

Armed Services Technical Information Agency

AD

19479

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

ATI No. 19479
ASTIA FILE COPY

THE ESTIMATION OF BIOLOGICAL POPULATIONS

By

Douglas G. Chapman

University of Washington

Technical Report No. 12

August 15, 1953

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

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

The Estimation of Biological Populations^{1,2}

By Douglas G. Chapman

University of Washington

Summary. A number of statistical models, underlying the methods used in the estimation of the sizes and other parameters of animal populations, are set up. The relevant estimation equations are given, with their variances and covariances. For the most part the theory is designed for large populations. In setting up the models, consideration has been given to the desideratum of having them conform as closely as possible to the actual practices of animal sampling. To what extent the models do agree with reality is one of the many open questions which are noted in this paper.

1. Introduction. The use of sampling methods in the enumeration of populations has become widely known and widely accepted only within the past generation. Yet it is easily perceived that total enumeration methods fail for all but the simplest of populations. Particularly is this true of biological populations which may be mobile in space, transient in time and difficult of access. The changes in space (immigration and emigration) and in time (recruitment and mortality) must often be evaluated to determine the total population size and in any case these changes are usually of interest in their own right.

¹ Presented as a special invited address at the joint meeting of the Institute of Mathematical Statistics and the Biometric Society (WNAF) at Stanford, California, June 19, 1953.

² Work done under the sponsorship of the Office of Naval Research.

In this survey, only those methods are considered for which it is possible to set up a reasonable statistical model and for which it is possible to assess the sampling errors. Attention is limited to methods that lead to absolute rather than relative estimates. Little work has been done to set up statistical models, as a basis of relative estimates, though for an important exception, attention is called to a paper of Neyman [22]. To give unity to this survey, only those methods that have been used in the study of macroscopic mobile populations are discussed.

Fixed sample methods have been used for the most part in the enumeration of other biological populations. However, even the enumeration of sessile populations can give rise to new statistical problems; many of these are noted in an important study of statistical problems in ecology, that recently has been initiated by Skellam [28]. A further reference in this field is to a paper by Hoel [13].

2. Tag-and-sample estimates: direct random sampling. When the population structure is undefined and unknown, it is not possible to select a fixed sample, as is the case say in ecology or in sampling human populations. The idea of using an associated variable of known distribution to build up a sample count into a total population was first proposed by Laplace [17]. He suggested determining the population of France from the known number of births in all parishes and from the fact that the ratio of births to total population could be determined for some parishes.

Petersen [23], a Danish biologist, first developed the procedure of marking fish to assist in studying their movements, migration, etc.

He later came to realize that the marked fish could play the same role for his populations as the births did for Laplace—though evidently he was unaware of Laplace's work.

When a mathematical model is set up to formalize this intuitive approach, it is usual to assume random sampling [i.e., sampling such that the properties "being tagged" and "being sampled" are independent]. It is much easier to make this assumption than to verify it. It is also standard to assume that the numbers tagged and the numbers sampled are parameters at the disposal of the experimenter.

A completely adequate model must take into account the birth rate with possible lag effects, a changing death rate, as well as emigration and immigration over the period, during which repeated tagging and sampling take place. It is apparent that the number of unknown parameters is large and that such a model must be indeed complex. Some simplifying assumptions are desirable.

The following model is not the most general possible; it does, however, cover many of the situations that have been studied and it leads to simple estimation procedures. It applies specifically to large populations and it is further assumed that either there are no recruits to the population (through birth or immigration) or that new recruits are distinguishable and may be eliminated from the samples.

Model I

$$\text{Unknown parameters} \quad \left\{ \begin{array}{l} N_0 = \text{total population size at time zero} \\ P = \text{probability that an animal alive at time } t \\ \text{survives and remains in the population at} \\ \text{time } t+1 \end{array} \right.$$

$$\begin{array}{l}
\text{Known parameters} \quad \left\{ \begin{array}{l} t_i = \text{number of animals tagged at the } i^{\text{th}} \text{ tagging} \\ \text{taking place at time } a_i \text{ (} i=1,2,3\dots m \text{)} \\ n_j = \text{number of animals sampled in the } j^{\text{th}} \text{ sample} \\ \text{taken at time } b_j \text{ (} j=1,2,3\dots r \text{)} \end{array} \right. \\
\\
\text{Random variables} \quad \left\{ \begin{array}{l} x_{ij} = \text{number of animals originally tagged at the} \\ i^{\text{th}} \text{ tagging and recovered in the } j^{\text{th}} \text{ sample} \\ \tau_{ij} = \text{number of animals originally tagged at the} \\ i^{\text{th}} \text{ tagging available for recovery at the} \\ \text{time of the } j^{\text{th}} \text{ sample} \\ N_j = \text{population size at the time of the } j^{\text{th}} \text{ sample} \end{array} \right. \\
\\
f(i) = \text{the smallest value of } j \text{ such that animals} \\
\text{tagged at the } i^{\text{th}} \text{ tagging have a positive} \\
\text{probability of being recovered in the } j^{\text{th}} \\
\text{sample}
\end{array}$$

