SI

INVERSE, MULTIPLE AND SEQUENTIAL SAMPLE CENSUSES

By

Douglas G. Chapman

University of Washington

Technical Report No. 9

December 12, 1952

Contract N8onr-520 Task Order II
Project Number NR-042-038

Laboratory of Statistical Research
Department of Mathematics
University of Washington
Seattle, Washington

Inverse, Multiple and Sequential Sample Censuses[1]

Douglas G. Chapman

University of Washington

The enumeration of populations by sampling methods has only recently come into widespread use, though the idea dates back at least to 1783 when it was proposed by Laplace [9]. Approximately a century later, Petersen [13] used the tag and sample method to enumerate a plaice population. The basic procedure consists of tagging or marking a number of individuals within the population and subsequently sampling, at random, the whole population. Estimates of the population are based on the random number of tag recoveries in the sample; the number of individuals marked and the sample size are known parameters. A study of such estimates and of tests for the population size has been made by the author [3], [4].

While this simple model suffices for many purposes it is apparent that more complex models will be often necessary or desirable. In many cases the tagging and sampling is carried out in several stages. Such a procedure is referred to, here, as a multiple sample census. Various point and interval estimation formulae based on this procedure have been given by Schnabel [17], Schumacher and Essmeyer [18], De Lury [7], [7a], Schaeffer [16] and the author. This type of procedure lends itself to a sequential procedure. Tests and estimation formulae based on such multiple and sequential procedures are formulated in this paper.

An even simpler modification of the single step census is to "invert" the sampling process i.e. to fix the number of tagged individuals to be recovered by sampling, rather than fixing the sample size. Some formulae using this idea were given recently by Bailey [1].

[1]Research partially supported by the Office of Naval Research.

INVERSE SAMPLE CENSUSES

The following notation will be used throughout:

N: the population size being studied

t_i: the number of marked individuals in the population at the time the i-th sample is taken

n_i: the number of individuals taken in the i-th sample

s_i: the number of tagged individuals recovered in the i-th sample

s_{ij}: the number of individuals recovered in the i-th sample that had been tagged at the j-th tagging.

For convenience we also define $n_0 = t_1$ and $s_0 = 0$. The subscript i may be taken to run from 1 to k. Where k is 1, i.e. the sample census is conducted in one stage, the subscripts will be omitted for simplicity.

If the inverted sampling plan is used the s_i are fixed parameters, the n_i are random variables. For the single sample case n is a negative binomial or negative hypergeometric random variable according to whether sampling is with or without replacement.

If sampling is with replacement, Pr(n) the probability of having to sample n individuals to obtain s marked ones is given by

$$(1) \quad Pr(n) = \binom{n-1}{s-1} \left(\frac{t}{N}\right)^s \left(1-\frac{t}{N}\right)^{n-s}$$

and $E(n) = \frac{sN}{t}$.

Hence $\frac{nt}{s}$ is an unbiased estimate of N, which, as the author has pointed out [4], is not true when n is fixed and s the random variable. The variance of $\frac{nt}{s}$ is $(N^2-Nt)s^{-1}$ and an unbiased estimate of this variance is

$$(2) \quad \sigma^2_{\frac{nt}{s}} = \frac{nt^2}{s^2}\left[\frac{n-s}{s+1}\right]$$

As noted in Appendix I

$$(3) \qquad s = \frac{\frac{nt}{s} - N}{\sqrt{\frac{N^2}{s} - \frac{Nt}{s}}}$$

is approximately distributed according to N(0,1) for large s. This fact may be used to set up confidence intervals and tests for N.

For s and $\frac{t}{N}$ both small, as will be more frequently the case, the second limiting distribution given in Appendix I will be more useful. For this range of the parameters $\frac{2nt}{N}$ has approximately a χ^2 distribution with 2s degrees of freedom. This is equivalent to the Poisson approximation to the binomial and is related to the results of Sandelius [15] concerning inverse sampling with random variables distributed according to a Poisson distribution.

If $\chi^2_{2s}(\varepsilon)$ denotes the ε-quartile of the χ^2 distribution with 2s degrees of freedom i.e.

$$P\left(\chi^2_{2s} \leq \chi^2_{2s}(\varepsilon)\right) = \varepsilon$$

then $(1-\varepsilon)$ confidence limits for N are given by

$$(4) \qquad \underline{N} = \frac{2tn}{\chi^2_{2s}(1-\varepsilon/2)} \; ; \; \overline{N} \quad \frac{2tn}{\chi^2_{2s}(\varepsilon/2)}$$

The problem of improving on these confidence limits in a procedure similar to that treated in [3] will be considered elsewhere.

If the sampling occurs without replacement

$$(5) \qquad Pr(n) = \frac{(n-1)!\; t!\; (N-t)!\; (N-n)!}{(s-1)!\; (t-s)!\; (n-s)!\; N!\; (N-n-t+s)!}$$

and

$$(6) \qquad E(n) = \frac{(N+1)s}{t+1}$$

so that

$$(7) \qquad \hat{N} = \frac{n(t+1)}{s} - 1$$

is an unbiased estimate of N. In contrast to the direct sample census estimate, this unbiasedness does not depend on the parameters s and t.

Furthermore

$$(8)\quad \sigma^2_{\frac{n(t+1)}{s}} = \frac{(N+1)(N-t)(t-s+1)}{s(t+2)} \doteq \frac{N^2}{s}$$

This approximate formula (which exaggerates the actual variance) may be useful in the determination of the choice of s: by an appropriate choice of s

$\frac{\sigma_{\hat{N}}}{N}$ can be fixed at any desired level.

For testing purposes we note that $\hat{N}$ is approximately normal with mean N and variance given by (8) for large s. A model similar to that used by David [6] can be constructed to prove that as s tends to infinity n is asymptotically normally distributed.

On the average the inverse sampling procedure is better than the direct sampling procedure. For if n', s', denote the fixed sample size and the random number of tags recovered in the direct procedure, and if s is chosen equal to $\frac{n't}{N}$, then

$$E(n) = \frac{1+\frac{1}{N}n'}{1+\frac{1}{t}} < n'$$

while $\sigma^2_{\hat{N}} < \frac{N^2}{s} = N^2\left(\frac{N}{n't}\right) < \sigma^2_{\frac{(n'+1)(t+1)}{s'+1}}$;

$\frac{(n'+1)(t+1)}{s'+1}$ is the almost unbiased estimate in the direct sampling case, Hence a more efficient estimate is obtained with less average effort.

