On Estimating the Size of Mobile Populations from Recapture Data
Author(s): Norman T. J. Bailey
Source: *Biometrika*, Vol. 38, No. 3/4 (Dec., 1951), pp. 293-306
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: https://www.jstor.org/stable/2332575
Accessed: 01-08-2024 23:26 UTC

ON ESTIMATING THE SIZE OF MOBILE POPULATIONS FROM RECAPTURE DATA

By NORMAN T. J. BAILEY, *Department of Human Ecology, University of Cambridge*

CONTENTS

1. Introduction

The estimation of the total size of plant or animal populations is of great importance in a variety of biological problems, which may relate to population growth, ecological adaptation, genetic constitution, natural selection and evolution, and so on. Obvious practical consequences are the maintenance of human food supplies and the control of insect pests. For human communities and populations of sessile organisms procedures employing fixed sampling units are available, but for mobile populations other methods must be used.

The basic technique, which seems to have been first used by Lincoln (1930) to estimate the total number of duck in North America, and is sometimes referred to as the 'Lincoln Index', is as follows. One catches, marks and releases a certain number of animals taken at random from the population. A further random sample is caught and the proportion of marked animals noted. Then the *total* number of marked animals released divided by the proportion of marked animals in the sample captured can clearly be used as an estimate of the total population size. The same method was adopted independently by Jackson (1933), who used it for estimating the true density of tsetse flies. Jackson (1937, 1939) subsequently extended his treatment to allow for both birth- and death-rates, advocating his 'negative' and 'positive' methods. Further work by Jackson (1940) took into account the effects of migration. Recapture methods have also been successfully applied to populations of Lepidoptera. Dowdeswell, Fisher & Ford (1940) caught, marked and released moths on several different days. On each day the marked insects were classified according to the day on which they had previously been released. Analysis of the data, which could be exhibited in a triangular array, or 'trellis' diagram, gave estimates of the size of the moth population for each day. The interpretation of trellis diagrams was further developed by Fisher & Ford (1947). They estimated daily numbers, taking into account a death-rate which was obtained from the average time interval between marking and recapture. A more detailed application of this procedure was made by Dowdeswell *et al.* (1949).

Although in the various papers cited above effective use has been made of *estimates* of population size and of birth- and death-rates, there has been little discussion for the most part of the *precision* of the results obtained. Jackson (1937, 1939) obtained both the population estimate and its variance for the 'negative' and 'positive' methods by fitting a curve

to a series of 'standardized' recapture rates. However, this was not really adequate, as it did not employ satisfactory weighting factors. Schnabel (1938) considered the problem of finding the total number of fish in a lake, and obtained the maximum-likelihood estimate of the population size for repeated sampling of the recapture type. Although approximations to the solution of the likelihood equations were given, there was no discussion of the precision of the results. De Lury (1947) has used a somewhat different approach to the same problem. Specimens caught are removed from the population, and it is assumed that catches are sufficiently large to produce an appreciably diminishing catch per unit effort with repeated catching. An appendix to his paper gives a full discussion of the appropriate statistical model.

It is the purpose of the present paper to consider the precision of estimates of population size, and of birth- and death-rates derived from recapture data. We shall deal first with simple recapture, considering both direct and inverse sampling, and then go on to a new discussion of Jackson's 'negative' and 'positive' methods. Although, in general, the simultaneous estimation of population size and birth- and death-rates from a large trellis diagram is exceedingly tedious, it will be shown that in at least one special case, that of 'triple capture' relating to catching on only three occasions, a satisfactory explicit solution is available. In certain circumstances, therefore, it might be worth while making a special attempt to collect data in this form.

We shall, in the present discussion, assume that we are dealing with closed regions, unaffected by migration. Further, when birth- and death-rates are involved, it will be assumed that the numbers of marked animals are sufficiently large for these rates to operate deterministically.

2. Simple recapture

(*a*) *Direct sampling*

Suppose that we capture, mark and release a animals out of a total population of size x. When the marked animals have freely mingled with the unmarked, we catch a random sample of size n, of which r are found to be marked. We shall assume that n is sufficiently small compared with x for us to be able to ignore the complications of sampling without replacement. The large sample theory of the maximum-likelihood approach will be appropriate if we envisage n, a and x all tending to infinity while keeping constant ratios to each other. The likelihood is

$$e^L = \binom{n}{r}\left(\frac{a}{x}\right)^r\left(\frac{x-a}{x}\right)^{n-r}. \tag{2·1}$$

Therefore

$$L = \text{const.} + (n-r)\log(x-a) - n\log x. \tag{2·2}$$

Although x is in fact discrete, it will be convenient to treat it as though it were a continuous parameter. Differentiating with respect to x gives

$$\frac{\partial L}{\partial x} = \frac{n-r}{x-a} - \frac{n}{x}. \tag{2·3}$$

Hence the maximum-likelihood estimate of x is

$$\hat{x} = an/r. \tag{2·4}$$

Schnabel (1938) discussed the solution of the corresponding likelihood equation for a set of independent samples, e.g.

$$\sum_i (n_i - r_i)/(x - a_i) = \sum_i n_i/x,$$

where the suffix i indicates the values of n, r and a appropriate to the ith sample. She did not, however, go on to examine the precision of the estimate thus obtained.

