FINDING THE SIZE OF A FINITE POPULATION

By D. A. Darling[1] and Herbert Robbins[2]

University of California, Berkeley

1. Introduction. There are fixed sample size methods for estimating the unknown size N of a finite population by tagging the elements of a first sample and then counting the number of tagged elements of a second sample [1, p. 43]. Less well known are sequential methods [2], which have the advantage that the total sample size automatically adjusts itself to the unknown N to assure a desired accuracy of the estimate. All these methods only provide estimates for which the *relative* error is likely to be small. Suppose, however, that we want P_N (estimate $= N$) $\geqq \alpha = .99$, say, no matter what the value $N = 1, 2, \cdots$. How can this be done?

If we take as our estimate of N the number of distinct elements actually observed, the problem is one of finding a stopping rule such that the probability of having observed all N elements by the time we stop is $\geqq \alpha$ for all $N \geqq 1$. A concept of asymptotic efficiency may be introduced by comparing as $N \to \infty$ the expected sample size for any such rule with the fixed sample size necessary to observe all N elements with probability α. We give a procedure which is asymptotically efficient in this sense. We do not discuss the problem of finding a procedure which minimizes the Bayes expectation of the sample size for a given prior distribution of N.

Before going on, the reader is invited to consider the following problem. Sample one element at a time with replacement, tagging each element observed so that it can be recognized if it appears again. Choose some large integer M, and stop sampling when for the first time a run of M consecutive tagged elements occurs; estimate N to be the total number of distinct elements observed. Can we choose M so large that P_N (estimate $= N$) $\geqq .99$ for all $N = 1, 2, \cdots$? (Answer at end of paper.)

2. A procedure based on individual waiting times. An urn contains N white balls ($N = 1, 2, \cdots$). We repeatedly draw a ball at random, observe its color, and replace it by a black ball. Eventually all N white balls will have been drawn and the urn will contain only black balls. The probability that this will occur at or before the nth draw is [1, pp. 92–93]

$$P_{N,n} = \sum_{i=0}^{N} (-1)^i \binom{N}{i} (1 - i/N)^n, \tag{1}$$

and if $N, n \to \infty$ so that $Ne^{-n/N} \to \lambda$, $0 < \lambda < \infty$, then

$$P_{N,n} \to e^{-\lambda}. \tag{2}$$

For any fixed $0 < \alpha < 1$ the smallest $n = n(N, \alpha)$ such that $P_{N,n} \geqq \alpha$ can be found from (1) by trial and error. For large N this is tedious, but it follows from

Received 7 December 1966.

[1] Supported in part by National Science Foundation Grant GP 6549.

[2] Visiting Miller Research Professor.

(2) that $n = N \log N + cN + o(N)$, where c is determined by the equation $e^{-e^{-c}} = \alpha$. Equivalently, if we denote by Y_N the number of the draw on which the last (Nth) white ball is drawn, then for any $-\infty < c < \infty$,

$$P_N((Y_N - N \log N)/N \leqq c) \to e^{-e^{-c}} \quad \text{as} \quad N \to \infty. \tag{3}$$

Suppose now that N is unknown and that we wish to find a rule for deciding when to stop drawing such that for a given $0 < \alpha < 1$,

$$P_N(\text{all } N \text{ white balls drawn by the time we stop}) \geqq \alpha \quad (N \geqq 1). \tag{4}$$

No fixed sample size will do this, so we must look for a sequential procedure.

Let Y_n denote the number of the draw on which the nth white ball appears ($n = 1, \cdots, N$); thus $Y_1 \equiv 1$, and we put $Y_0 \equiv 0$, $Y_{N+1} \equiv \infty$ by convention. Define the waiting times

$$X_n = Y_{n+1} - Y_n \qquad (n = 0, \cdots, N), \tag{5}$$

so $X_0 \equiv 1$ and $X_N \equiv \infty$. The random variables $X_1, \cdots, X_{N-1}$ are independent, with the geometric distributions

$$P_N(X_n > j) = (n/N)^j \qquad (j = 0, 1, \cdots), \tag{6}$$

and

$$E_N(X_n) = N/(N - n), \qquad \mathrm{Var}_N(X_n) = nN/(N - n)^2. \tag{7}$$

Let (b_n) be any sequence of positive integers, and define B_n to be the event $X_n > b_n$ ($n = 1, \cdots, N$). Since $X_N \equiv \infty$, B_N is certain. Let J = first $n \geqq 1$ such that B_n occurs; then $1 \leqq J \leqq N$. Suppose we agree to stop drawing as soon as B_J occurs. All N white balls will have been drawn by the time we stop if and only if $J = N$. Hence the left hand side of (4) equals

$$P_N(J = N) = P_N(\textstyle\bigcap_1^{N-1} B_n') = \prod_1^{N-1} P(X_n \leqq b_n) = \prod_{n=1}^{N-1} \{1 - (n/N)^{b_n}\}. \tag{8}$$