On the other hand if the experimenter knows absolutely nothing about the possible population size, then by an improper choice of t and s, E(n) may be extremely large. Moreover

$$\sigma^2_n = \frac{s(N+1)(N-t)(t-s+1)}{(t+1)^2(t+2)} \doteq s\left(\frac{N}{t}\right)^2$$

is very large; this may be regarded as an undesirable feature of the procedure.

These difficulties may be partly overcome by a modification of the inverse sampling plan as follows: the number of untagged individuals to be recaptured is predetermined, rather than the number of tagged individuals. In other words, n-s is chosen in advance of sampling: s and n are now both random variables, though completely dependent. For convenience write $n-s=\delta$.

$$(9)\qquad Pr(n) = \frac{(n-1)!\ t!(N-t)!\ (N-n)!}{(n-\delta)!\ (t-n+\delta)!\ (\delta-1)!\ (N-t-\delta)!\ N!}$$

is derived in the usual manner. However the obvious estimate again is no longer strictly unbiased. For

$$(10)\qquad E\left(\frac{n}{n-\delta+1}\right) = E\left(\frac{n}{s+1}\right)$$

$$= \sum_{n}^{t+\delta}\left[\frac{n!\ (t+1)!(N+1-t-1)!(N+1-n-1)!}{(n-\delta+1)!(t+1-n-1+\delta)!(\delta-1)!(N-1-t+1-\delta)!(N+1)!}\right]\frac{N+1}{t+1}$$

$$= \frac{N+1}{t+1}\left(1 - \frac{(N+1-\delta)!\ (N-t)!}{(N+1)!\ (N-t-\delta)!}\right)$$

This result is analogous to that obtained by the author in [4]: by the same argument and formulae given there (pp. 145-146) it follows that the second term in the parentheses is negligible provided $\frac{t\delta}{N} > \log N$. For such values of δ, t, N the estimate

$$(11)\qquad \tilde{N} = \frac{n(t+1)}{s+1} - 1$$

has bias less than 1 in absolute value.

To determine the variance of $\tilde{N}$, $(s+1)^{-2}$ is expressed in an inverse factorial series as in [4] so that

$$(\tilde{N}+1)^2 = (t+1)^2\left[n(n+1)-n\right]\left(\frac{1}{(s+1)(s+2)} + \frac{1}{(s+1)(s+2)(s+3)} + \frac{2}{(s+1)(s+2)(s+3)(s+4)} + R\right)$$

To evaluate the expectation of this latter series let

$n_{(i)} = n(n-1)(n-2)\ldots(n-i+1)$ and observe that for any $i \leq j$ $\qquad (j-i=\varepsilon)$

$$(12)\qquad E\left(\frac{(n+i-1)_{(i)}}{(s+j)_{(j)}}\right) = \frac{(N+i)_i\ (N-t)_\varepsilon}{(t+j)_j\ (\delta-1)_\varepsilon}\ (1-\eta_j)$$

Here η_j stands for j terms of a form similar to the one in the brackets of formula (10). This formula is derived by a direct summation exactly parallel to that used in obtaining (10).

If η_j is neglected and the remaining terms on the right hand side of (12) are written q_{ij} then

$$(13)\quad \sigma^2_{\tilde{N}} = E(\tilde{N})^2 - N^2 = (t+1)^2\left[(q_{22}+q_{23}+2q_{24}+\ldots)-(q_{12}+q_{13}+2q_{14}+\ldots)\right]-2N-2-N^2$$

Table 1 gives a tabulation of $\frac{\sigma_{\tilde{N}}}{N}$ for various representative values of N, t and δ, obtained by means of this formula.

The average sample size required by this procedure can be determined simply from (9). In fact again by direct summation

$$(14)\quad E(n+i-1)_{(i)} = \frac{(\delta+i-1)_i\ (N+1)_i}{(N-t+i)_i}$$

so that

$$(15)\quad E(n) = \frac{\delta(N+1)}{(N-t+1)}$$

and

$$(16)\quad \sigma^2_n = \frac{\delta t(N+1)(N-\delta-t+1)}{(N-t+1)^2(N-t+2)} \doteq \frac{\delta t}{N-t}$$

Either of these formulae can be obtained from (6) and (8) by an interchange of N-t and t and by replacing s by δ. It is seen immediately that the tremendous variation possible in the earlier inverse sampling model is now eliminated.

Since δ will usually be reasonably large the most appropriate approximation to use for testing purposes is the normal distribution. In particular, it is desirable to work with n using formula (15) and (16). Writing

$\frac{t}{N+1} = p'$ and using a trivial modification of the approximation of (16)

$$\frac{\left(n - \frac{\delta}{1-p'}\right)}{\sqrt{\frac{\delta p'}{1-p'}}} \text{ is approximately } N(0,1)$$

and confidence limits for p' (and hence N) are obtained from the quadratic equation

$$(17) \qquad \left[n(1-p') - \delta\right]^2 = k_\alpha^2 \, \delta p'(1-p').$$

The approximation may be improved slightly if the exact formula for σ_n^2 is used but this involves solving a higher degree equation.

DIRECT MULTIPLE SAMPLE CENSUSES

In the most usual type of direct multiple sample census, at each stage a group of individuals are drawn from the population. Those that are tagged are noted; those not tagged are tagged and then the whole group returned to the population. This sampling may take place without replacement. Even where the sampling is with replacement it may be desirable to set up the experiment so that the model appropriate to sampling without replacement is correct. This may be done by ignoring recaptures in the same sampling period. It is frequently desirable to do so to avoid possible nonrandomness involved in such recaptures.