Differentiating (2·3) with respect to x, and then taking the expectation provides the required information in the case of a single sample. Thus

$$I_x = -E\frac{\partial^2 L}{\partial x^2} = an/\{x^2(x-a)\}. \tag{2·5}$$

Using the value of x given by (2·4) we can then write

$$\operatorname{var}\hat{x} = I_x^{-1} = a^2 n(n-r)/r^3. \tag{2·6}$$

Now $\hat{x}$ has an infinite expectation if we admit the value $r = 0$. On the other hand, this value will occur so rarely in large samples, for which $m = Er$ is not small, that we can choose to exclude it. With this convention it can be shown that the relative bias in the expectation of $\hat{x}$ is of order m^{-1}. For, adopting the technique employed by Anscombe (1948), we can expand the right-hand side of (2·4) as a Taylor series in powers of $t = r - m$. Taking expectations of both sides then leads to an asymptotic formula for $E\hat{x}$ $(r \neq 0)$. We find

$$\underset{r\neq 0}{E}\,\hat{x} \sim x\left(1 + \frac{x-a}{an} + O(m^{-2})\right) = x\left(1 + \frac{1}{m} + O(m^{-2})\right), \tag{2·7}$$

if m is small compared with n.

Suppose $x = 1000$, and $n = a = 100$. Then $m = 10$, and although the sample size is as large as 100, there is an average relative bias of the order of 10 % in the maximum-likelihood estimate of x. A slightly adjusted estimate which suggests itself is

$$\breve{x} = a(n+1)/(r+1). \tag{2·8}$$

The expectation of $\breve{x}$ can be found exactly. We have

$$E\breve{x} = \sum_{r=0}^{n} \frac{a(n+1)}{(r+1)} \binom{n}{r} \left(\frac{a}{x}\right)^r \left(\frac{x-a}{x}\right)^{n-r}$$

$$= x \sum_{r=0}^{n} \binom{n+1}{r+1} \left(\frac{a}{x}\right)^{r+1} \left(\frac{x-a}{x}\right)^{n-r}.$$

Therefore

$$E\breve{x} = x\left\{1 - \left(\frac{x-a}{x}\right)^{n+1}\right\}. \tag{2·9}$$

For large n

$$\left(\frac{x-a}{x}\right)^{n+1} \sim e^{-m}, \tag{2·10}$$

which is quite small even for moderate m. For the values given above, the relative bias is now less than $2{\cdot}5 \times 10^{-5}$. A satisfactory expression for the variance of $\breve{x}$ is somewhat more difficult to find. However, using the expansion technique referred to above we can derive the asymptotic series

$$E\breve{x}^2 \sim \frac{a^2(n+1)^2}{m^2}\left(1 + \frac{1-2p}{m} + \ldots\right), \tag{2·11}$$

where $p = a/x$. We now obtain the variance from the relation

$$\operatorname{var}\breve{x} = E\breve{x}^2 - (E\breve{x})^2. \tag{2·12}$$

We could obtain a single-series expression for the variance by substituting (2·11) in (2·12), and writing

$$(E\check{x})^2 \sim x^2 = a^2n^2/m^2 = \frac{a^2(n+1)^2}{m^2}\left(1-\frac{2}{n}+\ldots\right),$$

but as this would involve the consideration of inverse powers of n, there seems little to be gained by this device.

Fortunately, there exists a nearly unbiased estimate of the variance of $\check{x}$, which is quite convenient for use with samples that are not too small.

It is easy to show that

$$E\,\frac{a^2(n+1)(n+2)}{(r+1)(r+2)} = x^2\left\{1-\left(\frac{x-a}{x}\right)^{n+2}-(n+2)\left(\frac{a}{x}\right)\left(\frac{x-a}{x}\right)^{n+1}\right\} \tag{2·13}$$

$$\sim x^2(1-m\,e^{-m}), \quad \text{for large } n. \tag{2·14}$$

Therefore, if we write

$$T = \frac{a^2(n+1)^2}{(r+1)^2} - \frac{a^2(n+1)(n+2)}{(r+1)(r+2)} = \frac{a^2(n+1)(n-r)}{(r+1)^2(r+2)}, \tag{2·15}$$

then

$$ET = E\check{x}^2 - x^2(1-m\,e^{-m}) \sim \sigma_{\check{x}}^2 + x^2 m\,e^{-m}, \tag{2·16}$$

neglecting quantities of order e^{-m}. Now we know from (2·5) that $\sigma_{\check{x}}^2$ is of order x^2/m, so that

$$ET \sim \sigma_{\check{x}}^2(1+m^2e^{-m}). \tag{2·17}$$

The relative bias of T in our previous example is thus of order $2{\cdot}5 \times 10^{-3}$, which should be quite satisfactory for practical purposes.

It is worth mentioning here that in certain ecological problems we may be more concerned to use the reciprocal of the population size, $1/x$, as the appropriate index, rather than the population size itself. It follows from (2·4) that the maximum-likelihood estimate of this reciprocal is r/an. Consideration of the binomial distribution in (2·1) shows that this estimate is unbiased and has variance $(x-a)/(anx^2)$.

