SEQUENTIAL SAMPLING TAGGING FOR POPULATION SIZE PROBLEMS[1]

By Leo A. Goodman

The University of Chicago

Summary. Let P be a finite population which may have some subsets of known size but whose total size N is unknown. We shall consider the problem of point and interval estimation, tests of hypotheses, and fiducial distributions of N for some sampling-tagging procedures. The problem of estimating the number of classes in a population [1], [2], when it is known that the same number of elements is contained in each class, may be considered within the general problem discussed.

1. Introduction. We shall be interested in the following sequential sampling-tagging procedure $S(L, n_i)$. Let $\{n_i\}$ be a sequence of positive integers and let $S(L, n_i)$ denote the procedure whereby:

(1) n_1 elements are drawn at random from P, the number of elements which are drawn from subsets of P of known size is observed, the sampled elements are tagged so that they may be distinguished from the remaining elements, and replaced in P;

(2) n_2 elements are drawn from P, the number of tagged elements and elements drawn from subsets of known size is observed, the sampled elements are tagged and replaced in P;

(3) $\cdots$;

this procedure is discontinued when a total of at least $L > 0$ tagged elements or elements from the known subsets have been drawn.

The following practical cases are instances of this general procedure. It is well known that the decennial census is not complete. One would, therefore, like to estimate the total number actually living in the United States; that is, consider the problem of finding out how many people were not enumerated in the census. We would draw a sample of people and investigate how many people in this sample had been enumerated. We would then list the nonenumerated people in this sample, draw another sample and investigate how many people in this second sample had been enumerated in the census or had been listed in the first sample, . . . ; this procedure is discontinued when a total of at least L people have been found who had been enumerated in the census or listed in one of the preceding samples.

Suppose one wished to estimate the total population size two years after the

Received 2/13/51, revised form 8/15/52.

[1] Most of the results described in this paper were presented to the Summer Seminar in Statistics at the University of Connecticut, August 22, 1950, and to meetings of the Institute of Mathematical Statistics at Chicago, Illinois, December 27, 1950.

census had been taken. One might divide the population into those people who were enumerated in the census and are still alive, those born since the census was taken, and then the remaining people. Death records might be used in conjunction with census records to determine how many people who were enumerated are now dead. Using these results in conjunction with birth records we might proceed as before—sampling, investigating, listing, resampling, ...

In some areas the number of people in some classes of population (e.g., certain professional groups) is known. We could then use this information and a sampling, investigating, listing, resampling, . . . procedure to estimate the number of people in that area. One might, of course, wish to estimate the total population size of an area by these methods without resorting to any previous information.

The question concerning the number of animals of a given kind in a specified region is often of importance in ecological work. One might set traps in the area to catch these animals. When the traps are filled the animals are tagged and released, and the traps are moved and reset. The number of tagged animals appearing in the traps is then observed, the newly trapped animals are tagged, released, and we again move and reset the traps, ...

Marking methods are also used to estimate insect and fish populations. This paper studies procedures and estimates which may be of use in such problems.

Another case which is of some archeological interest deals with the problem of determining how many days there were in the calendar of some ancient civilization. By observing the days marked on gravestones and dealing with these days as though they were samples drawn at random from the total population of N days in the annual calendar, we might estimate N by observing more graves and taking note of days which occur more than once (marked elements reappearing).

In this study, we do not allow for population changes occurring during sampling nor for nonrandom sampling.

If $S(L, n_i)$ or a nonsequential fixed number of random samples procedure is applied, it is found that nondifferentiated tags (similar tags for each sample) are sufficient for the estimation problem and that the general problem may be reduced to the nondifferentiated case where P contains no subsets of known size. Given $S(L, n_i)$, there exists a minimum variance unbiased estimator (m.v.u.e.) of N which may be determined as the quotient of two determinants and simplified, by combinatorial methods, in special cases. Tables which shorten computation appear in [3].

If $\{n_i\}$ is bounded, as N approaches infinity, the limiting distribution of t^2/N, where t is the total number of elements drawn before the procedure ceases, is χ^2 with $2L$ degrees of freedom. Calculations indicate that the exact distribution differs only slightly from χ^2 when $N = 365$, $n_i = 1$, $L = 1$. Using the χ^2 approximation to the exact distribution, we find that the asymptotic m.v.u.e. of N is $t^2/2L$, and that approximate one- or two-sided confidence intervals may be obtained. The approximate fiducial distribution of N is given and certain tests of hypotheses concerning sizes of one and two populations are considered.

