AD NUMBER: AD0038816
LIMITATION CHANGES
TO: Approved for public release; distribution is unlimited.
FROM: Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; 25 Aug 1954. Other requests shall be referred to Office of Naval Research, Arlington, VA 22203.
AUTHORITY
ONR LTR, 26 OCT 1977

Armed Services Technical Information Agency

Because of our limited supply, you are requested to return this copy WHEN IT HAS SERVED YOUR PURPOSE so that it may be made available to other requesters. Your cooperation will be appreciated.

AD 388166

NOTICE: WHEN GOVERNMENT OR OTHER DRAWINGS, SPECIFICATIONS OR OTHER DATA ARE USED FOR ANY PURPOSE OTHER THAN IN CONNECTION WITH A DEFINITELY RELATED GOVERNMENT PROCUREMENT OPERATION, THE U. S. GOVERNMENT THEREBY INCURS NO RESPONSIBILITY, NOR ANY OBLIGATION WHATSOEVER; AND THE FACT THAT THE GOVERNMENT MAY HAVE FORMULATED, FURNISHED, OR IN ANY WAY SUPPLIED THE SAID DRAWINGS, SPECIFICATIONS, OR OTHER DATA IS NOT TO BE REGARDED BY IMPLICATION OR OTHERWISE AS IN ANY MANNER LICENSING THE HOLDER OR ANY OTHER PERSON OR CORPORATION, OR CONVEYING ANY RIGHTS OR PERMISSION TO MANUFACTURE, USE OR SELL ANY PATENTED INVENTION THAT MAY IN ANY WAY BE RELATED THERETO.

Reproduced by

DOCUMENT SERVICE CENTER

KNOTT BUILDING, DAYTON, 2, OHIO

UNCLASSIFIED

THE ESTIMATION OF THE SIZE OF A STRATIFIED POPULATION

By

D. G. Chapman and C. O. Junge, Jr.

University of Washington
and
Washington State Department of Fisheries

Technical Report No. 16

August 25, 1954

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

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

1. Summary The frequently used tag-sample procedure of estimating the size of a biological population may be invalid if the population has variable stratification. In this paper conditions are given under which the usual estimates, as well as one proposed by Schaefer [1] for this situation, are valid. Also a new estimate is proposed and studied.

2. Introduction and Notation A common method of estimating the size of mobile animal populations is based upon sampling the population, after a known number are marked or tagged. The mathematical theory of estimates based on this procedure have been widely studied--numerous references are given in [2]. It is only important to note here that all of the theory is based upon the assumption that the sample is random with respect to the marked members of the population.

If however the population is stratified, by time or by area, for example, there is no a priori reason why this assumption should be valid. For example the population may be migrating through a river system with a time lapse between marking and sampling. Thus the probability that a member of the population is marked may depend on the tagging rate and on the migration rate. This type of situation was considered by Schaefer [1] and he set up an estimates for the total population size. An example involving stratification by area has been noted in [3]. Where populations are stratified by areas with partial mixing occurring between areas with time, the mixing or migration rate may be of interest as well as the total population size. The procedure outlined below yields estimates of the population movement as well as the total population size.

The following notation is required:

N_{ij} = number of individuals that are in stratum i at the time of tagging and in stratum j at time of sampling

t_{ij} = number of tagged individuals in stratum i at tagging time and in stratum j at sampling time

n_{ij} = number of sampled individuals in stratum i at tagging time and in stratum j at sampling time

s_{ij} = number of tagged individuals, tagged in stratum i and subsequently recovered in a sample from stratum j

$(i,j = 1,\ldots,r)$

Summation over any subscript will be denoted by replacing the subscript by a dot. Thus

$t_{i.} = \sum_{j=1}^{r} t_{ij}$ number of tags put out in stratum i

$n_{.j} = \sum_{i=1}^{r} n_{ij}$ number sampled in stratum j.

The N_{ij}, $N_{i.}$, $N_{.j}$ and $N_{..}$ (the total population size) are regarded as unknown parameters, the $t_{i.}$, $n_{.j}$ as known parameters, the t_{ij} and n_{ij} are unobservable random variables while the s_{ij} are the observed random variables. It is assumed that all the parameters are positive.

If stratification is disregarded then the usual estimate of $N_{..}$ is

$$(1)\quad \hat{N}_0 = \frac{n_{..}\, t_{..}}{s_{..}}$$

though for small $s_{..}$

$$(2)\quad \hat{N}_1 = \frac{(n_{..}+1)\,(t_{..}+1)}{s_{..}+1}$$

is preferable (see [4]).