(*b*) *Inverse sampling*

Now in the above treatment we studied the mean and variance of estimates of the population size x in relation to repeated sampling for which the total size of the samples drawn was regarded as fixed. Various writers (e.g. Haldane, 1945) have considered the method of *inverse sampling*, according to which random sampling is continued until a certain predetermined number of individuals with a particular attribute has been obtained. The use of inverse sampling with simple recapture results in an appreciable simplification—an unbiased estimate of population size and an exact value of the sampling variance are available, even in the most general case of sampling without replacement. It should, of course, be mentioned in passing that it may not always be possible to adopt this method of sampling in practice.

Suppose we catch, mark and release a animals in a total population of size x. We then draw a random sample by the inverse method, continuing our catching until we have m marked animals altogether. The total number of animals in the sample is n, which is now the random variable under consideration.

The probability of obtaining a sample of size n is the probability of drawing first a sample of size $(n-1)$ containing $(m-1)$ marked animals, followed by the drawing of just one further

marked animal. Having regard to the existence of sampling without replacement, the likelihood is clearly

$$P(n) = \frac{\binom{a}{m-1}\binom{x-a}{n-m}}{\binom{x}{n-1}} \frac{(a-m+1)}{(x-n+1)} \tag{2·18}$$

$$= \frac{a}{x}\binom{a-1}{m-1}\binom{x-a}{n-m}\Big/\binom{x-1}{n-1}, \tag{2·19}$$

where $0 \leqslant m \leqslant a$ and $m \leqslant n \leqslant x+m-a$.

It follows from (2·19) that

$$\sum_{n=m}^{x+m-a}\binom{x-a}{n-m}\Big/\binom{x-1}{n-1} = \frac{x}{a\binom{a-1}{m-1}}\sum_{n=m}^{x+m-a}P(n) = x\Big/\left\{a\binom{a-1}{m-1}\right\}. \tag{2·20}$$

The expectations of n and $n(n+1)$, which we shall have occasion to use below, are readily derived as follows:

$$En = \sum_{n=m}^{x+m-a}\frac{na}{x}\binom{a-1}{m-1}\binom{x-a}{n-m}\Big/\binom{x-1}{n-1}$$

$$= a\binom{a-1}{m-1}\sum_{n=m}^{x+m-a}\binom{x-a}{n-m}\Big/\binom{x}{n}. \tag{2·21}$$

The summation on the right-hand side of (2·21) is obtained immediately from (2·20) on writing $x+1$, $a+1$, $n+1$ and $m+1$ for x, a, n and m respectively. Therefore

$$En = a\binom{n-1}{m-1}(x+1)\Big/(a+1)\binom{a}{m} = \frac{m(x+1)}{(a+1)}. \tag{2·22}$$

Similarly,

$$En(n+1) = \sum_{n=m}^{x+m-a}\frac{n(n+1)a}{x}\binom{a-1}{m-1}\binom{x-a}{n-m}\Big/\binom{x-1}{n-1}$$

$$= (x+1)a\binom{a-1}{m-1}\sum_{n=m}^{x+m-a}\binom{x-a}{n-m}\Big/\binom{x+1}{n+1}. \tag{2·23}$$

To evaluate the summation on the right-hand side of (2·23) we write $x+2$, $a+2$, $n+2$ and $m+2$ for x, a, n and m in equation (2·20). Therefore

$$En(n+1) = (x+1)a\binom{a-1}{m-1}(x+2)\Big/(a+2)\binom{a+1}{m+1} = \frac{m(m+1)(x+1)(x+2)}{(a+1)(a+2)}. \tag{2·24}$$

Now, strictly regarded, x can take only integral values. Thus the greatest value attained by the likelihood is derived by considering the ratio

$$\frac{P(n \mid x+1)}{P(n \mid x)} = \frac{(x-a+1)(x-n+1)}{(x+1)(x-a-n+m+1)}, \tag{2·25}$$

using (2·19). It follows from (2·25) that

$$\left.\begin{aligned} P(n \mid x+1) &\gtrless P(n \mid x), \\ \text{according as} \quad x &\lessgtr \frac{an}{m} - 1. \end{aligned}\right\} \tag{2·26}$$

Unless an/m is an integer, the likelihood attains its greatest value when x is equal to the integral part of an/m. If an/m is integral the greatest value is attained at both $\left(\frac{an}{m}-1\right)$ and $\frac{an}{m}$.

In practice x is usually sufficiently large for us to be able to ignore its discreteness. In any case we are led to consider the estimates of the general type an/m. It follows from (2·22) that

$$x' = \frac{n(a+1)}{m} - 1 \tag{2·27}$$

is an unbiased estimate of x, for any values of x, n, a and m whatever. In large samples x' approximates to the maximum-likelihood solution. The exact small sample variance of x' is also easily evaluated, using the results in (2·22) and (2·24). We find

$$\operatorname{var} x' = \frac{(a-m+1)(x+1)(x-a)}{m(a+2)}. \tag{2·28}$$

3. Jackson's 'negative' and 'positive' methods

(*a*) *The 'negative' method*

The 'negative' method was first described by Jackson (1937), and simply consists in the repeated catching, marking and releasing of animals on successive occasions, making no record of the number of recaptures until the last occasion. Under certain circumstances, when the preliminary marking can be carried out by relatively unskilled workers while the final scoring of the numbers of recaptures requires a more elaborate set-up, this procedure may be very suitable.