It is shown that any sequence of successive independent repetitions of the sequential procedure, $S(L_1, n_i)$, $S(L_2, n_i)$ $\cdots$ may be improved upon by a single $S(L, n_i)$. The $S(L, n_i)$ also compares favorably with other procedures considered.

A numerical illustration of the procedure $S(L, n_i)$ is also presented.

A list of references to discussions of the various practical considerations involved in sampling-tagging programs and of certain results thereof appears in [4] and [5] where other methods of sampling-tagging are analyzed.

2. Theorems on sufficient statistics and the existence of unbiased estimators.

Let P be a set of N elements and suppose that a notion of equivalence has been defined so that the elements of P may be said to belong to $T + 1$ different classes $P(0)$, $P(1)$, $P(2)$, $\cdots$, $P(T)$. The number of elements in $P(j)$ is $N(j)$, $\sum_{j=0}^{T} N(j) = N$, where the value of $N(j)$ is known for all but one class, say $P(0)$. Let K samples of n_i elements, $i = 1, 2, \cdots, K$, be drawn in order from P in such a manner that, before the ith sample is drawn, the n_{i-1} elements appearing in sample $i - 1$ are so labeled and then replaced in P. (Henceforth, this procedure shall be designated by $F(K, n_i)$.) Suppose h denotes the history of an element; that is, h is an index denoting the set of tags which have already been placed on an element. Let $n_i(j, h)$ be the number of elements from class $P(j)$ which appeared in the ith sample with the set of tags corresponding to the hth history; $n_i(j, 1)$ being the number of elements from class P (j)which first appeared in the ith sample. We have

THEOREM 1. *If the sampling-tagging procedure is $F(K, n_i)$ then the statistic $\sum_{i=1}^{K} n_i\ (0, 1)$ is sufficient for estimating $N(0)$.*

PROOF. A somewhat stronger result holds; namely, if $\Pr\{n_i(j, h); N(j), n_i, K\}$ denotes the joint probability function of all possible $n_i(j, h)$ (not for a fixed j and h), then

$$(1)\qquad \Pr\{n_i(j, h); N(j), n_i, K\} = C\left\{\sum_{i=1}^{K} n_i(0, 1),\ N(0)\right\} \cdot \frac{h(K)}{\prod_{i=1}^{K} C\{n_i, N\}},$$

where henceforth, $C\{a, b\} = b!/(b - a)!$, and where $h(K)$ is a functon of $n_1(j, h), n_2(j, h), \cdots, n_K(j, h), n_1, n_2, \cdots, n_K$ and $N(1), N(2), \cdots, N(T)$. Equation (1) may be obtained by direct calculations of the probabilities by means of standard combinatorial methods. The theorem then follows from the factorization conditions for sufficient statistics.

By Theorem 1, we see that for $F(K, n_i)$ no information is gained by tagging elements appearing from $P(1)$, $P(2)$, $\cdots$, $P(T)$ nor by using different tags for different samples.

Suppose we consider the general class of procedures where random samples are drawn from P in such a manner that if it is decided that an ith sample is to be drawn, the n_{i-1} elements of sample $i - 1$ are first so labeled and replaced in P before the ith sample is drawn. The total number x of samples drawn will be determined by some stopping rule which may depend on the sample results. We shall define a sufficient statistic to be one for which the conditional distribution, when x is fixed, can be factored as usual. Clearly by Theorem 1, $\sum_{i=1}^{x} n_i(0, 1) = \mu$ is a sufficient statistic and in estimating $N(0)$ it will be sufficient to use μ and x. For some procedures (stopping rules) there may exist a function $g(\mu)$ of the total number μ of untagged elements drawn from $N(0)$ which is such that $g(\mu) = x$. For such procedures it will be sufficient to use μ in estimating $N(0)$.

Corollary 1. *Given $S(L, n_i)$, then $\mu = \sum_{i=1}^{x} n_i(0, 1)$ is sufficient for estimating $N(0)$.*

Proof. Let $g(\mu)$ be the least integer g such that

$$\sum_{i=1}^{g} n_i \geqq L + \mu.$$

This function $g(\mu)$ is such that $g(\mu) = x$. Q.E.D.

In view of the preceding results, without any loss of generality, we shall, henceforth, consider the problem of estimating the size of a population, which is not divided into classes, by means of nondifferentiating tagging methods. We let μ_i denote the number of untagged elements drawn in the ith sample of n_i elements.