It is clear that we can satisfy (4) by choosing the b_n properly. One way is the following. Define b_1^* = smallest integer b_1 such that

$$1 - (\tfrac{1}{2})b_1 \geqq \alpha.$$

Then (4) holds for $N = 2$ (and for $N = 1$ no matter what the sequence (b_n)). If $b_1^*, \cdots, b_{j-1}^*$ have been defined, set b_j^* = smallest integer b_j such that

$$\prod_{n=1}^{j-1} \{1 - (n/(j+1))^{b_n^*}\} \cdot \{1 - (j/(j+1))^{b_j}\} \geqq \alpha.$$

The sequence (b_n^*) gives a "step-wise minimal" solution of the inequalities

$$P_N(J = N) = \prod_{n=1}^{N-1} \{1 - (n/N)^{b_n}\} \geqq \alpha \qquad (N = 2, 3, \cdots). \tag{9}$$

(It is not *uniformly* minimal, since if (b_n) satisfies (9) then we can increase $b_1, \cdots, b_{j-1}$ sufficiently so that a smaller b_j will still satisfy (9).) Although the b_n^* are not given by an explicit formula they can be computed numerically for any given α.

Instead of doing this we shall find a lower bound for the left hand side of (9)

for the particular sequence

$$b_n = [cn] + 1 \qquad (n = 1, 2, \cdots), \tag{10}$$

where c is a suitable positive constant. We shall show that if c is chosen large enough then (9) holds, and shall find the limit of $P_N(J = N)$ as $N \to \infty$.

For the sequence (10) let β be any constant $0 < \beta < 1$. Then

$$P_N(J = N) = \prod_{1 \leqq n \leqq \beta N} \{1 - (n/N)^{b_n}\} \cdot \prod_{\beta N < n \leqq N-1} \{1 - (n/N)^{b_n}\} = Q_N \cdot R_N, \tag{11}$$

where

$$Q_N \geqq \prod_{1 \leqq n \leqq \beta N} (1 - \beta^{cn}) \geqq \prod_{n=1}^{\infty} (1 - \beta^{cn}), \tag{12}$$

and since $\log (1 - x) \leqq -x$,

$$\begin{aligned} R_N &= \prod_{1 \leqq i < (1-\beta)N} \{1 - (1 - i/N)^{[c(N-i)]+1}\} \\ &\geqq \prod_{1 \leqq i < (1-\beta)N} \{1 - (1 - i/N)^{c(N-i)}\} \\ &= \prod_{1 \leqq i < (1-\beta)N} \{1 - e^{c(N-i)\log(1-i/N)}\} \\ &\geqq \prod_{1 \leqq i < (1-\beta)N} \{1 - e^{-ci(1-i/N)}\} \\ &\geqq \prod_{1 \leqq i < (1-\beta)N} \{1 - e^{-ci\beta}\} \geqq \prod_{i=1}^{\infty} (1 - e^{-ci\beta}). \end{aligned} \tag{13}$$

Defining for $0 \leqq x < 1$ the function

$$\varphi(x) = \prod_1^{\infty} (1 - x^n) \geqq 1 - \sum_1^{\infty} x^n = 1 - x/(1 - x) \to 1 \quad \text{as} \quad x \to 0, \tag{14}$$

we have the uniform lower bound

$$P_N(J = N) \geqq \varphi(\beta^c) \cdot \varphi(e^{-c\beta}) \qquad (N = 2, 3, \cdots). \tag{15}$$

In particular, if we choose β to be the root $\beta_0 = 0.56 \cdots$ of the equation

$$\beta = e^{-\beta}, \tag{16}$$

then

$$P_N(J = N) \geqq \varphi^2(e^{-c\beta_0}). \tag{17}$$

Choosing c to satisfy $\varphi^2(e^{-c\beta_0}) = \alpha$ it follows that (9) holds.