As mentioned in the introduction above, Jackson's method of fitting a curve to 'standardized' recapture rates is not really satisfactory owing to the lack of proper weighting factors. The present section will develop the maximum likelihood approach.

Suppose that on the jth day previous to the last day a_j freshly marked animals are released. Then, if we can assume a constant death-rate γ, $a_j e^{-\gamma j}$ of these animals will survive to stand a chance of recapture on the last day. Of a total of n captures on the last day, let r_j have been marked on the jth day previously and let r_0 be unmarked. With animals bearing more than one mark it will be convenient to take notice of the earliest mark only, although this will result in the loss of a slight amount of information. Then if x is the size of the population on the last day, we can write the likelihood as

$$e^L \propto \prod_{j=1} \left(\frac{a_j e^{-\gamma j}}{x}\right)^{r_j} \left(\frac{x - \sum_{j=1} a_j e^{-\gamma j}}{x}\right)^{r_0}. \tag{3·1}$$

Therefore

$$L = \text{const.} - n \log x - \gamma \sum_{j=1} j r_j + r_0 \log (x - \sum_{j=1} a_j e^{-\gamma j}). \tag{3·2}$$

Let us write

$$\left.\begin{aligned} F(\gamma) &= \sum_{j=1} a_j e^{-\gamma j}, \\ F'(\gamma) &= -\sum_{j=1} j a_j e^{-\gamma j}, \\ F''(\gamma) &= \sum_{j=1} j^2 a_j e^{-\gamma j}. \end{aligned}\right\} \tag{3·3}$$

Further, put
$$\sum_{j=1} jr_j = A. \tag{3·4}$$

Then the likelihood equations are
$$\frac{\partial L}{\partial x} = -\frac{n}{x} + \frac{r_0}{x - F(\gamma)} = 0 \tag{3·5}$$

and
$$\frac{\partial L}{\partial \gamma} = -A - \frac{r_0 F'(\gamma)}{x - F(\gamma)} = 0. \tag{3·6}$$

We can eliminate x from (3·5) and (3·6) to give
$$\sum_{j=1} c_j e^{-\hat{\gamma} j} = 0, \tag{3·7}$$

where
$$c_j = a_j\left(j - \frac{A}{n - r_0}\right). \tag{3·8}$$

The quantities c_j are readily determined from the observations, and the required root, $\hat{\gamma}$, of (3·7) is easily obtained in practice by the usual iterative procedure. Substituting this value of γ in (3·5) then gives
$$\hat{x} = nF(\hat{\gamma})/(n - r_0). \tag{3·9}$$

Further differentiation of (3·5) and (3·6), followed by taking expectations, gives the amounts of information
$$\left.\begin{aligned} I_{xx} &= nF/\{x^2(x-F)\}, \\ I_{x\gamma} &= -nF'/\{x(x-F)\}, \\ I_{\gamma\gamma} &= n\{F'^2 + F''(x-F)\}/\{x(x-F)\}, \end{aligned}\right\} \tag{3·10}$$

where F, F' and F'' are calculated for $\gamma = \hat{\gamma}$. It now follows from (3·10) that
$$\operatorname{var} \hat{x} = \frac{x^2}{n}\left\{\frac{xF''}{FF'' - F'^2} - 1\right\}, \tag{3·11}$$

and
$$\operatorname{var} \hat{\gamma} = xF/\{n(FF'' - F'^2)\}. \tag{3·12}$$

It is, of course, important to realize that the foregoing treatment assumes that the death-rate operates 'deterministically'. This will be satisfactory if the numbers of marked animals involved are sufficiently large; otherwise, 'stochastic' fluctuations may become important, and it is possible that, under such conditions, the variances of the estimates would be larger.

Example. Jackson (1939, p. 241) gives some data for the 'negative' method carried out on a tsetse fly population. In our notation we have:

j	6	5	4	3	2	1	0
a_j	1262	1086	1299	1401	1198	1183	—
r_j	1	5	17	28	48	70	1389

The total number, n, captured on the last day ($j = 0$) was 1558. It is not clear what convention was adopted with regard to flies marked more than once, but for the purpose of this illustration we will assume that on the last day notice was taken of the first mark only.