Theorem 2. *Given $S(L, n_i)$, there exists a minimum variance unbiased estimator* (m.v.u.e.) $M(\mu)$.

Proof. It is clear that if $\mu = \sum_{i=1}^{x} \mu_i$, then $\Pr\{\mu; N, n_i\}$ is zero for $\mu > N$ and is positive for $n_1 \leqq \mu \leqq N$. Hence, an unbiased estimator $M(\mu)$ would satisfy the system of equations

$$N = \sum_{\mu=n_1}^{N} M(\mu) \Pr\{\mu; N, n_i\} \tag{2}$$

for $N = n_1, n_1 + 1, n_1 + 2, \cdots$. This system of equations defines an estimator $M(\mu)$ uniquely since its values can be determined recursively for $\mu = n_1, n_1 + 1, n_1 + 2, n_1 + 3, \cdots$. That $M(\mu)$ is the minimum variance unbiased estimator follows from Corollary 1 and Blackwell's result [6] that given any other unbiased estimate, one can obtain an unbiased sufficient estimate whose variance is at least as small. The uniqueness of an unbiased sufficient estimate insures that $M(\mu)$ is the desired estimate. Q.E.D.

The reader will see by examining the preceding proof that a result similar to Theorem 2 for somewhat more general procedures than $S(L, n_i)$ will still hold. This result indicates one advantage of sequential procedures since Chapman ([4], pp. 149–150) has shown that no unbiased estimators exist for the procedures he was considering (fixed number of samples).

Corollary 2. *For $S(L, n_i)$ the m.v.u.e. is*

$$M(\mu; L, n_i) = \frac{|a_{ij}|}{|b_{ij}|}, \qquad i, j = 1, 2, \cdots, \mu - n_1 + 1,$$

where

$$b_{ij} = \begin{cases} 1, & i = \mu - n_1 + 1 \\ \dfrac{C(n_1 + i - 1, n_1 + j - 1)}{\prod_{s=1}^{g(n_1+i-1)} C(n_s, n_1 + j - 1)} & i < \mu - n_1 + 1, \end{cases}$$

with $b_{ij} = 0$ *when undefined and* $a_{ij} = b_{ij}$ *for* $i < \mu - n_1 + 1, a_{\mu-n_1+1,j} = n_1 + j - 1$, *and* $g(\mu)$ *is as in Corollary* 1.

Proof. In view of the proof of Theorem 2, we need only show that $M(\mu;\ L, n_i)$ satisfies the system of equations (2). Since the number of samples K is a function of μ, and hence fixed for each μ, equation (1) may be used to give

$$\Pr\{\mu; L, N, n_i\} = \frac{C(\mu, N)}{\prod_{s=1}^{g(\mu)} C(n_s, N)} h(\mu, n_t, L).$$

Now since

$$\sum_{\mu=n_1}^{N} \Pr\ \{\mu; L, N, n_t\} = 1,$$

we have

$$\frac{C(n_1, n_1)}{\prod_{s=1}^{g(n_1)} C\ (n_s, n_1)} h(n_1, n_t, L) = 1,$$

$$\frac{C(n_1, n_1 + 1)}{\prod_{s=1}^{g(n_1)} C(n_s, n_1 + 1)} h(n_1, n_t, L) + \frac{C(n_1 + 1, n_1 + 1)}{\prod_{s=1}^{g(n_1+1)} C(n_s, n_1 + 1)} h(n_1 + 1, n_t, L) = 1,$$

$$\frac{C(n_1, n_1 + 2)}{\prod_{s=1}^{g(n_1)} C(n_s, n_1 + 2)} h(n_1, n_t, L) + \frac{C(n_1 + 1, n_1 + 2)}{\prod_{s=1}^{g(n_1+1)} C(n_s, n_1 + 2)} h(n_1 + 1, n_t, L) + \frac{C(n_1 + 2, n_1 + 2)}{\prod_{s=1}^{g(n_1+2)} C(n_s, n_1 + 2)} h(n_1 + 2, n_t, L) = 1,$$

$$\vdots$$

Using Cramér's rule, we have

$$h(\mu, n_t, L) = \frac{|\,b_{ij}\,|}{|\,c_{ij}\,|},$$

where

$$c_{ij} = \frac{C(n_1 + i - 1, n_1 + j - 1)}{\prod_{s=1}^{g(n_1+i-1)} C(n_s, n_1 + j - 1)}$$

