Sequential Search Algorithm for Estimation of the Number of Classes in a Given Population

Klass, Michael Jay ; Nowicki, Krzysztof

2016

Working Papers in Statistics
No 2016:1

Department of Statistics
School of Economics and Management
Lund University

Sequential Search Algorithm for Estimation of the Number of Classes in a Given Population

MICHAEL J. KLASS, UNIVERSITY OF CALIFORNIA, BERKELEY

KRZYSZTOF NOWICKI, LUND UNIVERSITY

Sequential Search Algorithm for Estimation of the Number of Classes in a Given Population

Michael J. Klass
University of California, Berkeley
and
Krzysztof Nowicki
Lund University, Sweden

Abstract

Let N be the number of classes in a population to be estimated. Fix any preassigned error probability $0 < \epsilon < \exp(-2)$ (roughly). We present a sequential search algorithm to estimate the exact value of N, with an error probability of at most ϵ, regardless of the value of N.
Key words and phrases: Unobserved species, estimation of population size, sequential estimation procedure, error probability

1 Introduction

Historically many people consider the classic "species problem", that of estimating the number N of categories in a given population based on a random sample, first introduced by Fisher *et al* (1943) and widely studied in ecology and later extended to many other applications: see, for instance, Thisted and Efron (1987) and Mao and Lindsay (2007). Bunge and Fitzpatrick (1993) provide a review of various statistical methods to estimate the number of unseen species.

The estimation of the number of unseen species is a question closely related to the problem of estimating the expected number of new species that will be seen if we take an additional sample of any given size, see Good and Toulmin (1956), Efron and Thisted (1976), Boneh *et al* (1998).

Several authors proposed estimation methods of the number of classes utilizing sequential random sampling with various stopping rules, see Goodman (1953), Samuel (1968, 1969), Holst (1971) and Nayak and Kundu (2007). In the algorithm we propose the stopping rule is connected to a preassigned error probability $P(N \neq \hat{N})$ where $\hat{N}$ is our estimate of N.

The problem addressed in this paper can be formalized as follows. Let N be a fixed but unknown positive integer. Let $X_1, X_2, ...$ be i.i.d. random variables which take values in $\{1, 2, ..., N\}$, each of the N outcomes being equally likely. Fix any preassigned error probability $0 < \epsilon < \exp(-2)$ (roughly). We want to estimate the exact value of N, with an error probability of at most ϵ, regardless of the value of N.

2 The search algorithm

To teach ourselves how to proceed, suppose we begin by asking: When might one decide it would no longer be advantageous to continue searching/ sampling (gathering observations) for items yet unseen?

Given that j different objects have already been recorded, let W_{j+1} denote the random waiting time describing the number of additional observations that happen to be needed until the $(j+1)^{\text{st}}$ item surfaced, with $W_{j+1} = \infty$ if there are only j items. Formally, define $T_1 = W_1 = 1$ and, having defined W_j for $1 \leq j \leq k$, let $T_j = W_1 + ... + W_j$ and define

$$W_{j+1} = \begin{cases} 1^{\text{st}} \quad i \geq 1 : X_{i+T_j} \notin \{X_1, .., X_{T_j}\} \\ \infty \text{ if no such } i \text{ exists.} \end{cases} \tag{1}$$

and $T_{j+1} = W_1 + ... + W_{j+1}$.