The estimate proposed by Schaefer in the notation given here is

$$(3)\quad \hat{N}_2 = \sum_{i=1}^{r} \sum_{j=1}^{r} \left(\frac{n_{.j}\, t_{i.}\, s_{ij}}{s_{i.}\, s_{.j}} \right)$$

The estimate derived below is most simply expressed in vector-matrix notation as

(4) $\hat{N}_3 = \vec{n}\, S^{-1} \vec{t}$

where $\vec{n} = (n_{.1}, n_{.2} \cdots n_{.r})$

$\vec{t} = (t_{1.}, t_{2.} \cdots t_{r.})$

and S is the matrix (s_{ij})

An alternative form of this estimate is

(5) $\hat{N}_3 = \sum_{i=1}^{r} \sum_{j=1}^{r} t_{i.}\, n_{.j}\, s^{ji}$

s^{ji} being the inverse element of s_{ij} in S.

The primary property of these estimates, or more precisely sequences of estimators, that will be studied in this paper is consistency. The property of consistency of estimates based on samples from a finite population has been variously defined; consequently it is necessary to make clear the definition that will be followed in this paper.

Following one such usage an estimate $\hat{N}$ of $N_{..}$ would be called consistent if $\hat{N} = N_{..}$ whenever all $n_{.j} = N_{.j}$, i.e. whenever the sample, taken without replacement exhausts the population. This usage makes the definition particular not only to the finiteness of the sample but also to the method of sampling. Moreover it is certainly satisfied in this problem if, whenever $s_{ij} = t_{ij}$ for all i, j $\hat{N} = N_{..}$, and it is easy to construct numerous estimates that satisfy this condition and are otherwise meaningless. The definition might be made more restrictive by some monotonicity requirement on the distribution of N, but this is difficult to formulate and to use.

Alternatively it is possible to define consistency within an infinite sequence of populations $N_{..}^{(k)}$ with $N_{..}^{(k)}$ tending to infinity in some prescribed manner. This is the device used by David [5] and by Madow [6] in proving asymptotic properties of sampling without replacement from finite populations.

We wish to set up models that involve a minimum of assumptions; in [general] the assumptions will specify only the expectations of various rand[om] variables in the model. With this in mind it is simpler to define cons[ist]ency as follows (the term "estimate" will be used as a tautology for "se[]quence of estimators" and plim for limit in probability).

<u>Definition 1</u> An estimate $\hat{N}$ of $N_{..}$ is consistent if plim $s_{ij} = E(s_{ij})$ for all i, j implies plim $\hat{N} = N_{..}$ (indentically in the parameters involv[ed]).

Many of the assumptions will be stated in terms of conditional expe[cta]tions. In this connection $E(X|y)$ or $E(X|Y)$ will be used to denote the [con]ditional expectation of the random variable X given that the random vari[able] $Y = y$.

This usage of conditional expectations suggests two additional defi[ni]tions.

<u>Definition 2</u> An estimate $\hat{N}$ of N is said to be unconditionally c[onsis]tent with respect to <u>the random variable Y</u> $[u.c.(Y)]$ if

$$\text{plim } s_{ij} = E(s_{ij} | Y) \quad \text{for all } i, j \text{ implies}$$
$$\text{plim } \hat{N} = N \quad \text{for all values of } Y.$$

<u>Definition 3</u> An estimate $\hat{N}$ of $N_{..}$ is said to be conditionally co[nsist]ent with respect to y $[c.c.(y)]$ if

$$\text{plim } s_{ij} | y = E(s_{ij} | Y = y) \quad \text{for all } i, j \text{ implies}$$
$$\text{plim } \hat{N} = N \quad \text{for } Y = y.$$

It may be noted that if the sample size is increased to equal the wh[ole] population then $s_{ij} = E(s_{ij})$ and consistency as defined above would imp[ly] finite consistency referred to above. Also in general the conditions im[posed] upon the sequence of populations, considered for example by David and h[] are such as to insure the convergence in probability of the random vari[ables] s_{ij} to their expectations, so that consistency in the sense of Definitio[n]

We wish to set up models that involve a minimum of assumptions; in general the assumptions will specify only the expectations of various random variables in the model. With this in mind it is simpler to define consistency as follows (the term "estimate" will be used as a tautology for "sequence of estimators" and p lim for limit in probability).

<u>Definition 1</u> An estimate $\hat{N}$ of $N_{..}$ is consistent if $p\lim s_{ij} = E(s_{ij})$ for all i, j implies $p\lim \hat{N} = N_{..}$ (indentically in the parameters involved).

Many of the assumptions will be stated in terms of conditional expectations. In this connection $E(X|y)$ or $E(X|Y)$ will be used to denote the conditional expectation of the random variable X given that the random variable $Y = y$.