For this situation the appropriate model is given by the following conditions

a) $t_{i+1} - t_i = n_i - s_i \qquad i = 1, 2, \ldots, k$

b) $t_1 (= n_0), n_1, n_2, \ldots, n_k$ are parameters

c) $t_2, t_3, \ldots, t_k; s_1, s_2, \ldots, s_k$ are random variables

$$(18)\quad \Pr(s_1, s_2, \ldots, s_k) = \prod_{i=1}^{k} \frac{\binom{t_i}{s_i}\binom{N-t_i}{n_i-s_i}}{\binom{N}{n_i}}$$

$$= \left[\prod_{i=1}^{k} \frac{t_i!\, n_i!\, (N-n_i)!}{N!\, s_i!\, (t_i-s_i)!\, (n_i-s_i)!}\right] \frac{(N-n_0)!}{(N-t_{k+1})!}$$

The remaining terms telescope because of (a). From formula (18) it is apparent that $\sum_{i=1}^{k} s_i$ is a sufficient statistic for N. Moreover the maximum likelihood estimate of N is easily derived from (18) [see Appendix II]. It is the solution of the equation

$$(19)\quad \prod_{i=0}^{k} \left(1 - \frac{n_i}{N}\right) = 1 - \frac{t_{k+1}}{N}$$

In view of the fact that N is much larger than any n_i a first approximation to the maximum likelihood estimate is

$$(20)\quad N_L = \frac{\sum_{i=0}^{k} \sum_{j=i+1}^{k} n_i\, n_j}{\sum_{i=1}^{k} s_i}$$

This formula is obtained by ignoring terms involving $\frac{1}{N^3}$ or higher powers of $\frac{1}{N}$ in (19). If the cubic terms are retained the approximate equation to be solved for the maximum likelihood estimate is

$$(21)\quad N^2\left(\sum_{i=1}^{k} s_i\right) - N\left(\sum_{i=0}^{k} \sum_{j=i+1}^{k} n_i n_j\right) + \sum_{\alpha=0}^{k} \sum_{\beta=\alpha+1}^{k} \sum_{\gamma=\beta+1}^{k} n_\alpha\, n_\beta\, n_\gamma = 0$$

The larger root of this equation (which will be denoted N_q) is the desired one. As shown in Appendix II, under certain conditions that will be frequently met with in practice, the maximum likelihood estimate will lie between N_q and N_L and furthermore the difference $N_L - N_q$ will be relatively small compared to N.

It is interesting to note that the same estimate is obtained by the method of moments. For, if $\sum_{i=1}^{k} s_i$ is denoted S_k, then

$$(22)\quad E(S_k) = \prod_{i=0}^{k}\left(1-\frac{n_i}{N}\right) - \left(1-\frac{\sum_{i=0}^{k} n_i}{N}\right)$$

$$= \sum_{i=2}^{k+1} (-1)^i \frac{k_i}{N^{i-1}}$$

where k_i = sum of products of $n_0, n_1, n_2, \ldots, n_k$ taken i at a time with all subscripts different. Either of these results is most easily obtained by induction.

The distribution of S_k is difficult to obtain in any simple useable form, so that it is difficult to evaluate the small sample properties of these estimates. It may be noted that these estimates are not strictly comparable with those of Schumacher and Essmeyer, Schnabel and the author's own, referred to in the introduction. These latter estimates which are generally of a χ^2 type are derived on the assumption (implicitly or explicitly) that the number of tags is a fixed parameter, not a random variable.

It is possible to write down simply an unbiased estimate which is valid in both of these models (i.e. whether the t_i are random variables or parameters) viz

$$(23)\quad N^* = \frac{1}{k}\sum_{i=1}^{k}\left[\frac{(n_i+1)(t_i+1)}{s_i+1} - 1\right]$$

provided the s_i (or more precisely $\frac{n_i t_i}{N}$) are sufficiently large [roughly, near 10 or greater in size]. For

$$E(N^*) = \frac{1}{k}\sum_{i=1}^{k} E_{t_i}\left[E\left[\frac{(n_i+1)(t_i+1)}{s_i+1} - 1 \;\middle|\; t_i\right]\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k} E_{t_i}[N] = N$$

under the restrictions noted (those given in [4] p. 145).

As a consequence of a theorem of Blackwell [2], however, it can be said that an unbiased estimate for N based on S_k has a smaller variance than N^*, within the model in which the t_i are cumulative and $t_{i+1} - t_i = n_i - s_i$.

Where the t_i are not random variables but fixed parameters, N^* is a desirable estimate (subject to the restrictions $\frac{n_i t_i}{N}$ not too small) and furthermore the variance of N^* may be computed from formula 33 of [4] and the fact that

$$(24) \quad \sigma^2_{N^*} = \frac{1}{k^2} \sum_{i=1}^{k} \sigma^2_{\frac{(n_i+1)(t_i+1)}{s_i+1}}$$

It will frequently happen that the $\frac{n_i t_i}{N}$ are too small to permit N^* being an unbiased estimate. For this situation, where the t_i are fixed, a slight modification of an estimate given by Schnabel [17] will be most useful viz

$$(25) \quad N_p = \frac{\sum_{i=1}^{k} n_i t_i}{\left(\sum_{i=1}^{k} s_i\right) + 1}$$

This is based on the fact that each s_i has approximately a Poisson distribution with parameter $\frac{n_i t_i}{N}$, and hence $\sum_{i=1}^{k} s$ has approximately a Poisson distribution with parameter $\frac{\sum_{i=1}^{k} n_i t_i}{N} = \lambda$ (say).

$E(N_p) = N(1 - e^{-\lambda})$ is easily derived. Furthermore $Ne^{-\lambda}$ will be certainly negligible if $\sum_{i=1}^{k} n_i t_i > N(\log N)$. This is a

much lesser restriction than that the same inequality hold for each $n_i t_i$.

Neglecting terms of the form $\lambda^j e^{-\lambda}$ ($j=0, 1, 2, 3$) an approximate formula for the standard deviation of N_p is derived, viz.,

$$(26) \qquad \sigma_{N_p} = \frac{N}{\sqrt{\sum_{i=1}^{k} n_i t_i}} \left(1 + \frac{2}{\lambda} + \frac{6}{\lambda^2}\right)^{\frac{1}{2}}$$

The derivation is exactly parallel to the derivation of (33) in [4], through the use of an inverse factorial series.

Many sample censuses will not fall strictly into either of these models: some of the unmarked individuals taken in a sample will be returned as marked members, but not all. A full treatment of this case is clearly complicated. Where the $\frac{n_i t_i}{n}$ are small, the estimate N_p is probably quite satisfactory.