The event of survival is assumed to be independent from animal to animal. For large N_0 it may be assumed that given τ_{ij} , N_j , (which are not observable r.v.) x_{ij} has a conditional Poisson distribution with expectation $(n_j \tau_{ij}) N_j^{-1}$. It then follows that, for large N_0 , x_{ij} has approximately a Poisson distribution with

$$(1) \quad E(x_{ij}) = \frac{\tau_{ij} n_j P^{-a_i}}{N_0}$$

More precisely this holds as a limiting result as $N_0 \rightarrow \infty$ in such a way that $\frac{nt}{N_0}$ is finite (> 0) but $t \left(\frac{n}{N_0} \right)^k \rightarrow 0$ for $k > 1$. The result is obtained by working with the conditional m.g.f. in a standard manner.

With this approximation it is straightforward to set up the maximum likelihood equations for N_0 and P , viz.

$$(2) \quad \hat{N}_o = \frac{\sum_{i=1}^m \sum_{j=f(i)}^r t_{ij} \hat{P}^{-a_i}}{x_{..}}$$

$$(3) \quad x_{..} \left(\sum_{i=1}^m \sum_{j=f(i)}^r a_i t_{ij} \hat{P}^{-a_i} \right) = \left(\sum_{i=1}^m \sum_{j=f(i)}^r a_i x_{ij} \right) \left(\sum_{i=1}^m \sum_{j=f(i)}^r t_{ij} P^{-a_i} \right)$$

where the dot subscripts denote the conventional summation notation.

Equation (3) is a polynomial in P that can be solved by the usual methods.

The inverse of the asymptotic variance-covariance matrix of $\hat{N}$ and $\hat{P}$ is

$$(4) \quad \begin{pmatrix} \frac{\sum_{i=1}^m \sum_{j=f(i)}^r t_{ij} P^{-a_i}}{N_o^3} & \frac{\sum_{i=1}^m \sum_{j=f(i)}^r a_i t_{ij} P^{-(a_i+1)}}{N_o^2} \\ \frac{\sum_{i=1}^m \sum_{j=f(i)}^r a_i t_{ij} P^{-(a_i+1)}}{N_o^2} & \frac{\sum_{i=1}^m \sum_{j=f(i)}^r a_i^2 t_{ij} P^{-(a_i+2)}}{N_o} \end{pmatrix}$$

It is convenient to display the parameters (t_i, n_j) and the observations (x_{ij}) of such a census in a triangular array—the so-called "trellis diagram" used by Dowdeswell, Fisher and Ford [10] but much more thoroughly studied by Leslie and Chitty [19] and by Leslie [18]. Model I departs primarily from that proposed by Leslie and Chitty in ignoring multiple recaptures. Leslie and Chitty show this represents a loss of information; for large N_o , however, the expected number of multiple recaptures is very small. In fact if this is not so, it

suggests that the stochastic variation of τ_{ij} and N_j may no longer be negligible. Moreover the multiple recaptures are often those most suspect from the point of view of randomness of the sample.

Leslie and Chitty, in common with other investigators, assumed that mortality and emigration are strictly deterministic. Thus they are able to write down the expected values of the various classes of tag recoveries as polynomials in P , and to assume a multinomial distribution for these tag recoveries. The maximum likelihood equations can then be formulated, though the solution of the equations can, in general, be accomplished only by iterative methods. They have studied a large number of problems in this manner and reference should be made to their papers for models appropriate to situations not considered here. A model based on the Poisson distribution can also be set up for most of these situations, which will be valid for large N_0 , even though space and time variations are stochastic variables, and which will often lead to simpler estimation equations. A complete treatment, considering this stochastic variation, has not been given for the case of small or moderate sized populations.

The formulae given above easily specialize to Jackson's "negative" census [14], (one in which several taggings are followed by a single sample, at which time only, are tag recoveries noted). Bailey [1] has given the maximum likelihood estimates and their asymptotic variance-covariance matrices for Jackson's various census schemes assuming deterministic birth and death rates. Jackson also set up a "positive" census scheme, which he used to estimate the rate of recruitment.

By defining a parameter B , as the probability that an individual alive at time t adds a new individual to the population by time $t+1$, and assuming that this event is independent of the event of survival, the model outlined above may be extended and the restrictions of no recruitment may be removed. The x_{1j} still have a Poisson distribution, to the same approximation as before and the maximum likelihood equations for N_0 , P and B are easily written down. The two equations involving $\hat{P}$ and $\hat{B}$ are polynomials jointly in $\hat{P}$ and $\hat{B}$. However, it seems hardly realistic to assume that the recruitment rate is proportional to the population size or that it is independent of survival. Another approach is noted later.