This usage of conditional expectations suggests two additional definitions.

<u>Definition 2</u> An estimate $\hat{N}$ of N is said to be unconditionally consistent with respect to <u>the random variable Y</u> [u.c.(Y)] if

$$p\lim s_{ij} = E(s_{ij} | Y) \quad \text{for all } i, j \text{ implies}$$
$$p\lim \hat{N} = N \quad \text{for all values of } Y.$$

<u>Definition 3</u> An estimate $\hat{N}$ of $N_{..}$ is said to be conditionally consistent with respect to y [c.c.(y)] if

$$p\lim s_{ij} | y = E(s_{ij} | Y = y) \quad \text{for all } i, j \text{ implies}$$
$$p\lim \hat{N} = N \quad \text{for } Y = y.$$

It may be noted that if the sample size is increased to equal the whole population then $s_{ij} = E(s_{ij})$ and consistency as defined above would imply the finite consistency referred to above. Also in general the conditions imposed upon the sequence of populations, considered for example by David and by Madow, are such as to insure the convergence in probability of the random variables s_{ij} to their expectations, so that consistency in the sense of Definition 1 would imply consistency within such a sequence.

3. Stratified Population Models It has been noted above that a basic assumption in most population estimation work is that of "randomness" of the sample, essentially that the properties of being sampled and being marked are independent. There is frequently no way to test the validity of this assumption and little reason to expect it to hold for a large heterogeneous population. While it may still not be possible to test the assumption, it may be more reasonable to assume some such "randomness" within small homogeneous substrata.

The minimum possible assumption appears to be

$$\text{I.} \qquad E(s_{ij} \mid n_{ij}, t_{ij}) = \frac{n_{ij} t_{ij}}{N_{ij}} \quad \text{for all } i, j.$$

This merely assumes that a random sample, on the average, is taken within the ij^{th} substratum. However a model constructed on assumption I appears to be inadequate to yield an estimate of $N_{..}$ for it involves $3r^2$ unknowns (n_{ij}, t_{ij}, N_{ij}'s) and there are only r^2 observable random variables (s_{ij}) plus 2r side conditions ($\sum_i n_{ij} = n_{.j}$, $\sum_j t_{ij} = t_{i.}$) to determine these. The information is inadequate, except in the trivial case $r=1$, so that it is necessary to make further assumptions to set up some structure relating the various substrata. In this respect it is sufficient that either

$$\text{II} \quad E(n_{ij}) = n_{.j} \left(\frac{N_{ij}}{N_{.j}}\right) \quad \text{for all } i, j \text{ with the distribution of } t_{ij} \text{ arbitrary}$$

or

$$\text{II}' \quad E(t_{ij}) = t_{i.} \frac{N_{ij}}{N_{i.}} \quad \text{for all } i, j \text{ with the distribution of } n_{ij} \text{ arbitrary.}$$

Assumption II states that, on the average, the various substrata are proportionately represented among the sample recovered while assumption II' states that the same property holds among the group tagged or marked.

It is seen that I and II together imply

$$\text{III} \quad E(s_{ij} \mid t_{ij}) = n_{.j} \frac{t_{ij}}{N_{.j}} \quad \text{for all } i, j.$$

Assumptions I and III also imply II but II and III do not imply I. For example consider II and III holding together with $E(s_{ij}|n_{ij}, t_{ij}) = \delta_{ij}n_{ij} + \varepsilon_{ij}n_{ij}^2$. Then it is trivial to determine $\delta_{ij}, \varepsilon_{ij}$ for each $E(n_{ij}^2)$ so that this assumption together with II and III are consistent. Consequently it follows that III alone does not imply I and II and is therefore a weaker assumption. However in an actual field situation it is likely that III will be satisfied only if I and II are.

If assumption II' is made rather than II then I and II' together imply

III': $E(s_{ij}|n_{ij}) = t_{i.} \frac{n_{ij}}{N_{i.}}$

Assumption III requires that, on the average, in the sample of size $n_{.j}$ from the population $N_{.j}$, the various tagged groups are proportionately represented. The dual assumption III' makes the same requirement but treats the tagged group as the sample and the subsequent recovery as the property of being marked. This appears to be a less reasonable practical assumption in that it requires predicting the future behavior of the animals marked. It might be of interest to note that whereas no effect due to tagging must usually be assumed (though seldom satisfied), I and II put no restriction on possible differential migration between tagged and untagged fish. Assumptions I and II', however, do require that the migration pattern into the different recovery strata be the same for tagged and untagged fish.