The quantity A in (3·4) is 349, and the six values of c_j are easily found to be

$$c_1 = -1260{\cdot}00, \quad c_4 = 2513{\cdot}45,$$
$$c_2 = -\ 77{\cdot}98, \quad c_5 = 3187{\cdot}31,$$
$$c_3 = \ 1309{\cdot}81, \quad c_6 = 4965{\cdot}86.$$

We now have to solve (3·7). Writing $e^{-\gamma} = \mu$ we take the trial values $\mu = 0{\cdot}55$ and $0{\cdot}54$; the right-hand side of (3·7) is then $+29{\cdot}193$ and $-13{\cdot}692$ respectively. Interpolating gives

$$\left.\begin{aligned} \hat{\mu} &= 0{\cdot}543, \\ \hat{\gamma} &= -\log_e \hat{\mu} = 0{\cdot}611. \end{aligned}\right\}$$

The next stage is to compute the F's from (3·3). We find

$$\left.\begin{aligned} F(\hat{\gamma}) &= \ 1416{\cdot}444, \\ F'(\hat{\gamma}) &= -2923{\cdot}873, \\ F''(\hat{\gamma}) &= \ 8327{\cdot}083. \end{aligned}\right\}$$

We are now in a position to substitute in (3·9), (3·11) and (3·12), obtaining finally the following estimates with attached standard errors:

$$\hat{x} = 13{,}060 \pm 1890,$$
$$\hat{\gamma} = \ 0{\cdot}611 \pm 0{\cdot}060.$$

(b) The 'positive' method

The 'positive' method was also first described by Jackson (1937). This consists in marking a large number of animals on a single occasion. On each of a number of later occasions the proportion of marked animals is recorded. The initial population can then be estimated from these data, taking the birth-rate into account. Jackson again used the method of fitting a curve to 'standardized' recapture rates, as in the 'negative' method, and the same objection applies.

The analysis of this type of data is of necessity more complicated as it entails sampling on a number of occasions, whereas in the 'negative' method sampling variation occurs only on the last day. We shall find that the maximum-likelihood treatment is more involved, but is fairly readily amenable to the usual method of calculating scores and information functions.

In order to render the mathematical treatment as simple as possible it is desirable that nearly all the animals captured on each occasion be released again, otherwise, unless numbers are very large, it would be necessary to take into account the depletion of marked animals by the process of recapture.

If we introduce a death-rate in this case we shall find that the corresponding factors in the likelihood cancel, as we should expect. It is, however, necessary to consider a birth-rate, β. Thus after j days the initial population of size x will have become $x e^{\beta j}$. Suppose that, initially, we release a marked animals, and that j days later we catch n_j animals, r_j of which are marked. It is easy to see that

$$L = \text{const.} - \log x \sum_{j=1} n_j - \beta \sum_{j=1} j n_j + \sum_{j=1} (n_j - r_j) \log (x e^{\beta j} - a). \tag{3·13}$$

The likelihood scores are $$S(x) = \frac{\partial L}{\partial x} = \sum_{j=1} \frac{n_j - r_j}{x - a e^{-\beta j}} - \frac{1}{x} \sum_{j=1} n_j, \tag{3·14}$$

and $$S(\beta) = \frac{\partial L}{\partial \beta} = x \sum_{j=1} \frac{j(n_j - r_j)}{x - a e^{-\beta j}} - \sum_{j=1} j n_j. \tag{3·15}$$

Proceeding in the usual way, we find

$$\left.\begin{aligned} I_{xx} &= \frac{a}{x^2} \sum_{j=1} n_j/(x e^{\beta j} - a), \\ I_{x\beta} &= \frac{a}{x} \sum_{j=1} j n_j/(x e^{\beta j} - a), \\ I_{\beta\beta} &= a \sum_{j=1} j^2 n_j/(x e^{\beta j} - a). \end{aligned}\right\} \tag{3·16}$$

Using (3·14), (3·15) and (3·16), we can follow the normal method of scoring. Thus if θ is a column vector representing approximate values of $\hat{x}$ and $\hat{\beta}$, if S is the corresponding vector of scores, and if I is the information matrix, then improved approximations to the maximum-likelihood estimates are given by θ_1, where

$$\theta_1 = \theta + I^{-1} S.$$

As with the death-rate in the previous section, so here we have assumed that numbers are sufficiently large for us to work with a deterministic birth-rate. With small numbers we should have to consider stochastic fluctuations.

4. TRELLIS DIAGRAMS

Jackson's 'negative' and 'positive' methods, discussed in the previous section, are clearly special cases of a more general procedure in which all, or at least the majority, of the animals caught on a particular day are marked and released; at the same time all captured animals are carefully scored with respect to the days on which they were previously marked. As mentioned in the last section, it is convenient, when an animal has more than one marking, to take notice of the earliest mark only, although this must result in a slight loss of information. The record showing the numbers of captures, releases and recaptures can then be set out in a triangular array. Such arrays have been used with great effect by Dowdeswell *et al.* (1940, 1949) and Fisher & Ford (1947) in the analysis of Lepidoptera populations. An example of such an array for three days only is shown in § 5 of the present paper. Fisher & Ford have analysed these trellis diagrams, taking into account a death-rate estimated from the average time interval between release and recapture. Daily estimates of population size can then be made. While this method is extremely valuable in practice it is difficult to assess the precision of the results obtained.

Let us introduce a general notation for these arrays. On the jth day let the total catch be n_j, of which n_{ji} were first marked on the ith day; while n_{j0} are unmarked. Further, let s_j freshly marked animals be released on the jth day. The total population on the jth day is to be represented by N_j, of which N_{ji} were first marked on the ith day, and N_{j0} are unmarked. Then we can attempt a large sample theory using the method of maximum likelihood, by adopting the following approach.