Another specialization of formulae (1) to (4) is to put $P=1$, i.e. assume mortality can be neglected. This situation is familiar to fishery biologists as a Schnabel type census named for the person who published a mathematical theory of estimates based on such a multiple census [27]. More precisely, as noted by the author in [6], for large N_0 ,

$$(6) \quad N_0 = \frac{\sum_{i=1}^m \sum_{j=f(i)}^r t_i n_j}{x_{..} + 1}$$

is approximately unbiased with standard deviation given by

$$(7) \quad \frac{\sigma_{\hat{N}_0}}{N_0} = \sqrt{\frac{N_0}{\sum_{i=1}^m \sum_{j=f(i)}^r t_i n_j}} \left(1 + 0 \left[\frac{\sum_{i=1}^m \sum_{j=f(i)}^r t_i n_j}{N} \right]^{-1} \right)$$

Also confidence limits for the Poisson parameter will yield confidence limits for N_0 in this case--see e.g. [4].

In the usual Schnabel census tagging is carried on simultaneously with the sampling process. More precisely after each sample is examined, the untagged individuals are tagged and then all are returned to the population. If this is strictly followed

$$t_1 = n_1 - \sum_{j=1}^{i-1} s_{1j}$$

and hence the t_1 are random variables. For large N_0 , the random variation of the t_1 may be neglected. In fact it has usually been disregarded in any case.

It is apparent that there may have to be some restriction on m and r to make the results given above meaningful. In particular if $m = r = 1$, no estimation of the parameters N_0 and P is possible—but estimation of N_0 is possible if $a_1 = 0$. This is the simple Petersen situation,—a single tagging followed by a single sample. The formulae in this case are seen not to depend, for large N_0 , on mortality assumptions. While the variance of $\hat{N}_0 = \frac{(n+1)(t+1)}{x+1}$, the almost unbiased estimate of N_0 , which is given by

$$(8) \quad \sigma_{\hat{N}_0}^2 = N_0^2 \left[\frac{N_0}{nt} + O\left(\frac{N_0}{nt}\right) \right] + \frac{N_0}{P}$$

is a function of P , the survival factor, for most practical purposes this may be disregarded.

3. Tag-and-sample methods: inverse sampling. A modification of the sampling procedure outlined above has been developed by Bailey [1], Goodman [12] and the author [6]. If the number of tags to be recovered, rather than the sample size, is predetermined, estimates are obtained

which are somewhat simpler and slightly more efficient. The most interesting of these results is that due to Goodman, who considered a multiple sample type of census for a situation where there is no recruitment and $P=1$ (such a population will hereafter be referred to as closed). His procedure is sequential in that the decision to stop sampling is a consequence of the observations.

Model II

Unknown parameter	$\left\{ \begin{array}{l} N_0 = \text{population size} \end{array} \right.$
Known parameters	$\left\{ \begin{array}{l} n_i = \text{the predetermined sequence of samples} \\ T_i = \text{the number of tagged individuals in the population at the time the } i^{\text{th}} \text{ sample is taken} \\ \quad (T_i \text{ cumulative total is to be distinguished from } t_i, \text{ the number of tags put out in the } i^{\text{th}} \text{ tagging. } (i=1, 2, 3, \dots)) \end{array} \right.$
	$\left\{ \begin{array}{l} x = \text{the predetermined number of tagged members to be recovered before the sampling experiment stops} \end{array} \right.$
	$\left\{ \begin{array}{l} r = \text{the number of samples taken before the } x \text{ tagged individuals are recovered} \end{array} \right.$
Random variables	$\left\{ \begin{array}{l} n = \sum_{i=1}^r n_i \end{array} \right.$

Sampling is assumed to be random with respect to tagged and untagged individuals. Then for large N_0

$$\begin{aligned}
 & \Pr (r \text{ samples are required to obtain } x \text{ tagged members}) \\
 &= \sum_{j=1}^x \Pr [x-j \text{ tags are recovered in first } r-1 \text{ samples}] \\
 & \quad \cdot \Pr [j \text{ tags are recovered in the } r^{\text{th}} \text{ sample}]
 \end{aligned}$$

$$= \sum_{j=1}^x e^{-\sum_{i=1}^{r-1} \lambda_i} \frac{\left(\sum_{i=1}^{r-1} \lambda_i\right)^{x-j}}{(x-j)!} \cdot e^{-\lambda_r} \frac{\lambda_r^j}{j!}$$

$$= e^{-\sum_{i=1}^r \lambda_i} \left[\frac{\left(\sum_{i=1}^r \lambda_i\right)^x}{x!} - \frac{\left(\sum_{i=1}^{r-1} \lambda_i\right)^x}{x!} \right]$$

where we have written λ_i for $\frac{n_i T_i}{N_0}$.