Starting with I and II the estimate $\hat{N}_j$ is easily derived, for summing

$E(s_{ij} \frac{N_{.j}}{n_{.j}}|t_{ij}) = t_{ij}$ over j yields

(7) $$E(\sum_{j=1}^{r} s_{ij} \frac{N_{.j}}{n_{.j}}) = t_{i.} \quad (i=1,2\ldots r)$$

(8) The set of equations $\sum_{j=1}^{r} s_{ij} \frac{\hat{N}_{.j}}{n_{.j}} = t_{i.}$ $(i=1,2,\ldots r)$ form a set of r equations in r unknowns, which has the solution

(9) $\left(\frac{\hat{N}}{n}\right) = S^{-1}\vec{t}$ provided $|S| \neq 0$

Here $\left(\frac{\hat{N}}{n}\right) = \left(\frac{\hat{N}_{.1}}{n_{.1}}, \frac{\hat{N}_{.2}}{n_{.2}}, \ldots \frac{\hat{N}_{.r}}{n_{.r}}\right)$

Summing the estimates of $\hat{N}_{.j}$ yields

(10) $\hat{N}_3 = \vec{n}' S^{-1}\vec{t}$ as an estimate of $N_{..}$.

If I and II' are made, the same procedure yields the equations

(11) $\sum_{i=1}^{r} s_{ij} \frac{\hat{N}_{i.}}{t_{i.}} = n_{.j}$ $(j=1,2,\ldots r)$ to estimate $N_{i.}$.

The estimate obtained in this way is also $\hat{N}_3$ since

$\vec{t}'(S')^{-1}\vec{n} = \vec{n}' S^{-1}\vec{t}$

Before studying the consistency of these estimates, one further estimation problem may be noted. In some situations, particularly migration studies, the N_{ij} will be of interest.

If assumptions I, II and II' are made then $N_{.j}$ and $N_{i.}$ are both estimable (estimates are given by (8) and (11)). Also the three assumptions imply

(12) $E(s_{ij}) = t_{i.}\, n_{.j} \frac{N_{ij}}{N_{i.} N_{.j}}$

so that an estimate of N_{ij} is

(13) $\hat{N}_{ij} = \frac{s_{ij}\, \hat{N}_{i.}\, \hat{N}_{.j}}{t_{i.}\, n_{.j}}$

4. <u>Consistency of the Estimates</u> In view of the continuity of the estimates $\hat{N}(s_{ij})$ for s_{ij} in the neighborhood of $E(s_{ij})$, it is necessary only to consider the conditions under which

(14) $\hat{N}\left[E(s_{ij})\right] = N$.

The basic assumptions are I and II (equivalent results are obtained if II is replaced by II') so that $E(s_{ij}|t_{ij}) = n_{.j} \frac{t_{ij}}{N_{.j}}$ with the distribution of t_{ij} arbitrary.

For $s_{ij} = E(s_{ij})$,

(15) $$\hat{N}_0 = \frac{n_{..}\,t_{..}}{\sum_j \frac{t_{.j}\,n_{.j}}{N_{.j}}}$$

Therefore $\hat{N}_0$ is u.c. (t_{ij}) if and only if $\frac{n_{.j}}{N_{.j}}$ = constant i.e. if the sampling is proportional to the population size at all stages.

Similarly, if I and II' are assumed, $\hat{N}_0$ is u.c. (n_{ij}) if and only if $\frac{t_{i.}}{N_{i.}}$ is constant, i.e., if the number tagged is proportional to the population size at all stages.

Now consider $\hat{N}_2$ with s_{ij} set equal to $E(s_{ij} \mid t_{ij})$. Then $\hat{N}$ is u.c. (t_{ij}) if

(16) $$\sum_i \sum_j \frac{t_{ij}\,t_{i.}\,n_{.j}}{t_{ij} \sum_\alpha t_{i\alpha}\left(\frac{n_{.\alpha}}{N_{.\alpha}}\right)} = N_{..}$$

which will be true provided $\frac{n_{.j}}{N_{.j}}$ = constant, i.e. proportional sampling takes place throughout the several stages.

Similarly, under assumptions I and II', $\hat{N}_2$ is consistent if and only if $\frac{t_{i.}}{N_{i.}}$ is constant. Therefore $\hat{N}_2$ has the same consistency properties as $\hat{N}_0$.

Now turning to $\hat{N}_3$, it is seen that substituting $E(s_{ij})$ for s_{ij} and $N_{.j}$ for $\hat{N}_{.j}$ equations (8) are satisfied. Hence if $|S| \neq 0$, the uniqueness of the solution of this set of linear equations assures that $\hat{N}_3$ is u.c. (t_{ij}), under I and II or u.c. (n_{ij}) under I and II'.