Suppose $N = n$ for some integer $n > 0$. For each n we use notation $W_{n,j+1}$ and $T_{n,j+1}$. Imagine that we have found j different objects already and have conducted another s observations without finding anything new. What information does this provide?

$$P(W_{n,j+1} > s) = (\frac{j}{n})^s, \quad \text{for } 1 \leq j \leq n-1. \tag{2}$$

Suppose our search has currently found j objects and then W_{j+1} exceeds $\lceil A(j+k) \rceil$ where k is to be determined. Let

$$t = 1^{\text{st}} \quad j \geq 1 : W_{j+1} > \lceil A(j+k) \rceil \tag{3}$$

We stop at such time t and declare that $\hat{N}$, our best guess as to the value of N, is t. At that point we have conducted

$$Q \equiv 1 + W_2 + \cdot + W_t + \lceil A(t+k) \rceil \tag{4}$$

searches. What is the maximum error probability incurred by this rule over all possible values of $N \geq 1$?
Let $\hat{N}$ be the integer which our procedure guesses. If $n = 1$,$\hat{N}$ is always 1. For $N = n \geq 2$

$$P(\hat{N} \neq n) = P(\bigcup_{j=1}^{n-1} \{W_{n,j+1} > \lceil A(j+k) \rceil \}). \tag{5}$$

To upper-bound this expression we introduce the following lemma.

Lemma 2.1 *Let E_{j+1}, $1 \leq j \leq n-1$ be a set of independent events and let E^*_{j+1}, $1 \leq j \leq n-1$ be a set of independent events such that $P(E_{j+1}) \leq P(E^*_{j+1})$ for $1 \leq j \leq n-1$. Then*

$$P(\bigcup_{j=1}^{n-1} E_{j+1})) \leq \sum_{j=1}^{n-1} P(E^*_{j+1}) - P(E^*_n) P(E^*_{n-1}). \tag{6}$$

Proof: First observe that

$$\begin{aligned} P(\bigcup_{j=1}^{n-1} E_{j+1}) &= 1 - P(\bigcap_{j=1}^{n-1} E^c_{j+1}) \\ &= 1 - \bigcap_{j=1}^{n-1} P(E^c_{j+1}) \quad \text{(by independence)} \\ &\leq 1 - \bigcap_{j=1}^{n-1} P((E^*_{j+1})^c) \quad \text{(by the assumption)} \\ &= P(\bigcup_{j=1}^{n-1} E^*_{j+1}). \end{aligned} \tag{7}$$

Moreover, by Boole's inequality

$$P(\bigcup_{j=1}^{n-1} E^*_{j+1}) \leq P(E^*_n \cup E^*_{n-1}) + \sum_{j=1}^{n-2} P(E^*_{j+1}) \tag{8}$$

and by independence $P(E^*_n \cup E^*_{n-1}) = P(E^*_n) + P(E^*_{n-1}) - P(E^*_n)P(E^*_{n-1})$ from which (6) follows.

$\square$

For $1 \leq j \leq n-1$, $A > 0$ and fixed $k \geq 1$ let

$$E_{n,j+1} = \{W_{n,j+1} > \lceil A(j+k) \rceil\} \tag{9}$$

and

$$P(E^*_{n,j+1}) = (\frac{j}{n})^{A(j+k)}. \tag{10}$$

Clearly, $P(E_{n,j+1}) \leq P(E^*_{n,j})$. Using Lemma 2.1, (5) can be upper-bounded by

$$P(\bigcup_{j=1}^{n-1} E_{n,j+1}) \leq P(\bigcup_{j=1}^{n-1} E^*_{n,j+1}) \leq \sum_{j=1}^{n-1} P(E^*_{n,j+1}) - P(E^*_{n,n})P(E^*_{n,n-1}). \tag{11}$$

Next we introduce upper-bounds for terms in (11).

Lemma 2.2 *For $1 \leq j \leq 2k$ and $n \geq j+1$*

$$P(E^*_{n,n-j+1}) \leq \exp(-Aj). \tag{12}$$

Proof: For $k \geq 1$, $1 \leq j \leq n-1$, $A > 0$,

$$\begin{aligned} P(W_{n,j+1} > \lceil A(j+k) \rceil) &= \Big(\frac{j}{n}\Big)^{\lceil A(j+k) \rceil} \quad \text{(by definition of } W_{n,j+1}) \\ &\leq \Big(\frac{j}{n}\Big)^{A(j+k)} \\ &= \exp(-A(j+k)\ln(\frac{n}{j})) \equiv P(E^*_{n,j+1}). \end{aligned} \tag{13}$$