for $i, j = 1, 2, \cdots, \mu - n_1 + 1$ (with $c_{ij} = 0$ when undefined). Suppose M_μ is the m.v.u.e. and let

$$M_\mu h(\mu, n_t, L) = f(\mu, n_t, L).$$

Hence, by equations (2)

$$\sum_{\mu=n_1}^{N} \frac{C(\mu, N)}{\prod_{s=1}^{g(\mu)} C(n_s, N)} f(\mu, n_t, L) = N$$

for $N = n_1, n_1 + 1, n_1 + 2, \cdots$ and

$$f(\mu, n_t, L) = \frac{| a_{ij} |}{| c_{ij} |}.$$

Therefore

$$M_\mu = M(\mu; L, n_t) = \frac{| a_{ij} |}{| c_{ij} |} \cdot \frac{| c_{ij} |}{| b_{ij} |} = \frac{| a_{ij} |}{| b_{ij} |}. \qquad \text{Q.E.D.}$$

3. The m.v.u.e. in special cases.

THEOREM 3. *For $S(L, 1)$, the m.v.u.e. is*

$$M(\mu; L) = \frac{K(\mu, L)}{K(\mu, L - 1)},$$

where $K(\mu, L) = \mu\sum_{s=1}^{\mu} s\sum_{v=1}^{s} v\sum_{w=1}^{v} w \cdots \sum_{z=1}^{y} z$, when there are L summations; for example, $K(\mu, 0) = \mu$, $K(\mu, 1) = \mu^2(\mu + 1)/2$.

In order to prove this result, we first prove

LEMMA 1.

$$\sum_{\mu=1}^{N} K(\mu, L) \frac{C(\mu, N)}{N^\mu} = N^{L+1},$$

$$\sum_{\mu=t}^{N} K(\mu, L) \frac{C(\mu, N)}{N^\mu} = \frac{C(t, N)}{t - 1} \sum_{s=0}^{L} \frac{K(t - 1, s)}{N^{t-L+s-1}}, \qquad 1 < t \leqq N.$$

PROOF. We first consider the special case $L = 0$. We wish to prove that

$$\sum_{\mu=t}^{N} \mu \frac{C(\mu, N)}{N^\mu} = \frac{C(t, N)}{N^{t-1}}, \qquad t = 1, 2, \cdots, N.$$

The equality clearly holds for $t = N$. Now let us suppose the equality holds for t fixed, and consider $t - 1$. The left side of the equality becomes

$$(t - 1) \frac{C(t - 1, N)}{N^{t-1}} + \sum_{\mu=t}^{N} \mu \frac{C(\mu, N)}{N^\mu}$$

$$= (t - 1) \frac{C(t - 1, N)}{N^{t-1}} + \frac{C(t, N)}{N^{t-1}} = \frac{NC(t - 1, N)}{N^{t-1}}$$

$$= \frac{C(t - 1, N)}{N^{t-2}}.$$

Hence, by inverse mathematical induction we have proven the lemma for the special case where $L = 0$ and $t = 1, 2, \cdots, N$. The lemma may now be proved in the general case by induction on L.

Now the proof of Theorem 3 follows as an application of Lemma 1 and the methods used to prove Corollary 2.

In general $M(\mu; L)$ may be expressed as a rational function in μ,

$$M(\mu; L) = \frac{\mu^2}{2L} + \left(\frac{2}{3} - \frac{1}{6L}\right)\mu + \frac{P_1}{P_2},$$

where P_1 and P_2 are polynomials of degree $2L - 1$. Using various recursion relationships, the exact calculation of $M(\mu; L)$ may be simplified by means of tables (see [3]).

Theorem 4. *Consider a modified $S(1, n_i)$, where the only change is the addition of the condition that no more than R samples are to be drawn. Then*

$$E\{M(\mu; n_i)\} = N - \frac{C(t(R) + 1, N)}{\prod_{s=1}^{R} C(n_s, N)},$$

where

$$M(\mu; n_i) = \frac{1}{2}\left[t^2(x - 1) - \sum_{i=1}^{x-1} n_i^2 + 2t(x - 1)\right] + \mu_x\left[\frac{1 + t(x - 1)}{1 + n_x - \mu_x}\right],$$

when x is the number of samples drawn before sampling ceased,

$$\mu = \sum_{s=1}^{x-1} n_s + \mu_x, \qquad t(x) = \sum_{s=1}^{x} n_s.$$

To show that this theorem is true we first prove

Lemma 2. *For any positive integers $n \leqq A < N$, then*

$$\sum_{j=A-n}^{A}\left[\frac{n(1 + A) + j}{A - j + 1}\right]\frac{C(N - n - j, N - A)C(j, A)}{(N - n - j)!j!} + C(n + 1, N - A)/n! = C(n, N)N/n!.$$