Consider the random matrix S with s_{ij} set equal to $E(s_{ij} \mid t_{ij})$--assuming I and II. Denote the determinant of this new matrix by Det $\left[E(S \mid t_{ij})\right]$. It is immediate that Det $\left[E(S \mid t_{ij})\right] \neq 0$ provided the matrix T: (t_{ij}) is nonsingular.

If in addition assumption II' is made then $E(T) = \left(t_{i.}\ \frac{N_{ij}}{N_{i.}}\right)$ which is nonsingular provided (N_{ij}) is nonsingular.

It might be possible to construct a test of the hypothesis: (N_{ij}) is singular. However the tediousness of such a test makes it hardly worthwhile. Perhaps of more meaning biologically is the hypothesis that there is random mixing i.e. that

$$(20) \quad N_{ij} \propto N_{i.} \, N_{.j}$$

(which would imply $\left|(N_{ij})\right| = 0$)

To construct such a test we make Assumption IV. The distribution of the n_{ij}, t_{ij} and the conditional distribution of the s_{ij} given n_{ij}, t_{ij} are multinomial with expectations given by the equations in I and II.

It is thus elementary to derive the variance of each s_{ij}, by working with conditional expectations.

$$(21) \quad \sigma^2_{s_{ij}} = \frac{t_{i.} n_{.j} N_{ij}}{N_{i.} N_{.j}} \left[1 + \frac{1}{N_{.j}} \left(\left[n_{.j} - 1 \right] \left[1 - \frac{N_{ij}}{N_{i.}} \right] t_{i.} \, \frac{N_{ij}}{N_{i.}} \right) \right]$$

Hence if N_{ij}, $N_{.j}$, $N_{i.}$ tend to infinity in such a way that $\frac{n_{.j}}{N_{.j}}$ $\frac{t_{i.}}{N_{i.}}$ tend to zero, then under the hypothesis the asymptotic variance of the s_{ij} is $t_{i.} \, n_{.j} \, N_{..}^{-1}$. Under the same restrictions the s_{ij} are asymptotically independent, while asymptotic normality is proven under weaker restrictions in the standard way. Thus under the restrictions that $n_{.j}$ and $t_{i.}$ be small relative to $N_{..}$ an approximate test of the hypothesis of complete mixing i.e. (20), is based on the statistic

$$(22) \quad \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{r} \frac{\left(s_{ij} - \frac{t_{i.} n_{.j}}{\hat{N}} \right)^2}{\frac{t_{i.} n_{.j}}{\hat{N}}}$$

where $\hat{N} = \frac{n_{..} t_{..}}{s_{..}}$

If the $n_{.j}$ are considerably larger than the $t_{i.}$ and are not small relative to $N_{..}$, this test should be used with caution, for the type 1 error may be much larger than the nominal significance level. This is partly due to the fact that

$\hat{N}_0$ is not exactly the modified minimum χ^2 estimate of $N_{..}$. The inflation of χ^2 in (22), caused by the underestimate of $\sigma^2 s_{ij}$, is more serious. The exact variance of the s_{ij} contains terms involving $N_{.j}$, $N_{i.}$, N_{ij} which cannot be estimated by the modified minimum χ^2 method. Hence no asymptotically efficient estimates of these parameters exist under the hypothesis. An approximate correction may be obtained by estimating the $N_{.j}$ from equations (8), and substituting these estimates in (21).

In many cases no test is necessary since the nature of the situation dictates that S is nonsingular. Thus if the stratification is with respect to time and the time periods are set up so that an animal marked in period i cannot be recovered in any period j, where $j<i$, then $s_{ij}=0$ for all $j<i$. Hence S is nonsingular provided all $s_{ii}\neq 0$; hence it certainly converges in probability to a nonsingular matrix.

5. Non-Existence of Unbiased Estimates. It is to be expected that if stronger assumptions are made, in particular assumptions about the distributions of the several random variables involved, estimates might be found with stronger properties. An assumption such as IV opens the possibility of obtaining maximum likelihood or minimum χ^2 estimates. However the modified minimum χ^2 estimates obtained by the use of Lagrange multipliers would require the solution of $r^2 + 2r$ linear equations. Even for r as small as 2 this is hardly feasible for general usage in the absence of special computing facilities. Whether those procedures could be simplified or other optimum estimates can be found remains an open question. However the following negative result may be of some interest.

Under assumption IV, no unbiased estimate of $N_{..}$ with finite variance exists.