Making the change of variable

$$u = 2 \sum_{i=1}^r \lambda_i \quad \Delta u = 2 \lambda_r$$

$$(10) \quad \Pr(a < u < b) = \sum_{r=r_1}^{r_b} e^{-\frac{u}{2}} \left[\left(\frac{u}{2}\right)^{x-1} \cdot \frac{1}{(x-1)!} \cdot \frac{\Delta u}{2} + o(\Delta u) \right]$$

Let $N_0 \rightarrow \infty$ in such a way that $\lambda_i \rightarrow 0$ while $\sum_{i=1}^{\infty} \lambda_i > 0$.

Using Dunham's lemma it can be shown that

$$(11) \quad \lim_{N_0 \rightarrow \infty} \Pr(a < u < b) = \frac{1}{2^x \Gamma(x)} \int_a^b e^{-\frac{u}{2}} u^{x-1} du$$

i.e. $u = 2 \sum_{i=1}^r \frac{n_i T_i}{N}$ has a limiting χ^2 distribution with $2x$ degrees of freedom.

It follows that the (asymptotic) minimum variance unbiased estimate of N_0 is

$$(12) \quad \hat{N}_0 = \frac{\sum_{i=1}^r n_i T_i}{x} \quad \text{and} \quad \sigma_{\hat{N}_0}^2 = \frac{N^2}{x}$$

The proof given above differs from that of Goodman: he considered the Schnabel type of census where tagging and sampling are performed in the same operation, i.e. all untagged individuals are tagged before the sample is returned to the population. What he showed, namely that $\frac{n^2}{N}$ has, asymptotically, a χ^2 distribution with $2x$ d.f., is equivalent to the above result. In this case it is simple to find the average sample size, i.e. $E(n)$, (for large N_0). For these results and other exact sample results reference is made to Goodman's paper cited, [12].

The simplicity of σ_{N_0} may make it particularly useful in designing the sample census. Up to the moment, however, the several inverse sampling schemes proposed, have not been tried out. How to choose the sequence $\{n_i\}$ in an optimum manner remains an open question. Nor has any attempt been made to set up a theory of inverse sampling for other than closed populations.

4. Tag-and-sample estimates: regression approach. The assumptions underlying Model I may fail for a variety of reasons—imperfect sampling, clustering of the populations, variation over the populations and over time, of the mortality (or emigration) rate, etc. In view of the considerable superimposed variability that may thus exist, in addition to strictly multinomial (or Poisson) variation, it is pertinent to ask whether a linear regression model might not be more appropriate.

Model III. The same notation as Model I is needed. However, the restriction that there be no recruitment may be removed. Hence it is more reasonable to regard $N_1, N_2, \dots, N_r$ as unknown parameters to be estimated. Furthermore the definition of P can be extended as follows: P = the average probability that an individual alive at time t survives and remains within the population to time $t+1$.

If the sampling is such that

$$(13) \quad E(x_{1j} | \tau_{1j}) = \frac{\tau_{1j} n_j}{N_j}$$

it follows that

$$(14) \quad E(x_{1j}) = \frac{t_{1j} n_j P^{b_j - a_1}}{N_j}.$$

The regression approach might be based upon the assumption that

$$(15) \quad E \left[\ln \frac{t_{1j} n_j}{x_{1j} + 1} \right] = (a_1 - b_j) \ln P + \ln N_j$$

and that $\ln(x_{1j} + 1)$ has a constant variance (approximately).

The factor $(x_{1j} + 1)$ is suggested by the fact that the reciprocal of a binomial or Poisson r.v. plus one is an (almost) unbiased estimate of the reciprocal of the parameter. Moreover such a device avoids the difficulties of occasional zeros—care must be exercised if the zeros are numerous or in sequence, for the assumption above may then be clearly invalid. The logarithmic transformation is suggested by the product nature of equation (14). However, it is also true that the variance of the logarithm of a variable that is distributed according to the Poisson law is constant up to terms of order λ^{-1} . Furthermore,

the logarithmic transformation has been extensively used in analysing data obtained from pelagic hauls or catches (cf. e.g. Windsor and Clark [30]).

Best linear unbiased estimates of $\ln P$ and $\ln N_j$ are found by the least squares method (under these assumptions). From these, estimates of P and N_j are obtainable which have optimum asymptotic properties though not necessarily optimum small sample properties. Interval estimates may also be obtained by postulating approximate normality of the $\ln(x_{1j}+1)$. Such interval estimates may be much more realistic than those based on Model I, if there is in fact superimposed variability due to the causes indicated or to other causes.

Model III represents in a sense an omnibus model. It has the advantage that an estimate of the extraneous variation can be made from the observations. On the other hand, it is imprecise and heuristic rather than rigorous. If the heterogeneities noted can be carefully assayed, if not controlled, it may be possible to set up a model which has this advantage and is at the same time more exact.