If we can assume that the N_j and N_{ji} are sufficiently large for us to ignore the effects of sampling without replacement, then the likelihood is evidently given by

$$e^L \propto \prod_{j=2}^{k} \prod_{i=0}^{j-1} (N_{ji}/N_j)^{n_{ji}}, \tag{4·1}$$

where the trellis diagram extends over k days in all. If we can further assume the existence of constant birth- and death-rates, we have

$$\left.\begin{aligned} N_{ji} &= s_i e^{-\gamma(j-i)} \quad (i = 1, \ldots, (j-1)), \\ N_j &= N_1 e^{(\beta-\gamma)(j-1)} \quad (j = 2, \ldots, k), \\ \text{and} \quad N_{j0} &= N_j - \sum_{i=1}^{j-1} N_{ji}. \end{aligned}\right\} \tag{4·2}$$

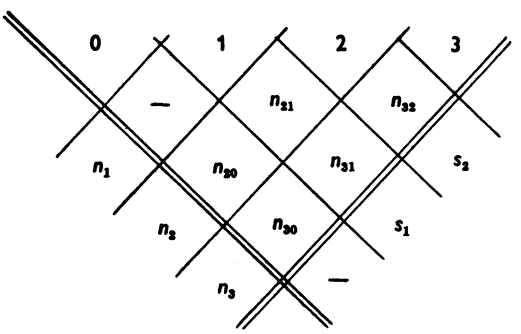
It will be realized that while the whole population is subject to both birth and death, the group of marked individuals suffers depletion by death but gains new recruits only from the release of newly marked specimens.

We can now substitute (4·2) in (4·1) and proceed with the ordinary maximum-likelihood estimation of β, γ and, say, N_1. This is a laborious undertaking in general, and in any case is hardly applicable to data where the numbers of recaptures are low, as with much of the Lepidoptera material, for the reasons discussed in § 2 above, though it might be worth while with the fairly abundant data such as is often available with tsetse fly populations.

However, this approach is susceptible of a fairly simple treatment if we restrict ourselves to catching on three occasions only. This will now be discussed in the following section.

5. The 'triple-catch' method

Suppose that we confine ourselves to catching on only three occasions. Then in the notation of § 4, the numbers caught, released and recaptured are as shown below:



Let us assume, as in the two previous sections, that it is legitimate to work with deterministic birth- and death-rates, which are taken to be β and γ respectively. Thus a group of N individuals gives rise to a group of size $N e^{(\beta-\gamma)t}$ after time t. Note, however, that newly born or emerged individuals are always unmarked, so that the *number* of marked individuals is subject only to the death-rate. Also suppose that only a few captured specimens are not subsequently released again, so that the numbers of animals marked on different days in the population are not appreciably affected.

Finally, let the time intervals between days 1 and 2, and between days 2 and 3, be t_1 and t_2 respectively. Then, in the notation of § 4, the expected numbers of marked and unmarked animals in the population are given by

$$\left.\begin{aligned} N_{21} &= s_1 e^{-\gamma t_1}, & N_{32} &= s_2 e^{-\gamma t_2}, \\ N_{20} &= N_1 e^{(\beta-\gamma)t_1} - s_1 e^{-\gamma t_1}, & N_{31} &= s_1 e^{-\gamma(t_1+t_2)}, \\ N_2 &= N_1 e^{(\beta-\gamma)t_1}, & N_{30} &= N_1 e^{(\beta-\gamma)(t_1+t_2)} - s_2 e^{-\gamma t_2} - s_1 e^{-\gamma(t_1+t_2)}, \\ & & N_3 &= N_1 e^{(\beta-\gamma)(t_1+t_2)}. \end{aligned}\right\} \tag{5·1}$$

The general maximum-likelihood treatment is simplified if we put

$$\left.\begin{aligned} N_2 &= N_1 e^{(\beta-\gamma)t_1} = x, \\ e^{\beta t_2} &= \lambda, \\ e^{-\gamma t_1} &= \mu. \end{aligned}\right\} \tag{5·2}$$

The likelihood can now be written

$$e^L \propto x^{-(n_2+n_3)}(x-\mu s_1)^{n_{20}}(\lambda x-\mu s_1-s_2)^{n_{30}}\lambda^{-n_3}\mu^{(n_{21}+n_{31})}. \tag{5·3}$$

The maximum-likelihood equations are therefore

$$\left.\begin{aligned} \frac{\partial L}{\partial x} &= -\frac{n_2+n_3}{x} + \frac{n_{20}}{x-\mu s_1} + \frac{n_{30}\lambda}{\lambda x-\mu s_1-s_2} = 0, \\ \frac{\partial L}{\partial \lambda} &= -\frac{n_3}{\lambda} + \frac{n_{30}x}{\lambda x-\mu s_1-s_2} = 0, \\ \frac{\partial L}{\partial \mu} &= \frac{n_{21}+n_{31}}{\mu} - \frac{n_{20}s_1}{x-\mu s_1} - \frac{n_{30}s_1}{\lambda x-\mu s_1-s_2} = 0. \end{aligned}\right\} \tag{5·4}$$