Proof Let $M(s_{ij}\,;\, n_{.j}, \frac{t_{ij}}{N_{.j}})$ $\quad M(t_{ij};\, t_{i.}\, \frac{N_{ij}}{N_{i.}})$

denote the conditional distribution of s_{ij} given t_{ij} and of the distribution of t_{ij} respectively. Then the probability distribution of the s_{ij} denoted by $P(s_{ij}; t_{i.}, n_{.j}, N_{ij}, N_{i.}, N_{.j})$, or more briefly $P(s_{ij})$ is

$$(23) \quad P(s_{ij}) = \prod_{j=1}^{r} \sum_{t_{ab}} M(s_{ij}, n_{.j}, \frac{t_{ab}}{N_{.j}}) \prod_{a=1}^{r} M(t_{ab}, t_{i.}, \frac{N_{aj}}{N_{a.}})$$

where the summation with respect to t_{ab} is over all partitions of $t_{a.}$ for each a.

Let

$$(24) \quad \pi_k(s) = \frac{P(s_{ij}; t_{i.}, n_{.j}, N_{ij}^{(k)}, N_{i.}^{(k)}, N_{.j}^{(k)})}{P(s_{ij}; t_{i.}, n_{.j}, N_{ij}^{(0)}, N_{i.}^{(0)}, N_{.j}^{(0)})}$$

where $N_{ij}^{(k)}$. $N_{i.}^{(k)}$, $N_{.j}^{(k)}$ represent admissable values of the unknown parameters.

Then it follows from Barankins' theorem [7] that no unbiased estimate of $N_{..}$ with finite variance can exist, if there does not exist a finite constant C such that the inequality

$$(25) \quad \left[\sum_{k=1}^{m} a_k (N_{..}^{(k)} - N_{..}^{(0)})\right]^2 \leq C \prod_j \sum_{s_{ij}} (\sum_k a_k \pi_k(s))^2 \, P(s_{ij}; t_{i.}, n_{.j}, N_{ij}^{(0)}, N_{i.}^{(0)}, N_{.j}^{(0)})$$

holds for every set $\pi_1(s)$, $\pi_2(s), \ldots \pi_m(s)$, for all real numbers a_1 , a_2 , ... a_m and for every integer m.

Now consider a sequence of $N_{ij}^{(k)}$ tending to infinity. For m=1 the left hand side is unbounded but since P (0)$\neq$0, the right hand side is bounded. Hence no finite C exists satisfying (25) and hence no unbiased estimate of $N_{..}$ exists with finite variance.

The theorem also remains true if the distributions are multihypergeometric rather than multinomial (as in general in practice they will be). In this connection it should be pointed out that formula (21) will then also involve finite sampling correction terms.

Furthermore it is true even if the parameter space is bounded. The same tool, Barankins' theorem i.e condition (25), is applicable but a somewhat more tedious argument is involved.

6. Asymptotic Variance of $\hat{N}_3$. The term asymptotic is used here in the same sense as in the definition of consistency in the previous section, i.e. as all s_{ij} converge to $E(s_{ij})$. We also make assumptions I, II and II'.

From theorems on matrix differentiation, [8],

$$\text{(26)} \quad \frac{\partial \hat{N}_3}{\partial s_{ab}} = \vec{n}' S^{-1} I_{ab} S^{-1} \vec{t}$$

where I_{ab} is the matrix with 1 in the abth place and 0 everywhere else. This reduces to

$$\text{(27)} \quad \frac{\partial \hat{N}_3}{\partial s_{ab}} = \left(\sum_j n_{.j} \frac{S_{aj}}{|S|}\right)\left(\sum_i \frac{S_{ib}}{|S|} t_{i.}\right)$$

$|S|$ being the determinant of S, S_{ij} the signed cofactor of s_{ij}.

Under the assumptions $E(s_{ij}) = \dfrac{t_{i.} n_{.j} N_{ij}}{N_{i.} N_{.j}}$

Then substituting $E(s_{ab})$ for s_{ab} it is readily seen that

$$\text{(28)} \quad \left(\frac{\partial \hat{N}_3}{\partial s_{ab}}\right)_{s_{ab} = E(s_{ab})} = \left(\frac{N_{a.}}{t_{a.}}\right)\left(\frac{N_{.b}}{n_{.b}}\right)$$

Finally then, for the several parameters N, n, t tending to infinity in such a way that $\dfrac{n_{.j}}{N_{.j}} \to 0$, $\dfrac{t_{i.}}{N_{i.}} \to 0$ (which imply the asymptotic independence of the (s_{ij}) the asymptotic variance of $\hat{N}_3$ is

$$\text{(29)} \quad \sum_i \sum_j \frac{N_{ij} N_{i.} N_{.j}}{t_{i.} n_{.j}}$$

If $r=1$ this reduces to $\frac{N_{..}^3}{nt}$, the asymptotic variance for sampling a homogeneous or non-stratified population.