The Poisson approximation will also be useful as a basis of constructing tests and confidence intervals for N. The confidence limits given in [3] may clearly be used (with nt replaced by $\sum_{i=1}^{k} \frac{n_i t_i}{N}$) if the t_i are fixed parameters. Within the model primarily considered in this section this will probably be reasonably satisfactory also.

If the experimenter observes s_{ij} the number of tags recovered in sample i that were placed in the j-th tagging rather than simply s_i ($s_i = \sum_{j=0}^{i-1} s_{ij}$) more information is apparently available. Actually the knowledge of s_{ij} is of no value in estimating N, or in constructing tests for N. For the joint probability of the s_{ij} is

$$(27) \qquad \Pr(s_{10}, s_{20}, s_{21}, \ldots, s_{k,k-1}) = \prod_{i=1}^{k} \frac{\prod_{j=1}^{i} \binom{n_j - s_j}{s_{ij}} \binom{N - t_i}{n_i - s_i}}{\binom{N}{n_i}}$$

$$= \left[\prod_{i=1}^{k} \left(\prod_{j=1}^{i} \binom{n_j - s_j}{s_{ij}}\right) \frac{1}{(n_i - s_i)! \binom{N}{n_i}}\right] \frac{(N - n_0)!}{(N - t_{k+1})!}$$

From the form of this probability distribution it is seen that

$S_k = \sum_{i=1}^{k} s_i = \sum_{i=1}^{k} \sum_{j=0}^{i-1} s_{ij}$ is still a sufficient statistic for N.

However the s_{ij} may be used to test the assumptions that the sampling is random (i.e. the probability of including an animal in the sample is independent of its being tagged) or that the population is constant.

Under the hypothesis tested, that N is constant, given t_j (j = 0, 1, 2,...,i-1) the s_{ij} have a multihypergeometric distribution or neglecting the sampling without replacement a multinomial distribution. However the usual χ^2 test does not follow immediately for the estimate of N will be determined from the random variables s_{ij} which are not independent for different i. The The author, however, conjectures that in view of the nature of the χ^2 test this is asymptotically negligible. Much more serious is the fact that in many cases at least, $E(s_{ij})$ are too small for a reasonable application of the test.

As an alternative procedure a non-parametric test is suggested. An array of the form

$$
\begin{array}{llll}
\dfrac{s_{10}}{n_1 n_0}, & \dfrac{s_{20}}{n_2 n_0} & \dfrac{s_{30}}{n_3 n_0}, \quad \ldots, & \dfrac{s_{k0}}{n_k n_0} \\
 & \dfrac{s_{21}}{n_2(n_1-s_1)}, & \dfrac{s_{31}}{n_3(n_1-s_1)}, \quad \ldots, & \dfrac{s_{k1}}{n_k(n_1-s_1)} \\
 & \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\
 & & & \dfrac{s_{kk-1}}{n_k(n_{k-1}-s_{k-1})}
\end{array}
\tag{23}
$$

may be formed, in which each element is a random variable with expectation $\frac{1}{N}$.

Within each row the random variables are independent. Between rows this is not the case; for the s_{ij}, for fixed i, are observations from the same sample. However the correlation between any two s_{ij}, say $s_{i\alpha}$, $s_{i\beta}$ is

$$-\frac{t_{i\alpha}}{N}\,\frac{t_{i\beta}}{N}\,\frac{N-n_i}{N-1}$$

which will be usually negligible. Moreover if $\frac{n_i(n_\alpha - s_\alpha)}{N}$ is small compared to n_i the conditional probability of $s_{i\alpha}$ given $s_{i\beta}$ is almost equal to the unconditional probability of $s_{i\alpha}$.

In view of this the sign test suggested by Moore and Wallis [12] to test for randomness in a sequence of independent observations from a common distribution may be appropriate. The test is based on the statistic D, the number of negative signs in the sequence of successive differences of observations. Moore and Wallis tabulated the probability distribution of D for small values of n, the number of observations. They conjectured and Mann [11] subsequently proved rigorously, that D is asymptotically normally distributed. In application of the test since

$$E(D) = \frac{n-1}{2} \qquad \text{and} \qquad \sigma_D^2 = \frac{n+1}{12}$$

$\left(\frac{D - \frac{n-1}{2}}{\sqrt{n+1}}\right)\sqrt{12}$ is taken to be distributed according to N(0, 1) for $n \geq 12$. If the array (28) is considered as a single sequence of observations the n of the Moore and Wallis test is here equal to $\frac{k(k+1)}{2}$.

In many cases the alternatives to randomness are essentially one-sided. For example, some that may be considered are:

(a) the tagged individuals die off more rapidly or disappear so as not to be available for sampling

(b) the tagged individuals disperse from the tagging location slowly and are more likely to be recaptured in the samples taken soon after the tagging rather than later

(c) there is a marked change in the composition of the population due to natural processes or to migration.

If any of these alternatives is true, the random variables in each row of the array will tend to increase. In this case a test based on the whole array as a single sequence has the following defect: if the alternatives are true, in each row the probability of a negative difference is greater than $\frac{1}{2}$, but the probability of a negative difference between the last element of any row and the first of the next row will be much less than $\frac{1}{2}$.

To avoid this it is necessary to consider each row separately, i.e., the array (28) may be considered as k sequences of observations decreasing in length from k to 1. A test of randomness may be made using the sum of the number of negative differences in these k sequences (actually k-1 since no difference is obtainable from the last row).

Let

$$(29) \qquad X = D_1 + D_2 + \ldots + D_{k-1}$$

where D_i = number of negative differences in row i.

Then $E(X) = \frac{k(k-1)}{4}$ and $\sigma_X^2 = \frac{(k+4)(k-1)}{4}$.

The asymptotic normality of X is immediate using the fact that D_{k-1} itself tends to be normally distributed as k tends to infinity while the initial terms of the sum (29) are asymptotically negligible as k becomes large.

Table II gives a partial tabulation of the distribution of X for $k = 4, 5, 6, 7$.

INVERSE MULTIPLE SAMPLE CENSUSES

As in the single stage census two models may be considered. In the first of these, sampling and tagging are carried out in exactly the same manner as in the direct multiple sample census, except that the number of tagged individuals at each stage is predetermined, rather than the sample size.