Proof. We have the following identity

$$(1 + y)^{N-A}(1 + y)^A\{n + (n + 1)y\} = (1 + y)^N n + (n + 1)(1 + y)^N y.$$

The coefficient of y^{N-n} in the expansion of the right side of the identity in powers of y is

$$C(n, N)/(n - 1)! + C(n + 1, N)/n! = C(n, N)N/n!,$$

the right side of the equality stated in Lemma 2. The coefficient of y^{N-n} in the expansion of the left side of the identity in powers of y may be seen to be equal to the left side of the equality in Lemma 2. Hence, the equality is proven. Q.E.D.

We may prove Theorem 4 for the special case $R = 2$ using Lemma 2 by setting $A = n_1$, $n = n_2$, and $j = n_1 - n_2 + \mu_2$. We may then proceed to prove the theorem in general by mathematical induction on R using again Lemma 2.

By this method of proof we also see that $M(\mu;\, n_i)$ is the m.v.u.e. for the original $S(L,\, n_i)$. The calculation of $M(\mu;\, n_i)$ simplifies considerably in the special case where $n_2 = n_3 = n_4 = \cdots = n$ (which is of some interest in applications). Similar kinds of results may be obtained for $L = 2$, and $n_2 = n_3 = n_4 = \cdots = n$.

4. Limit theorems. An interesting dual relationship exists between the family of sampling procedures $S(L,\, n_i)$ and $F(K,\, n_i)$ which will be useful. Namely, if $H(L;\, K,\, n_i,\, N)$ is the distribution function of the total number L of tagged elements appearing when the sampling procedure $F(K,\, n_i)$ is used, and if $G(K;\, L,\, n_i,\, N)$ is the distribution function of the total number K of samples drawn before sampling ceased when $S(L,\, n_i)$ is applied, then

$$G(K; L, n_i, N) = 1 - H(L - 1; K, n_i, N).$$

We also have, then, that

$$\begin{aligned} h(L; K, n_i, N) &= H(L; K, n_i, N) - H(L - 1; K, n_i, N) \\ &= G(K; L, n_i, N) - G(K; L + 1, n_s, N) \end{aligned}$$

and

$$\begin{aligned} g(K; L, n_i, N) &= G(K; L, n_i, N) - G(K - 1; 1, n_i, N) \\ &= H(L - 1; K - 1; n_i, N) - H(L - 1; K, n_i, N). \end{aligned}$$

Henceforth, these relationships will be designated as Relation A.

Throughout the following sections $[n_i]$ is to be any bounded sequence of positive integers. We now have

THEOREM 5. *Let $K(N)$ be any integer-valued function such that*

$$\lim_{N\to\infty} \frac{t^2[K(N)]}{N} = y,$$

where

$$t[x] = \sum_{i=1}^{x} n_i.$$

Then

$$\lim_{N\to\infty} h(L; K, n_i, N) = \frac{e^{-y/2}}{L!}\left(\frac{y}{2}\right)^L,$$

where $h(L;\, K,\, n_i,\, N)$ is as in Relation A.

In order to prove this result, we need the following lemma:

LEMMA 3.

$$h(L; K, n_i, N) = \frac{C(t[K] - L, N)}{\prod_{s=1}^{K} C(n_s, N)}\, p(n_i, L),$$

where $p(n_i, L)$ is a polynomial in n_i such that

$$p(n_i, L) = \frac{t[K]^{2L}}{L!2^L} + 0(t[K])^{2L-1},$$

when $K = K(N)$.

This lemma may be proved directly from the fact that

$$t[K]^L = \sum n_{i_1}n_{i_2} \cdots n_{i_L} + 0(t[K])^{L-1},$$

where the summation is taken over all sets $(i_1, i_2, \cdots, i_L)$ such that $i_s \neq i_t$ for $s \neq t$ and $i_s = 1, 2, \cdots, K$. This statement may be proved by induction on L.

Theorem 5 now follows from Lemma 3 and the fact that

$$\lim_{N\to\infty} \frac{C(t(K[N]), N)}{\prod_{s=1}^{K[N]} C(n_s, N)} = e^{-y/2},$$

a result obtained using Stirling's formula.