We can improve (15) somewhat for large N. Write (13) in the form

$$\log R_N = \sum_{i=1}^{\infty} a_{i,N}, \tag{18}$$

where

$$\begin{aligned} a_{i,N} &= \log \{1 - (1 - i/N)^{[c(N-i)]+1}\} \quad \text{for} \quad 1 \leqq i < (1 - \beta)N, \\ &= 0 \qquad\qquad\qquad\qquad\qquad\quad \text{for } i \geqq (1 - \beta)N. \end{aligned} \tag{19}$$

For any fixed $i = 1, 2, \cdots$

$$\lim_{N \to \infty} a_{i,N} = \log (1 - e^{-ci}) = a_i, \text{ say}, \tag{20}$$

and for $1 \leqq i < (1 - \beta)N$ we have as in (13),

$$(21)\quad 0 \geqq a_{i,N} \geqq \log\{1 - (1 - i/N)^{c(N-i)}\}$$
$$\geqq \log(i - e^{-ci(1-i/N)}) \geqq \log(1 - e^{-ci\beta}),$$

so this holds for *all* i, N, and

$$(22)\qquad \sum_{i=1}^{\infty} \log(1 - e^{-ci\beta}) = \log \varphi(e^{-c\beta}) > -\infty.$$

By the dominated convergence theorem,

$$(23)\quad \lim_{N\to\infty} R_N = \lim_{N\to\infty} \exp[\sum_{i=1}^{\infty} a_{i,N}] = \exp[\sum_{i=1}^{\infty} a_i]$$
$$= \prod_{i=1}^{\infty}(1 - e^{-ci}) = \varphi(e^{-c}).$$

Since by (12), $\varphi(\beta^c)R_N \leqq P_N(J = N) \leqq R_N$, it follows that

$$\varphi(\beta^c)\varphi(e^{-c}) \leqq \liminf_{N\to\infty} P_N(J = N) \leqq \limsup_{N\to\infty} P_N(J = N) \leqq \varphi(e^{-c}),$$

and since β can be arbitrarily near 0, and $\varphi(\beta^c) \to 1$ as $\beta \to 0$,

$$(24)\qquad \lim_{N\to\infty} P_N(J = N) = \varphi(e^{-c}).$$

Thus from (17),

$$(25)\qquad \varphi^2(e^{-c\beta_0}) \leqq P_N(J = N) \to \varphi(e^{-c}) \quad \text{as} \quad N \to \infty.$$

For any $\epsilon > 0$, if we increase the first $j = j(\epsilon)$ terms of (10) we can clearly strengthen (25) to read

$$(26)\qquad \varphi(e^{-c}) - \epsilon \leqq P_N(J = N) \to \varphi(e^{-c}) \quad \text{as} \quad N \to \infty.$$

To see how efficient this procedure is, let us look at its sample size

$$(27)\qquad S = X_0 + \cdots + X_{J-1} + b_J \leqq X_0 + \cdots + X_{N-1} + b_N.$$

From (7),

$$(28)\qquad E_N(S) \leqq N(1 + \tfrac{1}{2} + \cdots + 1/N) + cN + 1,$$

which is somewhat greater than the fixed sample size $n = [N \log N + cN]$ for which we have seen that $P_{N,n} \to e^{-e^{-c}}$. Now

$$0 < \varphi(e^{-c}) < 1 - e^{-c} < e^{-e^{-c}} < 1,$$

so (26) shows that for $N \to \infty$ the probability of having drawn all the white balls by the sequential procedure is somewhat less than the corresponding probability for the fixed sample size $n = [N \log N + cN]$. Of course, the latter procedure requires a knowledge of N; moreover, the ratio of the two error probabilities is small for large values of c;

$$(29)\qquad (1 - e^{-e^{-c}})/(1 - \varphi(e^{-c})) \to 1 \quad \text{as} \quad c \to \infty.$$

Nevertheless, the fact remains that the sequential procedure is somewhat inefficient for any fixed c and large values of N. We therefore ask, is there any

sequential procedure such that for fixed $-\infty < c < \infty$,

(30) Sample size $\leqq N \log N + cN$,

P_N (all N white balls drawn by the time we stop) $\to e^{-e^{-c}}$ as $N \to \infty$?

An affirmative answer is given in the next section.

