbf{0}}$, that is, $d(\underset{\sim}{\mathbf{N}}) > 1$, then we consider two cases and use Lemma 3.2.

*Case 1:* $\underset{\sim}{\mathbf{N}}$ is of the form described in Lemma 3.2(a). In this case, as the proof of Lemma 3.2 will show, for any $1 \le i \le m$, $d(\mathbf{N}_{(\mathbf{i})}) = 2$, so that the induction hypothesis will actually give

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) > \Phi(t, N, m, \underset{\sim}{\mathbf{N}}^{(i)}), \quad \forall t \ge m. \tag{3.30}$$

But we also have, according to Lemma 3.2(a), that, for some $j$, $(\underset{\sim}{\mathbf{N}}^{(i)})^{(j)} = \underset{\sim}{\mathbf{N}}_{\mathbf{0}}$, so that, by (3.29), we have

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}^{(i)}) \ge \Phi(t, N, m, \underset{\sim}{\mathbf{N}}_{\mathbf{0}}), \quad \forall t \ge m. \tag{3.31}$$

From (3.30) and (3.31), we get

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) > \Phi(t, N, m, \underset{\sim}{\mathbf{N}_0}), \quad \forall t \geq m. \tag{3.32}$$

*Case 2:* In all other cases, by Lemma 3.2(b), there is an $i$, $1 \leq i \leq m$, such that $d(\underset{\sim}{\mathbf{N}}^{(i)}) < d(\underset{\sim}{\mathbf{N}})$. Moreover, it is also clear that, for the same $i$, we must have $d(\underset{\sim}{\mathbf{N}}_{(i)}) > 1$ (since otherwise $\underset{\sim}{\mathbf{N}}^{(i)}$ will be the same as $\underset{\sim}{\mathbf{N}}$), so that

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) > \Phi(t, N, m, \underset{\sim}{\mathbf{N}}^{(i)}), \quad \forall t \geq m. \tag{3.33}$$

If $d(\underset{\sim}{\mathbf{N}}^{(i)}) \leq 1$, then $\underset{\sim}{\mathbf{N}}^{(i)} = \underset{\sim}{\mathbf{N}_0}$ (up to permutation), so that we are done. If not, we can use Lemma 3.2 on the vector $\underset{\sim}{\mathbf{N}}^{(i)}$ to get another abundance vector, $\underset{\sim}{\mathbf{N}}^*$, say, with $d(\underset{\sim}{\mathbf{N}}^*) < d(\underset{\sim}{\mathbf{N}}^{(i)})$, which is accessible from $\underset{\sim}{\mathbf{N}}^{(i)}$ in at most two steps, so that by (3.29),

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}^{(i)}) \geq \Phi(t, N, m, \underset{\sim}{\mathbf{N}}^*), \quad \forall t \geq m. \tag{3.34}$$

However, since $d(\underset{\sim}{\mathbf{N}}) \leq N - m$, therefore, if we repeat this process of strictly reducing the $d$-value at successive stages, while the $\Phi$-values remain nonincreasing all the while, we will only need finitely many stages until we get an abundance vector with $d$-value $\leq 1$, that is, we reach the vector $\underset{\sim}{\mathbf{N}_0}$, thus proving finally that

$$\Phi(t, N, m, \underset{\sim}{\mathbf{N}}) > \Phi(t, N, m, \underset{\sim}{\mathbf{N}_0}), \quad \forall t \geq m, \tag{3.35}$$

and that completes the proof of the theorem. □

*Proof of Lemma* 3.2. (a) In this case, it is easy to see that $\underset{\sim}{\mathbf{N}}^{(i)}$, for any $i$, will have one coordinate equal to $N'$, one coordinate equal to $N' + 2$, and the other coordinates all equal to $N' + 1$, so that $d(\underset{\sim}{\mathbf{N}}^{(i)}) = d(\underset{\sim}{\mathbf{N}}) = 2$, for all $i$. However, it is equally easy to see that, if we now start with any one of these $\underset{\sim}{\mathbf{N}}^{(i)}$'s and if the $j$th coordinate of $\underset{\sim}{\mathbf{N}}^{(i)}$ equals $N' + 1$, then $(\underset{\sim}{\mathbf{N}}^{(i)})^{(j)} = (N' + 1, \ldots, N' + 1)$, so that $d((\underset{\sim}{\mathbf{N}}^{(i)})^{(j)}) = 0$.

(b) For the proof, we may clearly replace the vector $\underset{\sim}{\mathbf{N}}$ by any of its permutations. We assume, therefore, that our original vector $\underset{\sim}{\mathbf{N}}$ is such that $N_1 \geq \cdots \geq N_m$, so that $d(\underset{\sim}{\mathbf{N}}) = N_1 - N_m$. Let us denote $N_0$ to be the unique positive integer such that $N = mN_0 + k$ where $0 \leq k \leq m - 1$.