In this model

a) $t_{i+1} - t_i = n_i - s_i \quad i = 1, 2, \ldots, k$

b) $t_1 (= n_0)$, s_1, $s_2, \ldots, s_k$ are parameters

c) $t_2, t_3, \ldots, t_k, n_1, n_2, \ldots, n_k$ are random variables,

and

$$\Pr(n_1, n_2, \ldots, n_k) = \prod_{i=1}^{k} \frac{((n_i-1)!\ t_i!(N-t_i)!(N-n_i)!}{(s_i-1)!\,(t_i-s_i)!\,(n_i-s_i)!\,N!\,(N-n_i-t_i+s_i)!}$$

There is now no non-trivial sufficient statistic for N. However, an unbiased estimate is easily found, namely

$$(30) \qquad \hat{N} = \frac{1}{k} \sum_{i=1}^{k} \frac{n_i(t_i+1)}{s_i} - 1$$

for

$$(31) \qquad E(\hat{N}) = \frac{1}{k} \sum_{i=1}^{k} \underset{\substack{n_j \\ j<i}}{E} \left[\underset{n_i}{E} \left(\frac{n_i(t_i+1)}{s_i} - 1 \,\middle|\, n_j,\ j<i \right) \right]$$

$$= \frac{1}{k} \sum_{i=1}^{k} \underset{\substack{n_j \\ j<i}}{E} [N] = N.$$

Using the approximate formula for the variance of $\hat{N}$ in the single sample case, (8), by a similar procedure to the derivation of (31) it is found that

$$(32) \qquad \sigma^2_{\hat{N}} \doteq \frac{N^2}{k^2} \sum_{i=1}^{k} \frac{1}{s_i} .$$

In this inverse sampling procedure there are several parameters at the disposal of the experimenter - n_0, k and $s_1, s_2, \ldots, s_k$. It may be desirable to choose them so as to minimize $E\left(\sum_{i=0}^{k} n_i\right)$ while holding $\frac{\sigma \hat{N}}{N}$ fixed. This would achieve a fixed precision of estimation while minimizing the effort expended. However after arbitrarily fixing n_0 and k, the optimum choice of the s_i necessitates the solution of an algebraic equation of high degree (the degree increases rapidly with k) which involves N in a complex fashion.

In lieu of general rules Table III gives the properties of a number of simple designs. The approximate formula (32) was used to calculate $\frac{\sigma \hat{N}}{N}$. $E(n_i)$ was calculated recursively from the formula

$$E(n_i) \doteq \frac{(N+1)s_i}{\sum_{j=1}^{i-1} E(n_j) - s_j + n_0 + 1}$$

In order that this give a reasonable approximation it is necessary that s_1 be not too small. For this reason some of the more interesting cases with the initial s_i very small are excluded.

In the second model dealing with the inverse sampling procedure, at each stage $n_i - s_i$ is predetermined. As before we write $n_i - s_i = \delta_i$. Thus

a) $t_{i+1} - t_i = \delta_i \quad (i = 0, 1, \ldots, k)$

b) $s_1, s_2, \ldots, s_k; n_1, n_2, \ldots, n_k$ are random variables.

Since t_i is a parameter at each stage of the sampling procedure there is now independence between the successive random variables. Consequently the results determined in the single sample case are easily generalized to this model. Under the restrictions on n and δ that (11) hold, it follows that

$$\bar{\bar{N}} = \frac{1}{k} \sum_{i=1}^{k} \frac{n_i(t_i + 1)}{s_i + 1} - 1 \tag{33}$$

is an almost unbiased estimate. Also the variance is determined from (13) and the usual formula for the sum of the variances. Similarly

$$(34) \qquad E(n) = \sum_{i=1}^{k} \frac{\delta_i (N+1)}{N - t_i + 1}$$

is derived from (15).

It may be mentioned that a maximum likelihood estimate may be derived for this model which is analogous to that given by formula (19). For again a telescoping occurs in the formula for the joint probabilities of the random variables viz.

$$Pr(n_1, n_2, \ldots, n_k) = \prod_{i=1}^{k} \left[\frac{(n_i - 1)!(t_i - 1)!(N - n_i)!}{(n_i - \delta_i)!(t_i - n_i + \delta_i)!(\delta_i - 1)!N!} \right] \left[\frac{N - t_1}{N - t_{k+1}} \right]$$

The maximum likelihood estimate of N is the solution of the equation

$$(35) \qquad \prod_{i=0}^{k} \left(1 - \frac{n_i}{N} \right) = \left(1 - \frac{t_{k+1}}{N} \right)$$

where we write n_0 for t_1. In this case there is no simple sufficient statistic: the maximum likelihood estimate is a function of all the n_i. Since this is so and since the solution of (35) has a complicated distribution, the estimate $\bar{N}$ seems more desirable.

SEQUENTIAL TESTS FOR N

The optimum sample census procedures evidently are sequential: furthermore the sequential procedure should permit a choice of the design at any stage rather than merely a choice of whether or not to take further observations. For simplicity only standard sequential procedures are considered here, primarily in relation to tests for N. Such tests may be useful in control and management problems.

For direct sample censuses with $E(s_i)$ small we consider n_i as preassigned and let k the number of stages in the census be a random variable. Since the s have approximately a Poisson distribution, Wald's

theory [9] is applicable in a routine manner.

In particular to test the hypothesis H: $N \leq N_0$ against the alternatives $N > N_0$ at a level of significance α, so that the power at a particular alternative N_1 is $1 - \beta$ we proceed as follows:

at any stage j after $n_1, n_2, \ldots, n_j$ individuals have been sampled and $s_1, s_2, \ldots, s_j$ tags recovered

$$(36)\quad \text{Reject H if } \sum_{i=1}^{j} s_i \geq \frac{\log A - \sum_{i=1}^{j} n_i t_i \left(\frac{1}{N_0} - \frac{1}{N_1}\right)}{\log N_0 - \log N_1}$$

$$(37)\quad \text{Accept H if } \sum_{i=1}^{j} s_i \leq \frac{\log B - \sum_{i=1}^{j} n_i t_i \left(\frac{1}{N_0} - \frac{1}{N_1}\right)}{\log N_0 - \log N_1}$$

Continue observations if $\sum_{i=1}^{j} s_i$ lies between the bounds in (36) and (37).

Here $A = \frac{1-\beta}{\alpha}$ $\qquad B = \frac{\beta}{1-\alpha}$.