Theorem 6. *Let $D(y)$ be the distribution function of $t^2(K)/N = y$ where $t(K) = \sum_{s=1}^{K} n_s$ when sampling proceeds according to $S(L; n_i)$. Then*

$$\lim_{N\to\infty} D(y) = \frac{1}{2^L\Gamma(L)} \int_0^y x^{L-1} e^{-x/2}\, dx.$$

The result follows by induction on L and use of Relation A applied to Theorem 5.

We also have

Corollary 3. *If $S(L, n_i)$ is used, then the limiting distribution of $t/\sqrt{N}$ is $\sqrt{\chi^2}$ with $2L$ degrees of freedom.*

Proof. This is a direct application of a theorem stated by Wilks ([7], p. 219).

Suppose the sampling procedure $S(L, n_i)$ was independently repeated m_L times, for $L = 1, 2, \cdots, \sum_{L=1}^{\infty} m_L = m$. By the preceding result it is clear that

$$dF(t_{Li}) = \prod_{L=1}^{\infty} \frac{\exp\left[-\sum_{i=1}^{m_L} t_{Li}^2/(2N)\right]}{(2N)^{Lm_L}\Gamma(L)^{m_L}} \prod_{i=1}^{m_L} (t_{Li}^2)^{L-1}\, d(t_{Li}^2)$$

is approximately true when dealing with large N, where t_{Li} is the number of elements drawn when $S(L, n_i)$ was performed for the ith time. Henceforth, we shall assume that the preceding statement is exactly so; that is, all statements made should now be interpreted as being only approximately true, the approximation being close when we are dealing with large N.

Since it is easy to see that $\sum_{L=1}^{\infty} \sum_{i=1}^{m_L} t_{Li}^2$ is sufficient for estimating N, we may use a result stated by Kendall ([8], Vol. 2, p. 54) and the Blackwell theorem [6] to show that

$$M = \frac{\sum_{L=1}^{\infty} \sum_{i=1}^{m_i} t_{Li}^2}{2 \sum_{L=1}^{\infty} L m_L}$$

is the m.v.u.e. (asymptotic) of N. It can easily be seen that

$$\sigma_M^2 = \frac{N^2}{\sum_{L=1}^{\infty} L m_L},$$

and hence we may obtain standard errors for our estimates of N.

We shall now see that the best results will be obtained when $\sum_{L=1}^{\infty} m_L = 1$; that is, when $S(L, n_i)$ is not repeated. More precisely:

THEOREM 7. *Consider all sequences of nonnegative integers* $m_1, m_2, m_3, \cdots$, *such that* $\sum_{L=1}^{\infty} L m_L$ *is a fixed integer. Suppose the sampling procedure* $S(L, n_i)$ *was independently repeated* m_L *times, for* $L = 1, 2, 3, \cdots$. *Then the variance of the m.v.u.e. (asymptotic) of* N *is the same for each sequence. The expected number of elements drawn is a minimum when* $\sum_{L=1}^{\infty} m_L = 1$.

PROOF. Since $\sigma_M^2 = N^2/\sum_{L=1}^{\infty} L m_L$, the variance of M the m.v.u.e. (asymptotic) of N is the same for each sequence. The expected number of elements drawn, divided by $\sqrt{N}$ approaches

$$\sum_{L=1}^{\infty} \sqrt{\pi}\, m_L \Gamma(2L)/\Gamma^2(L) 2^{2L-3/2}$$

as N becomes large. The following inequality may be proved directly:

$$\frac{\Gamma(2\sum_{L=1} L m_L)}{\Gamma^2\left(\sum_{L=1}^{\infty} L m_L\right) 2^{2\Sigma L m_L}} < \sum_{L=1}^{\infty} m_L \frac{\Gamma(2L)}{\Gamma^2(L) 2^{2L}},$$

equality holding only in the trival case where $\sum_{L=1}^{\infty} m_L = 1$. Hence, the expected number of elements drawn is a minimum when $\sum_{L=1}^{\infty} m_L = 1$. Q.E.D.

Hence, we see that the amount of information about N is greater when $S(L, n_i)$ is performed once if the expected number of elements sampled is held constant. Since M is also an efficient estimator of N, we have the result that $(-1 + M/N) \cdot \sqrt{\sum_{L=1}^{\infty} L m_L}$ approaches normality, as $\sum L m_L$ becomes large in a sense more rapidly when we confine ourselves to not repeating $S(L, n_i)$.