Replacing j by $n-j$ and n by x, for $x \geq j+1$ let

$$f(x) = -(x-j+k)\ln(\frac{x}{x-j}). \tag{14}$$

Notice that $P(E^*_{n,n-j+1}) = \exp(Af(n))$. Then

$$\begin{aligned} f'(x) &= -\ln(\frac{x}{x-j}) - (x-j+k)(\frac{1}{x} - \frac{1}{x-j}) \\ &= -\ln(\frac{x}{x-j}) - \frac{k-j}{x} + \frac{k}{x-j} \end{aligned} \tag{15}$$

and

$$\begin{aligned} f''(x) &= -\frac{1}{x} + \frac{1}{x-j} + \frac{k-j}{x^2} - \frac{k}{(x-j)^2} \\ &= \frac{j}{x(x-j)} + \frac{(k-j)(x-j)^2 - kx^2}{x^2(x-j)^2} \\ &= \frac{jx^2 - j^2x - jx^2 - 2j(k-j)x + j^2(k-j)}{x^2(x-j)^2} \\ &= \frac{j(j-2k)x + j^2(k-j)}{x^2(x-j)^2} \\ &\leq 0 \ \big(\text{for } k \leq j \leq 2k \text{ or if } 1 \leq j \leq k-1 \text{ since } (2k-j)x \geq (k-j)j\big). \end{aligned} \tag{16}$$

Therefore $f(x)$ is concave. Notice that $\lim_{x\to\infty} f'(x) \to 0$. Therefore $f'(x) > 0$ for all $x > j+1$. Consequently

$$\begin{aligned} \sup_{x \geq j+1} f(x) &= \lim_{x\to\infty} f(x). \\ &= \lim_{x\to\infty} x \ln(1 - \frac{j}{x}) \\ &= -j. \end{aligned} \tag{17}$$

Hence

$$\sup_{n \geq j+1} P(E^*_{n,n-j+1}) = \lim_{n\to\infty} P(E^*_{n,n-j+1}) = \lim_{n\to\infty} \exp(Af(n)) = \exp(-Aj). \tag{18}$$

□

Theorem 2.1 *For all $n \geq 2$, $A \geq 2$, $k = 4$*

$$P(\bigcup_{j=1}^{n-1} E^*_{n,j+1}) < \exp(-A) + \exp(-2A) + \exp(-3A). \tag{19}$$

Proof: We treat various n separately.

For $n = 1$ the error probability is zero.

Case 1: $2 \leq n \leq 4$ Inequality (12) combined with Boole's inequality yields

$$P(\bigcup_{j=1}^{n-1} E^*_{n,j+1}) \leq \exp(-A) + \exp(-2A) + \exp(-3A) \tag{20}$$

Cases $n \geq 5$
Invoking (11) and (12), for $n \geq 5$

$$\begin{aligned} P(\bigcup_{j=1}^{n-1} E^*_{n,j+1}) &< \exp(-A) + \exp(-2A) + \exp(-3A) - P(E^*_{n,n})P(E^*_{n,n-1}) \\ &+ \sum_{j=1}^{n-4} P(E^*_{n,j+1}). \end{aligned} \tag{21}$$

Case $n = 5$

Since

$$P(E^*_{5,2}) = \exp(-5A\ln(5)), \tag{22}$$

$$\begin{aligned} P(E^*_{5,5})P(E^*_{5,4}) &= \exp(-(\Gamma_{5,5} + \Gamma_{5,4})A) \\ &> \exp(-5.4A) \text{ (by (36) and (37) below)} \\ &> \exp(-5A\ln(5)) \\ &= P(E^*_{5,2}) \end{aligned} \tag{23}$$

so by (21) Theorem 1.1 holds for $n = 5$.