The bounds in (36) and (37) are themselves random variables: this is so because the s_i are dependent in the model considered. However the conditional distribution of s_j given $s_i (i < j)$ is of the Poisson form, approximately, with parameter $\frac{n_j t_j}{N}$. The setting up of the sequential probability ratio test does not require independence of the observations: it depends on the conditional distribution of the observed random variable at each stage. However the various optimum properties and associated results that Wald and others have proved for the sequential probability ratio test are not necessarily valid without independence. However approximate formulae could be determined for most situations from the formulae for the operating characteristic curve and for E(n) in the standard Poisson situation. Some of these may be found in [8].

In conclusion it may be reiterated that the sample censuses studied here have been limited to situations where the population is essentially stationary. This excludes populations where any substantial immigration or emigration occurs. It does not necessarily exclude populations which are changing due to natural causes e.g. birth and death. If those born subsequently to the initial tagging operation can be distinguished from the rest of the population, without undue effort and if the death rate is the same among the tagged and untagged groups, then the sample census yields valid information on the population size at the time of first tagging.

For suppose the probability of survival from time of tagging to time of sampling is p. Now consider the almost unbiased estimate in the direct single sample census model, where at the time of sampling t', $(N-t)' = u'$, N' actually are surviving.

$$\text{Then } E\left[\frac{(n+1)(t+1)}{s+1} - 1 \,\middle|\, t',\ u'\right] = (N'+1)\left(\frac{t+1}{t'+1}\right) - 1$$

$$= \left[\frac{(t'+1)\,u'}{t+1}\right](t+1) - 1$$

$$\left[1 \quad \frac{u'}{t'+1}\right](t+1) - 1$$

Since t', u' are independent and since it is reasonable to assume they have a binomial distribution

$$(38) \quad E\left[\frac{(n+1)(t+1)}{s+1} - 1\right] = \left[1 + \frac{(N-t)p}{(t+1)p}\right](t+1) - 1$$

$$= N$$

The denominator in the right hand side of (38) comes from the fact that, by direct summation

$$E\left(\frac{1}{t'+1}\right) = \frac{1}{(t+1)p}\left[1 - (1-p)^{t+1}\right] = \frac{1}{(t+1)p} \quad \text{for large t, and}$$

moderate p.

Similarly the unbiased estimate $\hat{N} = \frac{nt}{s}$ of the inverse sample census is unbiased whether or not mortality occurs in the population (provided tagged and untagged are proportionally affected). Even the variance of these estimates is only slightly modified unless the mortality is excessive. For example, the approximate formula (3) now becomes

$$\sigma^2_{\hat{N}} = \frac{N^2}{s} + N\left(\frac{1-p}{p}\right)\left(\frac{1}{s} + 1\right)$$

which is approximately $\frac{N^2}{s}$ unless p is very small.

However if there is natural mortality the multiple sample census may be seriously affected--estimates of N in such a case may be meaningless. In fact this forms a basis for the possible estimation of natural mortality in an animal population--a fact utilized by Leslie and Chitty in a recent paper [10].

The references cited below are those referred to in the body or appendix of the paper. No attempt has been made to compile a complete bibliography, even of the methodology in this field. In fact nowhere is such a bibliography available apparently; however the monographs of Kicker [14] and Schaeffer [16] give a large number of useful references.

References

1. Norman T. J. Bailey, "On estimating the size of mobile populations from recapture data", Biometrika 38 (1951) pp. 292-306.

2. D. Blackwell, "Conditional Expectation and Unbiased Sequential Estimation", Annals of Math. Stat. 18 (1947) p. 105-110.

3. D. G. Chapman, "A Mathematical Study of Confidence Limits of Salmon Populations Calculated from sample tag ratios", Internat. Pac. Salmon Fisheries Comm. Bulletin #2 1948 pp. 69-85.

4. —, "Some properties of the hypergeometric distribution with applications to zoological sample censuses", University of California Publications in Statistics 1 (1951) pp. 131-160.

5. H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946.

6. F. N. David, "Limiting distributions connected with certain methods of sampling human populations", Stat. Res. Mem. II (1938) pp. 69-90.

7. D. B. DeLury, "On the estimation of biological populations", Biometrics 3 (1947) pp. 145-167.

7a.—, "On the planning of experiments for the estimation of fish populations", I. Fish. Res. Bd. Can., 8 (1951) pp. 281-307.

8. L. H. Herbach, "Bounds for some functions used in sequentially testing the mean of a Poisson distribution", Annals of Math. Stat. 19 (1948) pp. 400-405.

9. P. S. Laplace, "Sur les naissances, les mariages et les morts", Histoire de l'Academie Royale des Sciences Annee 1783, p. 693, Paris (actually published in 1786).

10. P. H. Leslie and Dennis Chitty, "The estimation of population parameters from data obtained by capture-recapture method. I The maximum likelihood equations for estimating the death rate", Biometrika 38 (1951) pp. 269-292.

11. H. B. Mann, "On a test for randomness based on signs of differences", Annals of Math. Stat. 16 (1945) pp. 193-199

12. G. Moore and W. Allen Wallis, "Time series significance tests based on signs of differences", Jour. Amer. Stat. Assoc. 38 (1943) pp. 143—159

13. C. G. J. Petersen, The yearly immigration of young plaice into the Limfjord from the German Sea, etc. Rept. Danish Biol. Sta. for 1895 6 (1896) pp. 1-48.

14. W. E. Ricker, Methods of Estimating Vital Statistics of Fish Populations, Indiana University Publications, 1948.

15. M. Sandelius, "An inverse sampling procedure for bacterial plate counts", Biometrics 6 (1950) pp. 291-2.

16. M. B. Schaeffer, "A study of the spawning populations of sockeye salmon in the Harrison river system, with special reference to the problem of enumeration by means of marked numbers", Internat. Pac. Salmon Fisheries Comm. Bulletin #4 (1951) pp. 1-207.

17. Z. E. Schnabel, "Estimation of the total fish population of a lake", Amer. Math. Monthly 45 (1938) pp. 348-352.

18. F. X. Schumacher and R. W. Eschmeyer, "The estimate of fish populations in lakes or ponds", Jour. Tenn. Acad. Sci. 18 (1943) pp. 228-249.