Applying the Schwartz inequality

$$(30) \quad \frac{N_{..}^2}{t_{..}} \leqq \sum_{i=1}^{r} \frac{N_{i.}^2}{t_{i.}}$$

and

$$(31) \quad \frac{N_{..}^2}{n_{..}} \leqq \sum_{j=1}^{r} \frac{N_{.j}^2}{n_{.j}}$$

From this it follows that the asymptotic variance of $\hat{N}_0$ is not greater than that of $\hat{N}_3$, if either tagging or sampling is proportionate i.e. if

$t_i \propto N_{i.}$ or $n_{.j} \propto N_{.j}$. Hence, if valid, $\hat{N}_0$ is a better estimate than $\hat{N}_3$.

Since, within this model simple estimates of the $N_{i.}$, $N_{.j}$ and N_{ij} are obtainable it is possible to use formula (29) to determine approximate confidence intervals or tests for $N_{..}$.

It is also possible to determine the asymptotic variance of $\hat{N}_{.j}$ by the same method. Under the same restrictions noted earlier in this section, this asymptotic variance is

$$(32) \quad A.V(\hat{N}_{.j}) = \frac{N_{.j}^2}{|\mathcal{N}|^2} \sum_a \sum_b \frac{\mathcal{N}_{aj}^2 N_{ab} N_{a.} N_{.b}}{t_{a.} n_{.b}}$$

where $\mathcal{N} = (N_{ij})$

and $\mathcal{N}_{aj}$ is the signed cofactor of N_{aj} in $\mathcal{N}$. This will be unpleasantly tedious to calculate even for r as small as 4.

7. Variable Number of Strata. In some situations the number of strata will change between the times of tagging and sampling. This may occur either where the distribution is by area or by time.

Suppose there are m strata at the time of tagging or marking, r strata at the time of sampling or recovery.

Consider $m > r$, with assumption III holding. The equations (8) yield m equations in r unknowns. The simplest device is the combination of some of the tagging periods or areas to form a system that has a unique solution. Of course, using assumption IV an optimum asymptotic solution could be found by determining the modified minimum χ^2 estimates. If assumption III' holds there are r equations in m unknowns, from (11), and no solution is possible.

In case $m < r$, these conclusions are reversed i.e. no estimation possible under Model III, estimation possible with assumption III'. In general it is reasonable to expect that either assumptions III or III' can be made so that estimation of $N_{..}$ is possible.

Another variation that may arise is that $t_{i.} = 0$ or $n_{.j} = 0$ for some i or j. It is easy to construct numerical examples to show that the conditional distribution of the s_{ij} given t_{ij} may be the same with different sets of the parameters N_{ij} and in particular with different $N_{..}$. Thus assumptions I, II and II' yield no information as to the estimation of $N_{..}$, though it is possible to estimate the total population of the strata where tagging or sampling takes place. It is conjectured that even if further assumptions are made such as IV, the amount of information as to $N_{..}$, is almost negligible.

References

[1] M. B. Schaefer, "Estimation of the size of animal populations by marking experiments." U. S. Fish and Wildlife Service Fisheries Bulletin 69 (1951) pp. 191-203

[2] Douglas G. Chapman, "The estimation of biological populations", Ann. Math. Stat. 25 (1954) pp. 1-15

[3] Karl W. Kenyon, Victor B. Scheffer and Douglas G. Chapman, "A population study of the Alaska fur-seal herd" U. S. Fish and Wildlife Service Special Scientific Report-Wildlife No 12 (1954) pp. 1-77

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

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

[6] William G. Madow, "On the limiting distributions of estimates based on samples from finite universes." Ann. Math. Stat. 19 (1948) pp. 535-545

[7] E. W. Barankin, "Locally best unbiased estimates," Ann. Math. Stat. 20 (1949) pp. 477-501

[8] Paul S. Dwyer and M. S. MacPhail, "Symbolic matrix derivatives", Ann. Math. Stat. 19 (1948) pp. 517-534

Chief of Naval Research
Office of Naval Research
Washington 25, D. C.
Attn: Code 432
(Mathematics Branch) 1

Dr. E. Lukacs
Head, Statistics Branch
Office of Naval Research
Department of the Navy
Washington 25, D. C. 2

Scientific Section
Office of Naval Research
Department of the Navy
1000 Geary Street
San Francisco 9, Calif
Attn: Dr. J. Wilkes 3

Director, Naval Research
Laboratory
Washington 25, D. C.
Attn: Technical Information
Officer 6