Case $n = 6$

Applying (21) for $n = 6$ we need to verify the inequality

$$P(E^*_{6,2}) + P(E_{6,3}) \leq P(E^*_{6,6})P(E^*_{6,5}). \tag{24}$$

Since $\Gamma_{n,n}$ and $\Gamma_{n,n-1}$ decrease in n,

$$\begin{aligned} P(E^*_{6,6})P(E^*_{6,5}) &= \exp(-A(\Gamma_{6,6}+\Gamma_{6,5})) \\ &> \exp(-A(\Gamma_{5,5}+\Gamma_{5,4}) \\ &> \exp(-5.4A) \\ &> \exp(-5A\ln(6)) + \exp(-6A\ln(3)) \\ &= P(E^*_{6,2})P(E^*_{6,3}), \end{aligned} \tag{25}$$

which confirms Theorem 1.1 for $n = 6$.

Case $n = 7$

Since $\Gamma_{n,n}$ and $\Gamma_{n,n-1}$ decrease in n,

$$\begin{aligned} P(E^*_{7,7})P(E^*_{7,6}) &> P(E^*_{5,5})P(E^*_{5,4}) > \exp(-5.4A) \\ &> \exp(-5A\ln(7)) + \exp(-6A\ln(\frac{7}{2})) + \exp(-7A\ln(\frac{7}{3})) \\ &= P(E^*_{7,2}) + P(E^*_{7,3}) + P(E^*_{7,4}) \text{ (for } A \geq 2), \end{aligned} \tag{26}$$

whence Theorem 1.1 holds for $n = 7$.

Case $n \geq 8$

We begin by considering $\sum_{j=1}^{n-4} P(E^*_{n,j+1})$. Splitting this sum into three parts,

$$\sum_{j=1}^{n-4} P(E^*_{n,j+1}) = P(E^*_{n,2}) + \sum_{2\leq j<\frac{n}{e}} P(E^*_{n,j+1}) + \sum_{\frac{n}{e}<j\leq n-4} P(E^*_{n,j+1}). \tag{27}$$

We will treat each of the sums separately. First,

$$\begin{aligned} &\sum_{2\leq j<\frac{n}{e}} P(E^*_{n,j+1}) = \sum_{j=2}^{\lceil\frac{n}{e}\rceil} \exp(-A(j+4)\ln(\frac{n}{j})) \\ &\leq \frac{\exp(-6A)}{1-\exp(-A)} \text{ (since } \frac{n}{j} \geq 1). \end{aligned} \tag{28}$$

Second,

$$\begin{aligned}
&\sum_{\frac{n}{e}<j\le n-4} P(E^*_{n,j+1}) = \sum_{\frac{n}{e}<j\le n-4} \exp(-A(j+4)\ln(\frac{n}{j}))\\
&= \sum_{\frac{n}{e}<n-j\le n-4} \exp(-A(n-j+4)\ln(\frac{n}{n-j}))\\
&\le \sum_{1\le i<n-\lceil\frac{n}{e}\rceil-3} \frac{\exp(2\ln(i)-A(n+1-i)\ln(\frac{n}{n-i-3)}))}{i^2} \qquad (29)\\
&\le \max_{1\le i\le n-\lceil\frac{n}{e}\rceil-3} \exp(2\ln(i)-A(n+1-i)\ln(\frac{n}{n-i-3}))\sum_{j=1}^{\infty}\frac{1}{j^2}\\
&\le \frac{\pi^2}{6}\exp\Big(-An\ln(\frac{n}{n-4})\Big) \text{ (by Lemma 3.2 if } k=4).
\end{aligned}$$

Hence,

$$\begin{aligned}
\sum_{j=1}^{n-4} P(E^*_{n,j+1}) &\le \exp(-5A\ln 8) + \frac{\exp(-6A)}{1-\exp(-A)} + \frac{\pi^2}{6}\exp\Big(-An\ln(\frac{n}{n-4})\Big)\\
&= \exp(-5A\ln 8) + \frac{\exp(-6A)}{1-\exp(-A)} + \frac{\pi^2}{6}P(E^*_{n,n-3}).
\end{aligned} \qquad (30)$$