19. A. Wald, Sequential Analysis, Wiley, New York, (1947).

Appendix

I To show the asymptotic normality of z as defined in formula (3) consider the moment generating function of z:

$$M_z(\theta) = p^s e^{\frac{ps}{\sqrt{s(1-p)}}} e^{-\sqrt{\frac{s\theta}{1-p}}} \left[1 - (1-p)e^{\frac{p}{\sqrt{s(1-p)}}}\right]^{-s}$$

By routine algebra and the usual manipulations it may be seen that $M_z(\theta)$ tends to $\frac{\theta^2}{2}$ as s tends to infinity (e.g. cf. Cramer 5 pp. 198-199).

On the other hand the moment generating function of the random variable 2np is

$$M_{2np}(\theta) = p^s e^{2ps\theta} \left[1 - (1-p)e^{2p\theta}\right]^{-s}$$

and $\ln M_{2np}(\theta) = 2ps\theta - s \ln \left[1 - 2\theta \quad h(p)\right]$

where $h(p) \to 0$ as $p \to 0$.

$$\lim_{p \to 0} M_{2np}(\theta) = (1 - 2\theta)^{-s}$$

which is the moment generating function of the χ^2 distribution with 2s degrees of freedom.

II Maximum Likelihood Equation.

The maximum likelihood estimate is derived by setting the ratio

$$(39) \quad q(n) = \frac{Pr(s_1, s_2, \ldots, s_k; N)}{Pr(s_1, s_2, \ldots, s_k; N-1)} = 1$$

Now

$$(40) \quad q(N) = \prod_{i=1}^{k} q_i(N) \quad \text{where}$$

$$(41) \quad q_i(N) = \frac{(N-n_i)\ (N-t_i)}{N\ (N-n_i-t_i+s_i)}$$

and also

$$(42) \qquad q(N) = \frac{1 - \frac{R_1}{N} + \frac{R_2}{N^2} - \frac{R_3}{N^3} \ldots (-1)^k \frac{R_k}{N^k}}{1 + \frac{S_k}{N} - \frac{R_1}{N}}$$

$$= \frac{1 + \frac{\frac{R_2}{N^2} - \frac{R_3}{N^3} \ldots (-1)^k \frac{R_k}{N^k}}{1 - \frac{R_1}{N}}}{1 + \frac{\frac{S_k}{N}}{1 - \frac{R_1}{N}}}$$

so that $q(N) \lesseqgtr 1$ according as

$$(43) \qquad S_k \gtreqless \frac{R_2}{N} - \frac{R_3}{N2} + \ldots (-1)^k \frac{R_k}{N^{k-1}}$$

also

$$(44) \qquad q_i(N) \lesseqgtr 1 \quad \text{according as } Ns_i \gtreqless n_i t_i$$

$$\text{Let } m = \min_i \frac{n_i t_i}{s_i} \qquad M = \max_i \frac{n_i t_i}{s_i}$$

In view of (40) and (44) the roots of (19) must lie between m and M. If one or more s_i are zero M will not be finite. However unless all s_i are zero (and hence $S_k = 0$) it is evident from the form of equation (43) that $q(N) < 1$ for sufficiently large N. $S_k = 0$ may be neglected since $Pr(S_k = 0)$ will be extremely small.

It is possible that equation (19) has several real roots in the interval (m, M). Those for which $q(N)$ is decreasing represent local maxima and one of these the extreme maximum. The large number of parameters gives rise to a diverse number of possibilities but since in general $N >> \sum_{i=1}^{k} n_i$ the following theorem will cover many cases that may arise

Theorem. If for $N > m$, $\frac{R_j}{N^{j-1}}$ is a monotone decreasing sequence and

if $\frac{\sum_{i=1}^{k} n_j}{S_k} < m$, $R_3 < \left(\frac{R_2}{2}\right)^2$, then the maximum likelihood estimate N^* satisfies the following inequalities

$$N_q < N^* < N_L$$

Proof.

$$\frac{\sum_{i=1}^{k} n_i t_i}{\sum_{i=1}^{k} s_i} > m$$

since the left hand side is a weighted harmonic mean of terms $\frac{n_i t_i}{s_i}$

and hence larger than the smallest term

$$\therefore N_L = \frac{R_2}{S_k} > \frac{\sum_{i=1}^{k} n_i t_i}{S_k} > m$$

For $N > N_L$, $S_k > \frac{R_2}{N}$ and $-\frac{R_2}{N^2} + \ldots - (-1)^k \frac{R_k}{N^{k-1}} < 0$

so that $N_L > N^*$

Since $R_3 > 0$ $\quad N_q < N_L$

If $N_q < m$ then $N_q < N^*$. Consider the contrary case $N_q > m$ and denote the other root of the quadratic equation (21) by N_q'.

$$N_q' = \frac{R_2}{2S_k}\left(1 - \sqrt{1 - \frac{4R_3}{R_2^2}}\right) < \frac{2R_2}{R_2 S_k} < \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} s_i} < m$$

[the first inequality follows from the fact that

$$1 - (1-X)^{\frac{1}{2}} < X \qquad 0 \leq X \leq 1 \quad]$$

Hence for $m < N < N_q$

$$S_k - \frac{K_2}{N} + \frac{K_3}{N^2} < 0 \qquad \text{and} \qquad \frac{K_4}{N^3} - \ldots (-1)^k \frac{K_k}{N^{k-1}} > 0$$

$\therefore N_q < N^*$.

Moreover $N_L - N_q = \frac{K_2}{2S_k}\left(1 - \sqrt{1 - \frac{4K_3}{K_2^2}}\right) < \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \varepsilon_i}$

provided only $K_3 < \left(\frac{K_2}{2}\right)^2$. The right hand side of this equation is approximately of the order $\frac{N}{n_i}$.

Table I

Numerical Comparison of Certain Inverse Sample Censuses with s (number of tagged individuals to be recaptured) fixed and with δ (number of untagged individuals to be recaptured) fixed.