This type of approach would give some flexibility to the assumptions underlying Jackson's positive census (where a single tagging is followed by a sequence of samples) or more generally to the "trellis diagram" census scheme where recruitment is to be taken into account by a single parameter. Redefining B as the average probability that an individual within the population at time t adds a new recruit to the population at time $t+1$, similar assumptions as those above lead to

$$(16) \quad E(x_{1j}) = \frac{t_1 n_j P^{b_j - a_1}}{N_0 (P+B)^{b_j}}$$

Hence estimates of P , B and N_0 could be derived from the least squares estimates of $\ln P$, $\ln (P+B)$ and $\ln N_0$ from the equation:

$$(17) \quad E \left(\ln \frac{t_{ij} n_j}{x_{ij} + 1} \right) = (a_i - b_j) \ln P + \ln N_0 + b_j \ln (P+B).$$

5. Dichotomy methods. A method of estimating population size, that has been used in wildlife research, and which may be useful in other fields is based on the change of sex ratio caused by a selective kill. The sex ratio is determined before and after the kill by sampling methods. Several references to field applications of the method are listed by Scattergood [25] in a general survey of methods of population estimation.

The estimation procedure may be based upon any dichotomy within the population, or even on external factors: all that is required is a sampling process followed by a selective removal of individuals from the population, and subsequently a further sampling process. Closed populations only will be considered.

Model IV

$$\begin{array}{ll} \text{Unknown parameters} & \left\{ \begin{array}{l} N_i = \text{population at time } t_i \quad (i=1, 2) \\ \text{made up of two classes X and Y} \\ X_i, Y_i = \text{size of classes X, Y at times } t_i \end{array} \right. \\ \\ \text{Known parameters} & \left\{ \begin{array}{l} n_i = \text{size of random samples taken at time } t_i \\ r_x = X_1 - X_2; r_y = Y_1 - Y_2; r = r_x + r_y \end{array} \right. \\ \\ \text{Random variables} & \left\{ \begin{array}{l} x_i = \text{number of elements of class X in sample } n_i \\ y_i = \text{number of elements of class Y in sample } n_i \end{array} \right. \quad (i=1, 2) \end{array}$$

Assuming sampling with replacement

$$(18) \quad P(x_1, x_2) = \binom{n_1}{x_1} \left(\frac{x_1}{N_1} \right)^{x_1} \left(1 - \frac{x_1}{N_1} \right)^{n_1 - x_1} \binom{n_2}{x_2} \left(\frac{x_2}{N_2} \right)^{x_2} \left(1 - \frac{x_2}{N_2} \right)^{n_2 - x_2}$$

Since it is assumed that X_2 , N_2 are expressible in terms of X_1 and N_1 and known parameters, estimates of X_1 and N_1 are easily found. The moment estimate

$$(19) \quad \hat{X}_1 = \frac{x_1(n_2 r - x_2 r)}{n_2 x_1 - n_1 x_2}$$

$$(20) \quad \hat{N}_1 = \frac{n_1(n_2 r - x_2 r)}{n_2 x_1 - n_1 x_2}$$

are also maximum likelihood estimates.

Formulae for the ^{inverse of the} asymptotic variance-covariance matrix are as follows:

$$(21) \quad \begin{pmatrix} \frac{n_1}{x_1 y_1} + \frac{n_2}{x_2 y_2} & -\frac{n_1}{N_1 y_1} - \frac{n_2}{N_2 y_2} \\ -\frac{n_1}{N_1 y_1} - \frac{n_2}{N_2 y_2} & \frac{n_1}{N_1^2} \frac{x_1}{y_1} + \frac{n_2}{N_2^2} \frac{x_2}{y_2} \end{pmatrix}$$

so that

$$(22) \quad \sigma_{N_1}^2 \text{ (asymptotic)} = \frac{\frac{P_1^2 x_2 y_2}{n_2} + \frac{P_2^2 x_1 y_1}{n_1}}{(P_1 - P_2)^2}$$

where $P_i = \frac{x_i}{N_i}$ ($i = 1, 2$).

These formulae may be used to determine the optimum theoretical allocation of sampling between the before and after samples. It is

also interesting to use them to compare the effort required for this type of census with that required for tag sample methods. A numerical study shows that the tag sample method has the advantage--assuming that the tags are sampled by the removal process. However, the evaluation is incomplete without some means of determining the relative costs of sampling and tagging. Moreover it is reasonable to suppose that the assumptions underlying the dichotomy method are more likely to be fulfilled than in the tag-sample method--questions of tag mortality and differential recapture rates do not arise.

In some situations it may be possible to sample two populations, e.g. a sport fish and a scrap fish. The sports fishery then serves as the selective removal factor in a very favorable situation since r_y will be zero. In this case X is the parameter that it is of interest to estimate.

The method may also be applied where the removal is done by the sampler. In this case it is more realistic to assume a succession of samples are taken. Again it is straightforward to set up the model for this situation and to derive the maximum likelihood equations for X and N . This naturally suggests a sequential estimation procedure where the decision to stop is determined by the sample results.