Next we show

$$\exp(-5A\ln(8)) + \frac{\exp(-6A)}{1-\exp(-A)} \le P(E^*_{n,n})P(E^*_{n,n-1}) - \frac{\pi^2}{6}P(E^*_{n,n-3}). \qquad (31)$$

For $n \ge 8$, $A \ge 2$

$$\begin{aligned}
&P(E^*_{n,n})P(E^*_{n,n-1}) - \frac{\pi^2}{6}P(E^*_{n,n-3}) \ge \exp(-4.5A) - \frac{\pi^2}{6}\exp(-5A)\\
&= \exp(-4.5A)(1-\frac{\pi^2}{6}\exp(-0.5A) \ge \exp(-4.5A)(1-\frac{\pi^2}{6}\exp(-1)) \qquad (32)\\
&> \exp(-5A\ln(8)) + \frac{\exp(-6A)}{1-\exp(-A)}
\end{aligned}$$

whence (31) holds for $n \ge 8$ and $A \ge 2$.

□

Remark 2.1 *Notice that for all* $j \geq 1$

$$\lim_{n\to\infty} P(E_{n,n-j+1}) = \lim_{n\to\infty} \exp(-\lceil A(n-j+4)\rceil \ln(\frac{n}{n-j})) = \exp(-Aj). \quad (33)$$

Hence as the number $N = n$ *of objects to be found tends to infinity the probability that we fail to guess their exact number tends to*

$$\begin{aligned} \lim_{n\to\infty} P(\bigcup_{j=1}^{n-1} E_{n,j+1}) &= \lim_{n\to\infty}(1 - P(\bigcap_{j=1}^{n-1} E^c_{n,j+1})) \\ &= \lim_{n\to\infty}(1 - \prod_{j=1}^{n-1}(1 - P(E_{n,n-j+1})) \quad (34) \\ &= 1 - \prod_{j=1}^{\infty}(1 - \exp(-Aj)). \end{aligned}$$

An alternative expression for this limit may be obtained by writing

$$\begin{aligned} \prod_{j=1}^{\infty}(1 - \exp(-Aj)) &= \exp(\sum_{j=1}^{\infty} \ln(1 - \exp(-Aj))) \\ &= \exp(-\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{\exp(-jkA)}{k}) \quad (35) \\ &= \exp(-\sum_{k=1}^{\infty} \frac{\exp(-kA)}{k(1-\exp(-kA))}). \end{aligned}$$

3 Appendix

We lower- and upper-bound $P(E^*_{n,k})$ for $k = n, n-1$, and $n-3$.

Lemma 3.1 *Let* $\Gamma_{n,k} = -\frac{1}{A}\ln(P(E^*_{n,k}))$. *For* $A \geq 2$ *and* $n \geq 2$,

$$\exp(-A(1 + \frac{7}{2n} + \frac{11}{6n^2} + \frac{5}{4n^2(n-1)})) \leq P(E^*_{n,n}) \leq \exp(-A(1 + \frac{7}{2n})) \quad (36)$$

$$P(-A(2 + \frac{6}{n} + \frac{20}{3n^2} + \frac{28}{3n^2(n-2)})) \leq P(E^*_{n,n-1}) \leq \exp(-A(2 + \frac{6}{n})) \quad (37)$$

and

$$\exp(-A(4 + \frac{8}{n} + \frac{64}{3n^2} + \frac{64}{n^2(n-4)})) \leq P(E^*_{n,n-3}) \leq \exp(-A(4 + \frac{8}{n})). \quad (38)$$