3. An asymptotically efficient procedure based on cumulative waiting times. We modify the sequential procedure of the previous section as follows. Let (a_n) be a sequence of positive constants, and define A_n to be the event that $Y_{n+1} > a_n$ $(n = 1, \cdots, N)$. Since $Y_{N+1} \equiv \infty$, A_N is certain. Let $I =$ first $n \geqq 1$ such that A_n occurs. Then $1 \leqq I \leqq N$. We agree to stop as soon as A_I occurs. Then (cf. (8))

$$
\begin{aligned}
P_N \text{ (all } N \text{ white balls drawn by the time we stop)} &= P_N(I = N) \\
&= P_N\Big(\bigcap_{n=1}^{N-1} A_n'\Big) = P_N\Big(\bigcap_{n=1}^{N-1} (Y_{n+1} \leqq a_n)\Big) \\
&= P_E\Big(\bigcap_{n=1}^{N-1} (X_0 + \cdots + X_n \leqq a_n)\Big).
\end{aligned} \tag{31}
$$

As before, we could define a "step-wise minimal" sequence (a_n^*) such that the expression (31) is $\geqq \alpha$ for every $N \geqq 2$, but since the events A_n are not independent the explicit computation of the a_n^* is difficult. Instead, as before, we shall choose an explicit sequence (a_n) and estimate the value of (31) when N is large. Our sequence is the following. Let c be any finite constant, let $n^* =$ smallest n such that $n \geqq e^{1-c}$, and define

$$
\begin{aligned}
a_n &= n + 1 \quad \text{for} \quad n = 1, \cdots, n^* - 1, \\
&= n \log n + cn \quad \text{for} \quad n \geqq n^*.
\end{aligned} \tag{32}
$$

It is clear that for this procedure the sample size is always $\leqq a_N$ $(= N \log N + cN$ for $N \geqq n^*)$. And we shall prove that $P_N(I = N) \to e^{-e^{-c}}$ as $N \to \infty$. (It will be seen from the proof that as in (26) for any $\epsilon > 0$ we can increase the first $j = j(\epsilon)$ values of (a_n) so as to make

$$e^{-e^{-c}} - \epsilon \leqq P_N(I = N) \to e^{-e^{-c}} \quad \text{as} \quad N \to \infty.)$$

It is easy to check that

$$
\begin{aligned}
a_n - a_{n-1} &\geqq 1 \quad \text{for all} \quad n \geqq 2, \\
&\geqq \log n + c \quad \text{for} \quad n \geqq n^*,
\end{aligned} \tag{33}
$$

and that the random variab es

$$w_n = X_{N-n}/N - 1/n \qquad (n = 1, \cdots, N) \tag{34}$$

are independent with

$$E_N(w_n) = 0, \qquad \mathrm{Var}_N\,(w_n) = 1/n^2 - 1/Nn < 1/n^2. \tag{35}$$

By Kolmogorov's inequality, for any $d > 0$,

$$(36)\quad P_N(w_1 + \cdots + w_n \leqq -d \text{ for some } n = 1, \cdots, N) \leqq (\textstyle\sum_1^\infty 1/n^2)/d^2 = \pi^2/6d^2.$$

We have

$$(37)\quad A_n' = (X_0 + \cdots + X_n \leqq a_n) = (w_{N-n} + \cdots + w_N \leqq b_n),$$
$$b_n = a_n/N - ((N-n)^{-1} + \cdots + 1/N).$$

And for $n \geqq 2$,

$$(38)\quad A'_{N-1} \cap (w_1 + \cdots + w_n \geqq b_{N-1} - b_{N-n}) \subset A'_{N-n}.$$

Now for $N \geqq n + n^*$,

$$b_{N-1} - b_{N-n} = ((a_{N-1} - a_{N-n})/N) + (1 + \tfrac{1}{2} + \cdots + 1/(n-1))$$
$$(39)\quad = (n-1)\log N/N - (1 + \tfrac{1}{2} + \cdots + 1/(n-1)) + c(n-1)/N + f(1 - n/N) - f(1 - 1/N),$$

where

$$(40)\quad 0 < f(x) = -x \log x < e^{-1} \quad \text{for} \quad 0 < x < 1.$$

Hence for $N \geqq n + n^*$,

$$(41)\quad b_{N-1} - b_{N-n} \leqq n \log N/N - \log n + \beta \quad (\beta = |c| + e^{-1}).$$

Put

$$(42)\quad p = d + \beta + 1, \qquad k = e^p.$$

Then as $N \to \infty$,

$$k \log N/N - \log k + \beta \to -d - 1,$$
$$(N(1 - p/\log N)/N) \log N - \log (N(1 - p/\log N)) + \beta \to -d - 1,$$

and for $N \geqq N_d$,

$$(43)\quad b_{N-1} - b_{N-n} \leqq -d \quad \text{for} \quad k \leqq n \leqq N(1 - p/\log N).$$