If there is dilution or elimination, the procedure is obviously vitiated. As yet no work has been done to extend the method to estimate these factors. Estimates of mortality for example might be based on a trichotomy or on an intermediate sampling during the removal process. The several sample scheme (sequential or not) would lend itself to this more complicated situation.

6. Methods based on the notion of effort. That the amount of effort expended in obtaining a given sample of a population is proportional to the population density has long been the basis of relative population estimates. Leslie and Davis [20] and independently DeLury [7] showed how absolute estimates could be determined from this information, when the successive samples are removed from the population—as for example occurs in the catch of a fishery. Except for this catch, the population is assumed closed. A model similar to DeLury's is as follows:

Model V

Unknown parameters	$\begin{cases} N_0 = \text{initial population size} \\ k = \text{average probability that an individual is captured by one unit of effort in any time interval} \end{cases}$
Known parameter	$\begin{cases} K_t = \text{total catch up to but not including the } t^{\text{th}} \text{ interval} \end{cases}$
Random variable	$\begin{cases} C_t = \text{catch per unit of effort during the } t^{\text{th}} \text{ interval} \\ (t = 1, 2, \dots, m) \end{cases}$

If the units of effort are independent it follows that

$$E(C_t) = k(N_0 - K_t)$$

With the further assumption that σ_C^2 is approximately constant (which is reasonable for large N_0 unless the cumulative catch represents a large segment of the population by the end of the experiment), least squares estimates of k and N_0 may be found. In particular

$$(23) \quad N_0 \approx \frac{\overline{C} \sum_{t=1}^m (K_t - \bar{K})^2}{\sum_{t=1}^m C_t (K_t - \bar{K})} + \bar{K}$$

If the further assumption of approximate normality of the C_t is made, confidence intervals for N are

$$(24) \quad \bar{K}_1 - \gamma_2 < N < \bar{K}_2 - \gamma_1$$

where γ_1, γ_2 are the roots of the equation,

$$(25) \quad \gamma^2 \left(\left[\sum C_t (K_t - \bar{K}) \right]^2 - q^2 \sum (K_t - \bar{K})^2 \right) - 2\bar{K} \left[\sum (K_t - \bar{K})^2 \right] \left[\sum C_t (K_t - \bar{K}) \right] \gamma + \left(\bar{K} - \frac{q^2}{n} \right) \left(\sum (K_t - \bar{K})^2 \right)^2 = 0$$

where $q = t \frac{(n-2) \cdot S_{C_t}}{1 - \frac{1}{n}}$ and $S_{C_t}^2$ is the estimated error variance

of C_t .

The confidence intervals are obtained by the Fieller technique [29], and hence there is a non-zero probability that the method will give interval estimates of the form $(0, \infty)$ or $(0, N_1)(N_2, \infty)$. Also it should be observed that the model is essentially a conditional one, i.e. conditional upon the values $K_1 K_2 \dots K_m$.

DeLury has also considered the possibility of weighting the least squares estimates, though he suggests that such a procedure may be meaningless if the sampling is not random. This is very likely the case in utilising commercial or sports catch record or in sampling schooling populations for example. For a further discussion of these points and of the method in general reference is made to [7] and [8].

For the case where the effort is constant, Moran [21] has set down a model based on the assumption of random sampling. The model may easily be extended to the case where the effort varies from period to

period. A somewhat more interesting extension is based on a combination of tag and sample and catch per unit of effort methods. The case of a closed population is still considered.

Model VI

Unknown parameters	$\left\{ \begin{array}{l} N_0 = \text{initial population size} \\ \lambda = \text{probability that a unit of effort captures one member of the population} \end{array} \right.$
Known parameters	$\left\{ \begin{array}{l} K_1 = \text{cumulative number removed from population up to but not including the } i^{\text{th}} \text{ sample} \\ e_i = \text{number of units of effort expended on the } i^{\text{th}} \text{ sample} \\ t_i = \text{the number of tagged individuals remaining in the population at the time the } i^{\text{th}} \text{ sample is taken} \end{array} \right.$
Random variables	$\left\{ \begin{array}{l} n_i = \text{size of the } i^{\text{th}} \text{ sample or catch, which is then removed from the population} \\ x_i = \text{number of tagged individuals in the } i^{\text{th}} \text{ sample} \\ (i=1, 2 \dots r) \end{array} \right.$

It is assumed that n_i has a Poisson distribution and given n_i, x_i also has a Poisson distribution. With the usual proviso that the units of effort are independent

$$(26) \quad E(n_i) = k(N_0 - K_1) e_i$$

$$(27) \quad \text{and} \quad E(x_i | n_i) = \frac{n_i t_i}{N - K_1}$$

Hence

Hence

$$(28) \quad \Pr(n_1, n_2, \dots, n_r; x_1, x_2, \dots, x_r) \\ = \prod_{i=1}^r \left[e^{-(N-K_1)\lambda} \frac{[(N-K_1)\lambda e_i]^{n_i}}{n_i!} e^{-\frac{n_i t_i}{N-K_1}} \left(\frac{n_i t_i}{N-K_1} \right)^{x_i} \cdot \frac{1}{x_i!} \right]$$