Proof: We derive upper- and lower- bounds of $P(E^*_{n,k})$ by bounding $\Gamma_{n,k}$ for $k = n$, $k = n-1$ and $k = n-3$.
First,

$$\begin{aligned}\Gamma_{n,n} &= (n+3)\ln(\frac{n}{n-1}) = -(n+3)\ln(1-\frac{1}{n}) \\ &= \sum_{j=1}^{\infty}\frac{1}{jn^{j-1}} + \sum_{j=1}^{\infty}\frac{3}{jn^j} \\ &= 1 + \sum_{j=1}^{\infty}(\frac{1}{j+1}+\frac{3}{j})\frac{1}{n^j}\end{aligned} \tag{39}$$

Clearly,

$$\Gamma_{n,n} \geq 1 + \frac{7}{2n} \tag{40}$$

and

$$\begin{aligned}\Gamma_{n,n} &= 1 + \frac{7}{2n} + \frac{11}{6n^2} + \sum_{j=3}^{\infty}\frac{4j+3}{(j+1)j}\frac{1}{n^j} \\ &\leq 1 + \frac{7}{2n} + \frac{11}{6n^2} + \frac{5}{4}\sum_{j=3}^{\infty}\frac{1}{n^j} \\ &= 1 + \frac{7}{2n} + \frac{11}{6n^2} + \frac{5}{4n^2(n-1)}\end{aligned} \tag{41}$$

which gives (36). Second,

$$\begin{aligned}\Gamma_{n,n-1} &= (n+2)\ln(\frac{n}{n-2}) = -(n+2)\ln(1-\frac{2}{n}) \\ &= 2 + \sum_{j=1}^{\infty}\frac{2^j}{n^j}(\frac{2}{j+1}+\frac{2}{j})\end{aligned} \tag{42}$$

which gives (37) by calculations similar to those used in (40) and (41). Third, we lower-bound $P(E^*_{n,n-3}) = \exp(-A\Gamma_{n,n-3})$ by upper-bounding

$$\begin{aligned}
\Gamma_{n,n-3} &= n\ln(\frac{n}{n-4}) = -n\ln(1-\frac{4}{n}) \\
&= 4 + \sum_{j=1}^{\infty} \frac{4^{j+1}}{(j+1)n^j} \\
&= 4 + \frac{8}{n} + \frac{64}{3n^2} + \sum_{j=3}^{\infty} \frac{4}{j+1}(\frac{4}{n})^j \\
&\leq 4 + \frac{8}{n} + \frac{64}{3n^2} + \sum_{j=3}^{\infty} (\frac{4}{n})^j \\
&= 4 + \frac{8}{n} + \frac{64}{3n^2} + \frac{64}{n^2(n-4)}
\end{aligned} \tag{43}$$

which gives (38).

□

Lemma 3.2 *Let $A \geq 2$, $k \geq 1$, $B = n-k+1$, $n > k$ and*

$$g(x) = 2\ln(B-x) - A(x+k)\ln(\frac{n}{x}) \tag{44}$$

Then

$$\sup_{\frac{n}{e} \leq x \leq B-1} g(x) = g(B-1) = -An\ln(\frac{n}{n-k}). \tag{45}$$

Proof: Toward this end,

$$\begin{aligned}
g'(x) &= -\frac{2}{B-x} - A\ln(\frac{n}{x}) + A + \frac{Ak}{x} \\
&= -\frac{2}{B-x} + A\ln(\frac{ex}{n}) + \frac{Ak}{x} \\
&\geq -\frac{2}{B-x} + 2\ln(\frac{ex}{n}) + \frac{2k}{x} \\
&\equiv 2h(x).
\end{aligned} \tag{46}$$