5. An exact distribution and comparisons with limit results. It is clear that the distributions of t^2/N and μ^2/N behave similarly for large populations. We believe that when the n_s is relatively small, then the use of t gives us a good approximation to the limiting results. However, when n_s is not small, then $\mu + L$ might be better. The following table shows that when the population size is only $N = 365$ (the archeologist's problem), $L = 1$, and $n_s = 1$, we obtain good approximations using μ, and still better approximations using $\mu + 1 = t$.

TABLE 1

Comparison of the exact distribution and the χ^2 approximation with 2 degrees of freedom for $N = 365$

Probability	.01	.02	.05	.1	.2	.3	.5	.7	.8	.9	.95	.98	.99
χ^2	.02	.04	.10	.21	.45	.71	1.39	2.41	3.22	4.61	5.99	7.82	9.21
$t^2/365$	.02	.04	.10	.22	.46	.70	1.33	2.30	3.17	4.38	5.80	7.41	8.59

For example, $\Pr\{\chi^2 \leqq .71\} = .3$, $\Pr\{t^2/365 \leqq .70\} = .3$, and $\Pr\{\mu^2/365 \leqq .62\} = .3$.

6. Approximate confidence intervals, fiducial limits, and tests of hypotheses. We repeat that all statements made in the following sections should be interpreted as approximately so, in so far as we shall assume that t^2/N has its limiting χ^2 distribution.

Since the distribution function of t^2/N is independent of N, we may set up one- or two-sided confidence intervals for N. It is easy to see that the theory of fiducial inference leads to the following fiducial distribution of N:

$$dF(N \mid t) = (t^2/N)^L \frac{e^{-t^2/2N}\, dN}{\Gamma(L) 2^L N}$$

and that limits obtained in this way for N are the same as those obtained by the confidence interval approach.

To test the hypothesis at a level of significance of ϵ that $N = N_0$ against the alternative that $N > N_0$, the region of rejection is $t^2 > c_1 N_0$ where c_1 is such that $\int_{c_1}^{\infty} dF(\chi^2 \mid 2L) = \epsilon$. A similar result holds for the alternative $N < N_0$.

To test the hypothesis at a level of significance of ϵ that $N = N_0$ against the alternative $N \neq N_0$, a uniformly most powerful unbiased test exists. The region of rejection is $t^2 < c_2 N_0$ and $t^2 > c_3 N_0$ where c_2 and c_3 are such that $\int_{c_2}^{c_3} dF(\chi^2 \mid 2L) = 1 - \epsilon$ and $c_2^L e^{-c_2/2} = c_3^L e^{-c_3/3}$. If we 'wish to test the hypothesis that two populations were of the same size, Fisher's Z distribution, $Z = \frac{1}{2} \log \frac{t_1^2 L_2}{t_2^2 L_1}$, with parameters $v_1 = 2L_1$ and $v_2 = 2L_2$, seems appropriate.

7. Comparisons. Let us first consider the nonsequential sampling procedure $F(1, n)$. It is easy to see that the maximum likelihood estimate of N in this case is the greatest integer less than or equal to $n^2/m = T$, where m is the number of marked elements appearing in the second sample. When $n = O(\sqrt{N})$, we have that the expected value of T/N approaches infinity. We have considered the case where $n = O\ (\sqrt{N})$ since the sequential sampling-tagging with which it is compared has the property that the expected number of elements drawn is $O\ (\sqrt{N})$. However, if we are willing to draw sample of size KN, then, although T is still not unbiased, we have that T/N converges in probability to one.

Suppose we consider the following sequential procedure. A sample of n_1 elements is drawn from a population P of N elements, tagged and replaced in P. Then elements are drawn one at a time and then replaced until a total of L tagged elements have been drawn. (This is similar to the sequential sampling-tagging procedure we have been discussing except that $n_2 = n_3 = \cdots = 1$ and we do not tag elements appearing after the first sample.) This scheme was studied by Haldane ([9], p. 222) for the purpose of estimating $p = n_1/N$, while we are interested in the estimation of n_1/p. We can without much difficulty show that if this sequential non-retagging method is used, then