It is worth observing that we are estimating three parameters, and have just three available degrees of freedom, one from the sample on day 2 and two from the sample on day 3. This situation has been discussed by Bailey (1951), who pointed out that the equations given by setting the expected numbers equal to their expectations provide maximum-likelihood solutions, and are often simpler to solve than the likelihood equations themselves. At the same time it is necessary to check that the likelihood equations regarded as linear functions of the observations, are linearly independent. The appropriate test, that the Jacobian matrix of the expectations with respect to the parameters to be estimated shall have rank equal to the number of those parameters, is easily seen to be satisfied in the present case. Thus we can either solve (5·4) directly, or we can solve the alternative equations, which, using (5·1) and (5·2), are evidently

$$\left.\begin{aligned} n_{21} &= \frac{N_{21}}{N_2} n_2 = n_2 \mu s_1/x, \\ n_{32} &= \frac{N_{32}}{N_3} n_3 = n_3 s_2/\lambda x, \\ n_{31} &= \frac{N_{31}}{N_3} n_3 = n_3 \mu s_1/\lambda x, \end{aligned}\right\} \tag{5·5}$$

omitting the other two linearly dependent equations. From (5·5) we have, immediately,

$$\left.\begin{aligned} \hat{x} &= s_2 n_2 n_{31}/n_{21} n_{32}, \\ e^{\beta t_2} = \hat{\lambda} &= n_{21} n_3/n_2 n_{31}, \\ e^{-\gamma t_1} = \hat{\mu} &= s_2 n_{31}/s_1 n_{32}. \end{aligned}\right\} \tag{5·6}$$

The large sample variances of these estimates can be most easily obtained from Fisher's formula for the variance of a function of the observations

$$\operatorname{var} T = \sum_i m_i \left(\frac{\partial T}{\partial a_i}\right)^2 - n\left(\frac{\partial T}{\partial n}\right)^2 \Bigg|_{a_i = m_i}, \tag{5·7}$$

where T is a function of the observations a_i, for which the expectations are m_i, the total sample size being n. In the present case there will be terms like $n(\partial T/\partial n)^2$ corresponding to both n_2 and n_3. Thus we have

$$\left.\begin{aligned} \operatorname{var} \hat{x} &= \hat{x}^2\left(\frac{1}{n_{21}} + \frac{1}{n_{32}} + \frac{1}{n_{31}} - \frac{1}{n_2}\right), \\ \operatorname{var} \hat{\lambda} &= \hat{\lambda}^2\left(\frac{1}{n_{21}} + \frac{1}{n_{31}} - \frac{1}{n_2} - \frac{1}{n_3}\right), \\ \operatorname{var} \hat{\mu} &= \hat{\mu}^2\left(\frac{1}{n_{32}} + \frac{1}{n_{31}}\right). \end{aligned}\right\} \tag{5·8}$$

The large sample variances of $\hat{\beta}$ and $\hat{\gamma}$ are therefore

$$\left.\begin{aligned} \operatorname{var} \hat{\beta} &= \operatorname{var} \hat{\lambda}/(\partial\lambda/\partial\beta)^2 = \left(\frac{1}{n_{21}} + \frac{1}{n_{31}} - \frac{1}{n_2} - \frac{1}{n_3}\right)\Big/ t_2^2, \\ \operatorname{var} \hat{\gamma} &= \operatorname{var} \hat{\mu}/(\partial\mu/\partial\gamma)^2 = \left(\frac{1}{n_{32}} + \frac{1}{n_{31}}\right)\Big/ t_1^2. \end{aligned}\right\} \tag{5·9}$$

The results given in (5·8) can be verified by the alternative procedure of calculating and inverting the information matrix, though this is rather more lengthy.

Having regard to the results of § 2, it is evident that the above formulae require n_{21}, n_{32} and n_{31} all to be fairly large. However, estimates of the type shown in (2·8) will be approximately unbiased for much smaller values. There is therefore some advantage in using the adjusted estimates

$$\left.\begin{aligned} \breve{x} &= s_2(n_2+1)\, n_{31}/(n_{21}+1)\,(n_{32}+1), \\ e^{\beta t_2} = \breve{\lambda} &= n_{21}(n_3+1)/n_2(n_{31}+1), \\ e^{-\gamma t_1} = \breve{\mu} &= s_2 n_{31}/s_1(n_{32}+1). \end{aligned}\right\} \tag{5·10}$$

As in § 2 there is some difficulty in calculating the population values of the variances of estimates given in (5·10), but again it is possible to obtain approximately unbiased *estimates* of the variances of $\breve{x}$, $\breve{\lambda}$ and $\breve{\mu}$. Consider, for example, the variance of $\breve{x}$ in (5·10). In a manner similar to that adopted in § 2 for simple recapture, we have

$$E\frac{(n_2+1)\,(n_2+2)\, n_{31}(n_{31}-1)}{(n_{21}+1)\,(n_{21}+2)\,(n_{32}+1)\,(n_{32}+2)} = E\frac{(n_2+1)\,(n_2+2)}{(n_{21}+1)\,(n_{21}+2)}\, E\frac{n_{31}(n_{31}-1)}{(n_{32}+1)\,(n_{32}+2)} \sim \frac{N_2^2}{N_{21}^2}\frac{N_{31}^2}{N_{32}^2} = \frac{x^2}{s_2^2}. \tag{5·11}$$