For $x = \frac{n}{e}$ we have $\ln(\frac{ex}{n}) = 0$. Hence

$$g'(x) \geq \frac{2}{x(B-x)}(Bk - (k+1)x) \geq 0 \tag{47}$$

for $\frac{n}{e} \le x \le \frac{Bk}{k+1}$. We need to prove that $g'(x) \ge 0$ for $\frac{Bk}{k+1} \le x \le B-1$. Consider

$$h'(x) = -\frac{1}{(B-x)^2} + \frac{1}{x} - \frac{k}{x^2} \tag{48}$$

For $\frac{Bk}{k+1} \le x < B$

$$\begin{aligned} h''(x) &= -\frac{2}{(B-x)^3} - \frac{1}{x^2} + \frac{2k}{x^3} \\ &< -\frac{2(k+1)^3}{B^3} + \frac{2(k+1)^3}{B^3k^2} \\ &\le 0 \quad \text{for } k \ge 1. \end{aligned} \tag{49}$$

Hence $h(x)$ is concave on $\frac{Bk}{k+1} \le x \le B-1$. Therefore

$$\inf_{\frac{Bk}{k+1} \le x \le B-1} = \min\{h(\frac{Bk}{k+1}), h(B-1)\} \ge 0 \text{ iff } h(B-1) \ge 0 \tag{50}$$

$$h(B-1) = -1 + \ln(\frac{e(n-k)}{n}) + \frac{k}{n-k} \equiv q(n) \tag{51}$$

$$\begin{aligned} q'(n) &= \frac{1}{n-k} - \frac{1}{n} - \frac{k}{(n-k)^2} \\ &= \frac{k}{n(n-k)} - \frac{k}{(n-k)^2} \\ &= \frac{k(n-k) - nk}{n(n-k)^2} < 0. \end{aligned} \tag{52}$$

Thus $q(n) \searrow \lim_{n\to\infty} q(n) = 0$ whence $h(B-1) > 0$ and so $g(B-1) = \sup_{\frac{n}{e} \le x \le B-1} g(x)$.

□

References

[1] Boneh, S., Boneh, A and Caron, R.J. Estimating the prediction function and the number of unseen species in a sampling with replacement, Journal of the American Statistical Association, **93(441)**, (1998), 372—379.

[2] Bunge, J and Fitzpatrick M. Estimating the number of species: a review, Journal of the American Statistical Association, **88(421)** (1993), 364—373.

[3] Efron, B and Thisted, R. Estimating the number of unseen species: How many words did Shakespeare know?, Biometrika **63(3)** (1976), 435—447.

[4] Fisher, R.A., Corbet, A.S. and Williams, C.B. The relation between the number of species and the number of individuals in a random sample of an animal population, Journal of Animal Ecology, **12** (1943), 42–58.

[5] Good, I.J. and Toulmin, G.H. The number of new species and the increase in population coverage, when a sample is increased, Biometrika, **43** (1956), 45–63.

[6] Goodman, L.A. Sequential sampling tagging for population size problems, the Annals Mathematical Statistics, **24** (1953), 56–69.

[7] Holst, L. Some asymptotic results for incomplete multinomial or Poisson samples, Scandinavian Journal of Statistics, **8** (1981), 243–246.

[8] Mao, C.X. and Lindsay, B.G. Estimating the number of classes, the Annals of Statistics, **35(2)** (2007), 917–930.

[9] Nayak, T.P. and Kundu, S. Comparison of stopping rules in sequential estimation of the number of classes in a population, Sequential Analysis, **26** (2007), 367–381.

[10] Samuel, E. Sequential maximum likelihood estimation of the size of a population, the Annals of Mathematical Statistics, **39(3)** (1968), 1057–1068.

[11] Samuel, E. Comparison of sequential rules of estimation of the size of a poulation, Biometrics, **25** (1969), 517–527.

[12] Thisted, R. and Efron, B. Did Sheakespeare write a newly-discover poem?, Biometrika **74(3)** (1987), 445—455.

http://journals.lub.lu.se/stat

LUND UNIVERSITY
School of Economics and Management

Working Papers in Statistics 2016
LUND UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT
Department of Statistics
Box 743
220 07 Lund, Sweden