$$E\{n_1 y/L\} = \mathrm{N},$$

and $\sigma^2_{(n_1 y)/L} = (N^2 - n_1 N)/L$, where y is the number of elements drawn, after the n_1 of the first sample, before sampling ceased. If we consider n_1 as fixed, then the total number of elements drawn is $O(N)$, whereas the variance is no better than that obtained in the sequential sampling-tagging case for only $O(\sqrt{N})$ elements. Now let us suppose that $n_1 = \sqrt{KN}$. Then $E(y) = L\sqrt{N/K}$ and the expected value of the total number of elements drawn is $\sqrt{NK}(1 + L/K)$. Since $K \neq 0$, it is easy to see that for a given L, the minimum value of $\sqrt{NK}$ $(1 + L/K)$ is obtained when $K = L$, and so we have $2\sqrt{NL}$. When N is large, we have seen in the sequential sampling-tagging procedure that the expected value of the total number of elements drawn is about $\sqrt{N}\ E(\sqrt{\chi^2})$, where χ^2 has $2L$ degrees of freedom. Since $E(\chi^2) = 2L$, and since $\sqrt{E(\chi^2)} \geqq E(\sqrt{\chi^2})$, we have that $2\sqrt{NL} \geqq \sqrt{2NL} \geqq E(\sqrt{\chi^2})\ \sqrt{N}$. Hence, we have shown in this case that, for a given variance, the expected value of the total number of elements drawn before cessation is smallest when the sequential sampling-tagging procedure is used for large populations. Let us now consider the case where $n_1 = KN$, $K < 1$. Then for any sequential nonretagging procedure, there exists a sequential sampling-tagging method which obtains estimators whose variance is smaller, and also has the property that the expected value of the total number of elements necessarily drawn in the process is smaller than in the given sequential nonretagging method used for large populations.

8. A numerical illustration. Seven samples of 100 ($n_1 = n_2 = \ldots = n_6 = n_7$) followed by samples of one ($n_8 = n_9 = \cdots = 1$) were drawn from a population of $N = 10{,}000$ random numbers [10]. Sampling ceased when a total of $L = 25$ elements had reappeared. On the basis of the sample results, we then estimated N. The standard deviation of our estimate is about $\sigma = 10{,}000/5 = 2{,}000$. The results of this sampling experiment are summarized in Table 2 which Mr. Sylvanus Tyler of Argonne National Laboratory was kind enough to prepare. We find that estimate of N is $t^2/50 = (756)^2/50 = 11{,}704.5$. Also an estimate of σ is $11{,}704.5/5 = 2{,}340.9$. The case where N is known is of little practical use but serves to illustrate the methods presented herein.

TABLE 2

Reappearance of tagged elements in samples from a population of 10,000 random numbers

Sample	Elements in Sample	Tagged Elements in Population	Elements Reappearing
1	100	0	0
2	100	100	4
3	100	196	4
4	100	292	1
5	100	391	3
6	100	488	2
7	100	586	7
8, 9, . . . , 72	65	679	4
Total.	765	740	25

9. Acknowledgements. Most of the research described herein was done while the author was a Research Training Fellow of the Social Science Research Council at the Mathematics Department, Princeton University. A part of the work was supported by the Office of Naval Research. The author is deeply indebted to Professors S. S. Wilks and J. W. Tukey for their encouragement and helpful comments. This writer also wishes to thank the referee for his very useful suggestions.

REFERENCES

[1] F. Mosteller, "Questions and answers," *Amer. Statistician*, Vol. 3, (1949), pp. 12–13.

[2] L. A. Goodman, "On the estimation of the number of classes in a population," *Ann. Math. Stat.* Vol. 20 (1949), pp. 572–579.

[3] L. A. Goodman, "The estimation of population size using sequential sampling tagging method," unpublished thesis, Princeton University, 1950.

[4] Douglas G. Chapman, "Some properties of the hypergeometric distribution with applications to zoological sample censuses," *Univ. California Publ. Stat.*, Vol. 1 (1951), pp. 131–160.

[5] NORMAN T. J. BAILEY, "On estimating the size of mobile populations from recapture data," *Biometrika*, Vol. 38 (1951), pp. 293–306.

[6] DAVID BLACKWELL, "Conditional expectation and unbiased sequential estimation," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 105–110.

[7] S. S. WILKS, *Mathematical Statistics*, Princeton University Press, 1943.

[8] M. G. KENDALL, *The Advanced Theory of Statistics*, Vols. 1 and 2, Charles Griffin and Company, Ltd., London, 1948.

[9] J. B. S. HALDANE, "On a method of estimating frequencies," *Biometrika*, Vol. 33 (1943–1946), p. 222.

[10] *Table of 105,000 Random Decimal Digits*, Interstate Commerce Commission, Statement No. 4914, File No. 261-A-1, Washington, D. C., 1949.