Consider, therefore, the statistic

$$U = \check{x}^2 - \frac{s_2^2(n_2+1)(n_2+2)n_{31}(n_{31}-1)}{(n_{21}+1)(n_{21}+2)(n_{32}+1)(n_{32}+2)}. \tag{5·12}$$

Using (5·11) we see that $$EU \sim E\check{x}^2 - x^2 \sim \operatorname{var}\check{x}. \tag{5·13}$$

Thus U is an approximately unbiased estimate of the variance of $\check{x}$. By comparison with (2·17) it is evident that the relative errors introduced are of order $m^2 e^{-m}$, where m is n_{21} or n_{32}. For $\check{\lambda}$ and $\check{\mu}$, we have similarly

$$V = \check{\lambda}^2 - \frac{n_{21}(n_{21}-1)(n_3+1)(n_3+2)}{n_2(n_2-1)(n_{31}+1)(n_{31}+2)}, \tag{5·14}$$

where $$EV \sim \operatorname{var}\check{\lambda}, \tag{5·15}$$

and the relative error is of order $n_{31}^2 \exp(-n_{31})$. Also

$$W = \check{\mu}^2 - \frac{s_2^2 n_{31}(n_{31}-1)}{s_1^2(n_{32}+1)(n_{32}+2)}, \tag{5·16}$$

where $$EW \sim \operatorname{var}\check{\mu}, \tag{5·17}$$

and the relative error is of order $n_{32}^2 \exp(-n_{32})$.

6. SUMMARY

The maximum-likelihood estimate of the population size in simple recapture is discussed and shown to have a relative bias of order m^{-1}, where m is the expected number of recaptures (rejecting cases when there are no recaptures). This bias may be serious even if the total sample size is fairly large. A slight adjustment of the estimate gives a quantity whose relative bias is of order e^{-m}, and is therefore satisfactory for much smaller values of m. It is difficult to find an adequate expression for the population value of the variance of this latter statistic. Fortunately, however, these exists an estimate of this variance, with relative bias of order $m^2 e^{-m}$, which is sufficiently good for many practical purposes.

It is also shown that with *inverse* sampling there is an appreciable simplification—an unbiased estimate of population size and an exact value of the sampling variance are available, and are valid for samples of any size in the most general case of sampling without replacement.

Birth- and death-rates have been taken into account by Jackson (1937, 1939) as exemplified in his 'negative' and 'positive' methods. The present paper discusses the maximum-likelihood treatment of both methods, which does not seem to have been given previously, providing in particular properly weighted estimates of the precisions involved.

The data collected by repeated marking and recapture on several occasions can be exhibited in the form of a triangular array or 'trellis' diagram. This was first done by Dowdeswell *et al.* (1939). A general maximum-likelihood treatment of such data is extremely laborious. If we restrict ourselves to collecting data on three occasions only—the 'triple-catch' method —then explicit solutions for the population size and birth- and death-rates are possible. As with simple recapture approximately unbiased estimates of these quantities and approximately unbiased estimates of the corresponding variances are also available. In view of the

simplicity of these results it seems that, in certain circumstances, it might be advantageous to make a special effort to collect data in just this form—concentrating on three occasions only—rather than producing a large trellis diagram which may be difficult to interpret.

I am indebted to Prof. R. A. Fisher for his interest in this paper and for many stimulating discussions.

REFERENCES

ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika*, **35**, 246–54.

BAILEY, NORMAN T. J. (1951). Testing the solubility of maximum likelihood equations in the routine application of scoring methods. *Biometrics*, **7**, (at Press).

DE LURY, D. B. (1947). On the estimation of biological populations. *Biometrics*, **3**, 145–67.

DOWDESWELL, W. H., FISHER, R. A. & FORD, E. B. (1940). The quantitative study of populations in the Lepidoptera. I. *Polyommatus icarus* Rott. *Ann. Eugen., Lond.*, **10**, 123–36.

DOWDESWELL, W. H., FISHER, R. A. & FORD, E. B. (1949). The quantitative study of populations in the Lepidoptera. 2. *Maniola jurtina* L. *Heredity*, **3**, 67–84.

FISHER, R. A. & FORD, E. B. (1947). The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. *Heredity*, **1**, 143–74.

HALDANE, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, **33**, 222–5.

JACKSON, C. H. N. (1933). On the true density of tsetse flies. *J. Anim. Ecol.* **2**, 204–9.

JACKSON, C. H. N. (1937). Some new methods in the study of *Glossina morsitans*. *Proc. Zool. Soc. Lond.* 1936, pp. 811–96.

JACKSON, C. H. N. (1939). The analysis of an animal population. *J. Anim. Ecol.* **8**, 238–46.

JACKSON, C. H. N. (1940). The analysis of a tsetse fly population. *Ann. Eugen., Lond.*, **10**, 332–69.

LINCOLN, F. C. (1930). Calculating waterfowl abundance on the basis of banding returns. *Circ. U.S. Dep. Agric.* no. 118, May 1930.

SCHNABEL, Z. E. (1938). The estimation of the total fish population of a lake. *Amer. Math. Mon.* **45**, 348–50.