Hence from (38) and (36),

$$P_N(A'_{N-n} \quad \text{for all} \quad 1 \leqq n \leqq N(1 - p/\log N))$$
$$(44)\quad \geqq P_N(A'_{N-1} \cap \cdots \cap A'_{N-k}) - P_N(w_1 + \cdots + w_n \leqq -d \text{ for some } n = 1, \cdots, N)$$
$$\geqq P_N(A'_{N-1} \cap \cdots \cap A'_{N-k}) - \pi^2/6d^2.$$

But from (3), as $N \to \infty$

$$(45)\quad P_N(A'_{N-1} \cap \cdots \cap A'_{N-k}) \geqq P_N(Y_N \leqq a_{N-k})$$
$$= P_N((Y_N - N \log N)/N \leqq ((a_{N-k} - N \log N)/N) \to e^{-e^{-c}},$$

since as $N \to \infty$

$$(a_{N-k} - N \log N)/N = ((N-k)\log(N-k) - c(N-k) - N\log N)/N$$
$$= (1 - k/N)\log(1 - k/N) + (1 - k/N)\log N$$
$$- \log N - c(1 - k/N) \to -c.$$

Hence

(46) $\liminf_{N\to\infty} P_N(A'_{N-n} \text{ for all } 1 \leqq n \leqq N(1 - p/\log N)) \geqq e^{-e^{-c}} - \pi^2/6d^2.$

We shall show in a moment that

(47) $$\lim_{N\to\infty} P_N(A_n' \text{ for all } 1 \leqq n \leqq pN/\log N) = 1.$$

It will then follow from (44) and (45) that

(48) $\liminf_{N\to\infty} P_N(A_n' \text{ for all } 1 \leqq n \leqq N-1) \geqq e^{-e^{-c}} - \pi^2/6d^2.$

Since d can be arbitrarily large, (48) holds without the last term. But

(49) $$P_N(A_n' \text{ for all } 1 \leqq n \leqq N-1) \leqq P_N(A'_{N-1}) \to e^{-e^{-c}}$$

by (45) for $k = 1$. Hence

(50) $$\lim_{N\to\infty} P_N(A_n' \text{ for all } 1 \leqq n \leqq N-1) = \lim_{N\to\infty} P_N(I = N) = e^{-e^{-c}}.$$

It remains only to prove (47). Now setting $a_0 = 1$

$$P_N(A_n' \text{ for all } 1 \leqq n \leqq pN/\log N) = P_N\Big(\bigcap_{1\leqq n\leqq pN/\log N}(X_0 + \cdots + X_n \leqq a_n)\Big)$$
$$\geqq P_N\Big(\bigcap_{1\leqq n\leqq (pN/\log N)}(X_n \leqq a_n - a_{n-1})\Big)$$
$$= \prod_{1\leqq n\leqq pN/\log N}\{1 - (n/N)^{a_n - a_{n-1}}\}$$
$$\geqq 1 - \sum_{1\leqq n\leqq pN/\log N}(n/N)^{a_n - a_{n-1}},$$

and by (33), as $N \to \infty$

$$\sum_{1\leqq n\leqq N^{\frac{1}{2}}/\log N}(n/N)^{a_n - a_{n-1}} \leqq \sum_{1\leqq n\leqq N^{\frac{1}{2}}/\log N}(n/N) \leqq 1/(\log N)^2 \to 0,$$
$$\sum_{N^{\frac{1}{2}}/\log N < n\leqq pN/\log N}(i/N)^{a_n - a_{n-1}} \leqq \sum_{N^{\frac{1}{2}}/\log N < n\leqq pN/\log N}(p/\log N)^{\log n + c}$$
$$\leqq (p/\log N)^c \sum_{n > N^{\frac{1}{2}}/\log N} n^{\log p - \log\log N}$$
$$\leqq (p/\log N)^c \cdot \int_{N^{\frac{1}{2}}}^{\infty}(dx/x^2) \to 0,$$

which completes the proof of (47).

The answer to the question in Section 1 is no; for any M,

$$\lim_{N\to\infty} P_N(\text{estimate} = N) = 0.$$

REFERENCES

[1] FELLER, W. (1957). *Introduction to Probability Theory and its Applications.* **1** (2nd ed.) Wiley, N. Y.

[2] GOODMAN, L. A. (1953). Sequential Sampling Tagging for Population Size Problems. *Ann. Math. Statist.* 56-69.