The maximum likelihood equations for k and N_0 are

$$(29) \quad \hat{\lambda} = \frac{\sum_{i=1}^r n_i}{\sum_{i=1}^r e_i (N-K_1)}$$

$$(30) \quad \sum_{i=1}^r \frac{n_i t_i}{(N-K_1)^2} + \sum_{i=1}^r \frac{n_i - x_i}{N-K_1} = \frac{\sum_{i=1}^r n_i \sum_{i=1}^r e_i}{\sum_{i=1}^r e_i (N-K_1)}$$

The inverse of the variance-covariance matrix of k and N_0 , expressed in terms of the K_1 is:

$$(31) \quad \begin{pmatrix} \frac{1}{k} \sum_{i=1}^r e_i (N_0 - K_1) & \sum_{i=1}^r e_i \\ \sum_{i=1}^r e_i & k \sum_{i=1}^r \frac{e_i}{N_0 - K_1} \end{pmatrix}$$

7. Further Problems. Each of the models set up and others that have been considered involves one or more assumptions which it is difficult or impossible to verify directly. For example underlying the tag-and-sample models there is the assumption that tagged members of the population behave similarly to the untagged members, at least in respect to recapture. A primary assumption of the methods based on effort is that catchability is constant.

Some empirical studies have been made to verify the estimates of populations by sampling methods. In some experiments conducted on fresh-water lakes the whole population has been poisoned out (a procedure that can hardly be recommended as an enumeration procedure except where the elimination of the existent populations has been the primary aim). The agreement has been satisfactory for some species but not for all—for example cf. Carlander [3]. It should be remarked that sampling methods have often been necessary in connection with the estimates determined from the dead recoveries.

Such methods of verification have at best limited application. It is necessary to design sampling experiment specifically for this purpose. In this connection it is suggested that combinations of the various methods outlined may be useful. This has been proposed by DeLury, [9]: his discussion of the underlying assumptions of sample census methods is particularly pertinent.

Such combinations, of which Model VI is an example, may also yield more information than the application of a single method. Of course if the sampling is being done by a succession of commercial catches, Model VI is the appropriate one rather than Model I—though the heterogeneities introduced by such commercial catch sampling may suggest a regression model, i.e. an extension of Model III.

In Model I the numbers tagged and sampled were regarded as parameters; in actual fact they may also be random variables. For example if the sampling is done by a commercial catch, the proportion of the catch from the population to be estimated may be determined only by sampling experiments. This situation complicates the interval estimates

and while a crude determination of a confidence interval for N_0 is possible by a sequence procedure, this patently wastes information. The several variations of this situation that may arise suggest the necessity of a study of confidence intervals in connection with compound distributions.

Referring again to Model I, it may be recognized from the outset that heterogeneity exists within the sampling procedure. If it is possible to subdivide the tagging and sampling into periods (by time or area for example), within each of which random sampling may be assumed then it is possible to obtain consistent estimates, though the interval estimation problem is unsolved. This situation was first considered by Schaeffer [26].

An obvious extension of Model V is to assume that the probability of capture, rather than being constant over the population, is itself a random variable. The distribution of the probability of capture may perhaps be related to the expected catch in any time interval. Additional information is available if different methods of capture are being used simultaneously--in fact in this case the restriction that the population is closed may be relaxed and an estimation procedure set up for the population size at each time interval.

As has been inferred, the interval estimation problem remains unsolved for many of the models, except for the large sample results. Correspondingly, the sample theory of tests in connection with such models has been given almost no attention. Some simple applications of the χ^2 test have been given by Leslie [18] and by the author [5]. As more intricate experiments are designed and more careful control plans

undertaken it will be necessary to consider tests for recruitment and mortality rates, for example.

The complexities of estimating the birth, death, emigration and immigration rates indicate that it will be necessary to set up special experiments to adequately determine these factors. Some of the experiments set up by Jackson [15] where marking and recovery were carried on in a series of adjacent areas were designed for this purpose. Random walk theory has been applied in one special situation by Gilmour, Waterhouse and McIntyre [11]. The study of birth and death processes, and of processes associated with random as well as migratory movement, is necessarily associated with the population estimation problem and the latter will be completely solved only when the problems associated with these stochastic processes are resolved.