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

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, New York 1

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

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

Office of Naval Research
Logistics Branch - Code 436
T-3 Building
Washington 25, D. C. 1

ASTIA Documents Service Center
Knott Building
Dayton 2, Ohio 5

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

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

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

Office of Technical Services
Department of Commerce
Washington 25, D. C. 1

Commander, Air Force Personnel
and Training Research Center
Attn: Director, Personnel Research Laboratory
Lackland Air Force Base, Texas 1

Institute for Numerical Analysis
405 Hilgard Avenue
Los Angeles 24, California 2

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

RAND Corporation
1500 Fourth Street
Santa Monica, California 1

Applied Mathematics and
Statistics Laboratory
Stanford University
Stanford, California 3

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

Professor W. G. Cochran
Department of Biostatistics
The Johns Hopkins University
Baltimore S. , Maryland 1

Professor Benjamin Epstein
Department of Mathematics
Wayne University
Detroit 1, Michigan 1

Professor Herbert Solomon
Teachers College
Columbia University
New York, New York 1

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

Professor J. Wolfowitz
Department of Mathematics
Cornell University
Ithaca, New York 1

Department of Mathematical Statistics
University of North Carolina
Chapel Hill, North Carolina 2

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

Professor S. S. Wilks
Department of Mathematics
Princeton, New Jersey 1

OFFICE OF NAVAL RESEARCH
PHYSICS BRANCH
TECHNICAL REPORTS DISTRIBUTION LIST
for
NR ~~010-604~~ 019-614

A. Government Distribution

1. Department of Defense

Asst. Sec. of Defense for Research and Development
Information Office Library Branch
Pentagon Building
Washington 25, D.C. (2 copies)

Armed Services Technical Information Agency
Documents Service Center
Knott Building
Dayton 2, Ohio (5 copies)

2. Department of the Navy

Chief, Bureau of Aeronautics
TD-4
Navy Department
Washington 25, D.C.

Chief, Bureau of Ordnance
Rea
Navy Department
Washington 25, D.C.

Chief, Bureau of Ships
Code 300
Navy Department
Washington 25, D.C.

Chief of Naval Research
Office of Naval Research
Washington 25, D.C.
Attn: Physics Branch (Code 421) (3 copies)

Director,
Naval Research Laboratory
Technical Information Officer (Code 2000) (6 copies)
Code 2021 (2 copies)
Washington 25, D.C.

Commanding Officer
Office of Naval Research Branch Office
150 Causeway Street
Boston 14, Massachusetts

Commanding Officer
Office of Naval Research Branch Office
346 Broadway
New York 13, New York

Office of Naval Research Branch Office
Commanding Officer
100 Geary Street
San Francisco 9, California

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

Commanding Officer
Office of Naval Research London Branch Office
Navy #100
Fleet Post Office
New York, New York (2 copies)

Librarian
U.S. Naval Post Graduate School
Monterey, California

Physics Department
U.S. Naval Post Graduate School
Monterey, California

U.S. Naval Electronics Laboratory Library
San Diego 52, California

Commanding Officer
U.S. Naval War College
Newport, Rhode Island

Director
Physics Research
Naval Ordnance Laboratory
White Oaks, Maryland

Commanding Officer
Physics Division
Naval Ordnance Test Station
Inyokern
China Lake, California

3. Department of the Air Force

Commanding General
Air Research and Development Command
Attn: Office of Scientific Research
Post Office Box 1395
Baltimore 3, Maryland

Commander
Air Force Cambridge Research Center
ATTN: CRHLP-1
230 Albany Street
Cambridge, Mass.

4. Department of the Army

Commanding Officer
Engineering Research & Development Laboratories
Fort Belvoir, Virginia
Attn: Technical Intelligence Branch

Office of Ordnance Research
Box CM, Duke Station
Durham, North Carolina (2 copies)

Signal Corps Engineering Laboratory
Fort Monmouth, New Jersey
Attn: Technical Information Officer

5. Department of Commerce

Office of Technical Services
Technical Reports Section
Department of Commerce
Washington 25, D.C. (2 copies)

Director
National Bureau of Standards
Washington 25, D.C.

Director
National Bureau of Standards
Corona, California

6. Other Agencies

National Research Council
Division of Physical Sciences
National Academy of Sciences
Washington, D.C.

Director
National Science Foundation
Washington 25, D.C.

U.S. Atomic Energy Commission
Technical Information Service
P.O. Box 62
Oak Ridge, Tennessee

Atomic Energy Commission
Washington 25, D.C.

Dr. R. C. Hermann
8621 Georgia Avenue
Applied Physics Laboratory
The Johns Hopkins University
Silver Spring, Maryland