First of all, $mN_m \leq N < m(N_0 + 1)$ implies that $N_m \leq N_0$. But, if $N_m = N_0$, then $N - N_m = (m - 1)N_0 + k$ where $0 \leq k \leq m - 1$, so that $N_{0:m} = N_0$ or $N_0 + 1$ according as $k <$ or $= m - 1$. It is now easy to see that $d(\underset{\sim}{\mathbf{N}}^{(m)}) = 0$ or 1 according as $k =$ or $> 0$. In any case, if $d(\underset{\sim}{\mathbf{N}}) > 1$, then $d(\underset{\sim}{\mathbf{N}}^{(m)}) < d(\underset{\sim}{\mathbf{N}})$.

For the rest of the proof, therefore, we may assume that $N_m \leq N_0 - 1$. Next, note that $d(\underset{\sim}{\mathbf{N}}) > 1$ really means that $N_1 > N_m + 1$. In particular, $mN_1 > N \geq mN_0$, implying that $N_1 \geq N_0 + 1$.

Suppose now that $N_1 = N_0 + 1$. Then $N - N_1 = (m-1)N_0 + k - 1$, where $0 \le k \le m-1$. In case $k \ge 1$, we see that $N_{0:1} = N_0$, so that $d(\underset{\sim}{\mathbf{N}}^{(1)}) = 1 < d(\underset{\sim}{\mathbf{N}})$. On the other hand, if $k = 0$, then $N - N_1 = (m-1)(N_0 - 1) + (m-2)$, so that $N_{0:1} = N_0 - 1$ and $\underset{\sim}{\mathbf{N}}^{(1)} = (N_1, N_0, \ldots, N_0, N_0 - 1)$. If we had $N_m < N_0 - 1$ to start with, then $d(\underset{\sim}{\mathbf{N}}^{(1)}) = N_1 - (N_0 - 1) < N_1 - N_m = d(\underset{\sim}{\mathbf{N}})$, and we are done. If, on the other hand, we had $N_m = N_0 - 1$, that is, we had the situation where $N_1 = N_0 + 1$, $N_m = N_0 - 1$, and $N = mN_0$, then either $\exists i$ $(2 \le i \le m-1)$ such that $N_i = N_0$, in which case clearly $\underset{\sim}{\mathbf{N}}^{(i)} = (N_0, \ldots, N_0)$, so that $d(\underset{\sim}{\mathbf{N}}^{(i)}) = 0 < d(\underset{\sim}{\mathbf{N}})$ or else $m = 2r$ and $N_1 = \cdots = N_r = N_0 + 1$, $N_{r+1} = \cdots = N_{2r} = N_0 - 1$. But this last contingency is impossible because the vector $\underset{\sim}{\mathbf{N}}$ is assumed to be not of the form considered in (a).

We now continue with the proof assuming that $N_m \le N_0 - 1$ and also $N_1 \ge N_0 + 2$.

The case $m = 3$ is relatively easy to handle and hence we do that first. We have a vector $\underset{\sim}{\mathbf{N}} = (N_1, N_2, N_3)$, with $N_1 \ge N_2 \ge N_3$ and $d(\underset{\sim}{\mathbf{N}}) = N_1 - N_3 > 1$. In fact, we actually have $N_3 \le N_0 - 1$ and $N_1 \ge N_0 + 2$, where $N_0$ is the unique positive integer such that $N_1 + N_2 + N_3 = 3N_0 + k$ with $0 \le k \le 2$. Now, if $N_2 > N_3 + 1$, then $N_2 + N_3 \ge 2(N_3 + 1)$, so that $N_{0:1} \ge N_3 + 1$ and hence $d(\underset{\sim}{\mathbf{N}}^{(1)}) = N_1 - N_{0:1} < N_1 - N_3 = d(\underset{\sim}{\mathbf{N}})$. On the other hand, if $N_2 \le N_3 + 1$, then using the fact that $N_3 + 1 \le N_0 \le N_1 - 2$, one gets $N_1 + N_2 \le 2(N_1 - 1)$. In case equality holds here, we have $\underset{\sim}{\mathbf{N}}^{(3)} = (N_1 - 1, N_1 - 1, N_3)$, so that $d(\underset{\sim}{\mathbf{N}}^{(3)}) = N_1 - 1 - N_3 < d(\underset{\sim}{\mathbf{N}})$. On the other hand, $N_1 + N_2 < 2(N_1 - 1)$ will imply that $N_{0:1} \le N_1 - 2$ and therefore $d(\underset{\sim}{\mathbf{N}}^{(3)}) \le N_{0:1} + 1 - N_3 \le N_1 - 1 - N_3 < d(\underset{\sim}{\mathbf{N}})$, and the proof for the case $m = 3$ is complete.