REFERENCES

A. Population Estimation

- [1] Norman T. J. Bailey, "On estimating the size of mobile populations from recapture data" , Biometrika 38 (1951) pp. 293-306.
- [2] ----- "Improvements in the interpretation of recapture data",
Journal of Animal Ecology 21 (1952) pp. 120-127.
- [3] Kenneth D. Carlander and William M. Lewis, "Some precautions in estimating fish populations", Prog. Fish. Cult. 10 (1948) pp. 134-137.
- [4] Douglas G. Chapman, "A mathematical study of confidence limits of salmon populations calculated from sample tag ratios",
Internat. Pacific Salmon Fisheries Comm. Bulletin No. 2 (1948) pp. 69-85.
- [5] ----- "Some properties of the hypergeometric distribution with applications to zoological sample censuses", Univ. California Publ. Stat. 1 (1951) pp. 131-160.
- [6] ----- "Inverse, multiple and sequential sample censuses",
Biometrics 8 (1952) pp. 286-306.
- [7] D. B. DeLury, "On the estimation of biological populations",
Biometrics 3 (1947) pp. 145-167.
- [8] ----- "On the planning of experiments for the estimation of fish populations", Jour. Fish. Res. Bd. Can. 8 (1951) pp. 281-307.
- [9] ----- "On the assumptions underlying estimates of mobile populations", Biostatistics Conference, Ames, Iowa, 1952 (to be published).

- [10] W. H. Dowdeswell, R. A. Fisher and E. B. Ford, "The quantitative study of population in the Lepidoptera. I. Polyommatus Icarus Rott." Ann. Eugen., Lond., 10 (1950) pp. 123-136.
- [11] D. Gilmour, D. F. Waterhouse and G. A. McIntyre, "An account of experiments undertaken to determine the natural population density of the sheep blowfly Lucilia cuprina Wied". Bull. Aust. Council Sci. Ind. Res. no. 195.
- [12] Leo A. Goodman, "Sequential sampling tagging for population size problems", Ann. Math. Statistics 24 (1953) pp. 56-69.
- [13] Paul G. Hoel, "The accuracy of sampling methods in ecology", Ann. Math. Statistics 14 (1943) pp. 289-300.
- [14] C. H. N. Jackson, "Some new methods in the study of Glossina morsitans. Proc. Zool. Soc. London 4 (1936) pp. 811-896.
- [15] ——— "The analysis of an animal population", J. Animal Ecology 8 (1939) pp. 238-246.
- [16] ——— "The analysis of a tsetse-fly population", Ann. Eugen., Lond., 10 (1940) pp. 332-369.
- [17] P. S. Laplace, "Sur les naissances, les mariages et les morts", Histoire de l'Academie Royale des Sciences Annee 1783 Paris p. 693 (actually published in 1786).
- [18] P. H. Leslie, "The estimation of population parameters from data obtained by means of the capture-recapture method, II The estimation of total numbers", Biometrika, 39 (1952) pp. 363-388.
- [19] P. H. Leslie and Dennis Chitty, "The estimation of population parameters from data obtained by means of the capture-recapture method, I The maximum likelihood equations for estimating the death-rate", Biometrika, 38 (1951) pp. 269-292.

- [20] P. H. Leslie and D. H. S. Davis, "An attempt to determine the absolute number of rats on a given area", J. Animal Ecology 8 (1939) pp. 94-113.
- [21] P. A. P. Moran, "A mathematical theory of animal trapping", Biometrika, 38 (1951) pp. 307-311.
- [21a] ——— "The estimation of death rates from capture-mark-recapture sampling", Biometrika 39 (1952) pp. 181-188.
- [22] J. Neyman, "On the problem of estimating the number of schools of fish", Univ. California Publ. Stat. 1 (1949) pp. 21-36.
- [23] 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, (cf. also later reports).
- [24] William E. Ricker, "Methods of estimating vital statistics of fish populations", Indiana Univ. Publ. Science Series No. 15 (1948).
- [25] Leslie W. Scattergood, "Estimating fish and wildlife populations: a survey of methods", Biostatistics Conference, Ames, Iowa, 1952 (to be published).
- [26] Milner B. Schaeffer, "Estimation of the size of animal populations by marking experiments", U. S. Fish and Wildlife Ser. Fish. Bull. 69 (1951) pp. 191-203.
- [27] Z. E. Schnabel, "Estimation of the total fish population of a lake", Amer. Math. Monthly 45 (1938) pp. 348-352.
- [28] J. G. Skellam, "Studies in statistical ecology. I Spatial pattern", Biometrika, 39 (1952) pp. 346-362.

- [29] E. C. Fieller, "The biological standardization of insulin", Supp. Jour. Royal Stat. Soc. 7 (1940) pp. 1-65.
- [30] C. P. Winsor and G. L. Clarke, "A statistical study of variation in the catch of plankton nets", Jour. Marine Research 3 (1940) pp. 1-34.

Technical Reports Distribution List
Contract N8 onr-520 Task II

Director
Office of Naval Research
Branch Office
346 Broadway
New York 19, N. Y.

Director
Office of Naval Research
Branch Office
1030 East Duane Street
Pasadena 4, California 1

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

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)

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

ANL Corporation
150 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