N	t	δ	s	E(n)[1]	$\sigma_{n/\delta}$[2]	$\sigma_{n/s}$[3]	$\sigma_{\frac{\hat{N}}{N}}$[4] exact	$\sigma_{\frac{\hat{N}}{N}}$[5] approximate	$\sigma_{\frac{\tilde{N}}{N}}$[6]
10^4	100	990	10	1000	3.2	316.2	.30	.32	.33
	100	4950	50	5000	7.1	707.1	.10	.14	.10
	500	475	25	500	5.0	100.0	.19	.20	.20
10^5	100	4995	5	5000	2.2	2236.1	.43	.45	.55
	500	4975	25	5000	5.0	1000.0	.19	.20	.20
10^6	1000	9990	10	10,000	3.2	3162.3	.31	.32	.35
	1000	19,980	20	20,000	4.5	4472.1	.22	.22	.23

1. δ and s were chosen so that $E(n)$ is the same (to the nearest integer) for both sampling plans, i.e., predetermining s or predetermining δ.
2. Calculated from the approximate formula $\sigma_n = \sqrt{\frac{\delta t}{N-t}}$
3. Calculated from the approximate formula $\sigma_n = N\sqrt{\frac{s}{t}}$
4. Calculated from formula (8) for the estimate where s is predetermined.
5. $\sigma_{\frac{\hat{N}}{N}} \doteq \sqrt{\frac{1}{s}}$ (the approximation to formula (8)).
6. Calculated from formula (13) for the estimate where δ is predetermined.

Table II

Cumulative Distribution of X for values of k=4, 5, 6, 7.

$P(X \leq x)$

	x	0	1	2	3	4	5	6	7
4		0.0035	0.0590	0.3056	0.6944	0.9410	0.9965	1.0000	1.0000
5		--	0.0012	0.0172	0.1052	0.3392	0.6608	0.8948	0.9828
6		--	--	0.0001	0.0020	0.0166	0.0627	0.2010	--
7		--	--	--	--	0.0001	0.0009	0.0323	0.1103

Table III

Standard Error of the Estimate of N and Total Effort Required by Various Multiple Inverse Sample Censuses.

k	s_1, s_2, s_3, s_4, s_5	n_0	$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{s_i}$	N = 10,000 E(n')*	N = 10,000 $\frac{\sigma_{\hat{N}}}{E(n')}$	N = 100,000 E(n')*	N = 100,000 $\frac{\sigma_{\hat{N}}}{E(n')}$
3	5, 10, 15, -- --	224	0.202	901	2.24	3385	0.60
3	10, 10, 10, -- --	317	0.183	923	1.98	4020	0.46
3	15, 10, 5, -- --	388	0.202	963	2.10	4595	0.44
4	5, 10, 15, 20, --	224	0.162	1131	1.43	3981	0.41
4	12, 12, 13, 13, --	347	0.142	1155	1.23	4728	0.30
4	20, 15, 10, 5, --	448	0.162	1208	1.34	5499	0.29
5	5, 10, 15, 20, 25	224	0.135	1363	0.99	4617	0.29
5	15, 15, 15, 15, 15	388	0.116	1394	0.83	5527	0.21
5	25, 20, 15, 10, 5	500	0.135	1452	0.93	6363	0.21
5	11, 13, 15, 17, 19	332	0.118	1377	0.86	5170	0.23
4	18, 19, 19, 19, --	425	0.116	1418	0.82	5794	0.20
4	7, 15, 23, 30, --	265	0.134	1389	0.96	4335	0.28
3	25, 25, 25, -- --	500	0.115	1464	0.79	6373	0.18
3	15, 25, 35, -- --	388	0.122	1425	0.86	5566	0.22
2	37, 38, -- -- --	609	0.115	1539	0.75	7248	0.16
2	25, 50, -- -- --	500	0.122	1513	0.81	6406	0.19
1	75, -- -- -- --	867	0.114	1732	0.66	9508	0.12

* n' = total number of individuals sampled including those marked or tagged initially i.e. n_0 .

Laboratory of Statistics
Department of Mathematics
University of Washington

Technical Reports Distribution List
Contract N8onr-520 Task II

Chief of Naval Research
Office of Naval Research
Washington 25, D. C.
Attn: Code 432
(Mathematics Branch) 5

Dr. E. Paulson, Head
Statistics Branch
Office of Naval Research
Department of the Navy
Washington 25, D. C. 5

Scientific Section
Office of Naval Research
Department of the Navy
1000 Geary Street
San Francisco 9, Calif. 2
Attn: Dr. J. Wilkes 1

Director, Naval Research
Laboratory
Washington 25, D. C.
Attn: Technical Information
Officer 9

Office of the Assistant Naval
Attache for Research
Naval Attache
American Embassy
Navy No. 100
Fleet Post Office
New York, N. Y. 2

Headquarters, USAF
Director of Research and
Development
Washington 25, D. C. 1

Director
Office of Naval Research
Branch Office
844 North Rush Street
Chicago 11, Illinois 1

Director
Office of Naval Research
Branch Office
346 Broadway
New York 13, N. Y. 1

Director
Office of Naval Research
Branch Office
1030 East Green Street
Pasadena 1, California 1

Chairman
Research and Development Board
The Pentagon
Washington 25, D. C. 1

Commander
U. S. Naval Ordnance Test Station
Inyokern, China Lake, Calif. 1

Chief of Naval Operations
Operation Evaluation Group-OP 374
The Pentagon
Washington 25, D. C. 1

Office of Naval Research
Department of the Navy
Washington 25, D. C.
Attn: Code 438
(Mechanics Branch) 2

Director
Office of Naval Research
Branch Office
495 Summer Street
Boston 10, Mass. 1

Professor Carl B. Allendoerfer
Department of Mathematics
University of Washington
Seattle 5, Washington 1

National Bureau of Standards
Institute for Numerical Analysis
405 Hilgard Avenue
Los Angeles 24, California 2

Professor W. Allen Wallis
Committee on Statistics
University of Chicago
Chicago 37, Illinois 1

Director, Applied Mathematics
and Statistics Laboratory
Stanford University
Stanford, California 3

Chief, Statistical Engineering
Laboratory
National Bureau of Standards
Washington 25, D. C. 1

RAND Corporation
1500 Fourth Street
Santa Monica, California 1

Professor J. Neyman
Statistics Laboratory
University of California
Berkeley, California 2

Assistant Chief of Staff, G-4
for Research and Development
U. S. Army
Washington 25, D. C. 1

Professor Herbert Solomon
Teachers College
Columbia University
New York, N. Y. 1