From now on, we assume that $m \ge 4$. We have a vector $\underset{\sim}{\mathbf{N}} = (N_1, \ldots, N_m)$ with $N_1 \ge \cdots \ge N_m$ and $N_1 + \cdots + N_m = N$. We also have $N_m \le N_0 - 1$ and $N_1 \ge N_0 + 2$ where $N_0$ is the unique positive integer such that $N = mN_0 + k$ with $0 \le k \le m-1$. Let us first consider $N_{0:1} = \left[\frac{N - N_1}{m-1}\right]$. Clearly, $N_{0:1} \ge N_m$. If $N_{0:1} > N_m$, then clearly $d(\underset{\sim}{\mathbf{N}}^{(1)}) = N_1 - N_{0:1} < N_1 - N_m = d(\underset{\sim}{\mathbf{N}})$, and we are done. If, on the other hand, we have $N_{0:1} = N_m$, we then consider $N_{0:m} = \left[\frac{N - N_m}{m-1}\right]$ and proceed to show that $N_{0:m} + 1 < N_1$, which will, of course, imply that $d(\underset{\sim}{\mathbf{N}}^{(m)}) \le N_{0:m} + 1 - N_m < N_1 - N_m = d(\underset{\sim}{\mathbf{N}})$. For this, note first that $N_{0:1} = N_m$ implies that $N_2 + \cdots + N_m = (m-1)N_m + l$, where $0 \le l \le m-2$, so that, $N_2 + \cdots + N_{m-1} = (m-2)N_m + l < (m-2)N_m + (m-1)$. Therefore, $N - N_m < N_1 + (m-2)N_m + (m-1)$. Using the fact that $N_m \le N_0 - 1 \le (N_1 - 2) - 1$, we get $N - N_m < N_1 + (m-2)(N_1 - 3) + (m-1) \le (m-1)(N_1 - 1) - (m-4) \le (m-1)(N_1 - 1)$, since $m \ge 4$. Thus, we have $N - N_m < (m-1)(N_1 - 1)$, which gives $N_{0:m} < N_1 - 1$, or, equivalently, $N_{0:m} + 1 < N_1$, as was to be proved. □

**Remark 3.2.** It is well known that an arbitrary $\underset{\sim}{\mathbf{p}}$-vector is majorized by $\underset{\sim}{\mathbf{p}}_0$ and hence the results on Schur concave functions will have direct application, provided we can establish Schur concavity of $P[T > t \mid m, \underset{\sim}{\mathbf{p}}]$ as a function of $\underset{\sim}{\mathbf{p}}$, which may be of independent interest. Referring to Lemma A.2 in Marshall and Olkin (1979), it amounts to establishing "local improvement by averaging over two neighboring allocations." This may once again require an induction argument, similar to the one used by us. However, our argument uses induction to directly hit upon the solution to the minimization problem, rather than going via Schur concavity.

**Remark 3.3.** Our result for the finite-population case asserts that under SRSWR with population size $N = ms$, $P[T > t \mid m, \underset{\sim}{\mathbf{p}}] \geq P[T > t \mid m, \underset{\sim}{\mathbf{p}_0}]$, with strict inequality holding unless $\underset{\sim}{\mathbf{p}} = \underset{\sim}{\mathbf{p}_0}$, and this holds for all $s = 1, 2, \ldots$. One may possibly try to derive the result for the infinite-population case from this by using a limiting argument with $s \uparrow \infty$. However, it presents a number of hurdles. Firstly, not all probability vectors $\underset{\sim}{\mathbf{p}}$ are admissible in the finite-population case. Secondly, as $s$ changes, the probability space also changes, and it does not seem obvious that the probabilities for finite population will, in the limit, give probability for the infinite-population case. Having explicit expressions for such probabilities would perhaps have helped, but we do *not* have such expressions at any stage, unless $m$ is really small! Of course, we have given a direct proof for the infinite-population case, which is fairly simple and perhaps interesting in its own right.



## ACKNOWLEDGMENTS

We are thankful to Professors Anil P. Gore and S. A. Paranjape of the Department of Statistics, Pune University, for suggesting this problem to one of us and for taking interest in this study. We would also like to thank the associate editor and the anonymous referee for seeking some clarifications that resulted in the Remarks 3.2 and 3.3, given above.

## REFERENCES

Apostol, T. M. (1974). *Mathematical Analysis*, New York: Addison-Wesley.

Basu, D. (1958). On Sampling With and Without Replacement, *Sankhyā Series A* 20: 287–294.

Feller, W. (1967). *An Introduction to Probability Theory and Its Applications*, Vol. 1, New York: Wiley.

Gore, A. P. and Paranjape, S. A. (1997). Effort Needed to Measure Biodiversity, *International Journal of Ecology and Environmental Sciences* 23: 173–183.

Gore, A. P. and Paranjape, S. A. (2001). *A Course in Mathematical and Statistical Ecology*, Budapest: Kluwer.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization with Applications*, New York: Academic Press.

Patil, G. P. and Taillie, C. (1982). Diversity as a Concept and its Measurement, *Journal of American Statistical Association* 77: 548–561.

Rao, T. J., Sinha, B. K., and Srivenkataramana, T. (2003). On Order Relations Between Selection and Inclusion, *Journal of Applied Statistical Science* 12: 67–73.