A M.S. thesis problem on sample size determination for mark-recapture experiments

D. S. Robson

BU-348-M December, 1970

Abstract

Estimation of population size by mark-recapture methods often entails sequential sampling and marking rather than the simple two-sample design of the classical mark-recapture experiment. In the case of natural populations where sampling involves capturing, the number which can be captured on any single occasion may be very limited, and in order to build up the desired sample size it becomes necessary to sample on a number of successive occasions. If the population remains constant during the sampling period, and if sampling is random, then a sufficient statistic is the number (U) of distinct population elements captured -- i.e., the number of elements in the union of all samples.

The sampling distribution of the sufficient statistic U cannot be expressed in closed form but can be closely approximated by a normal distribution. If preliminary estimates of population size N and the sampling rate $p = n/N$ (for any one occasion) are available then this normal approximation may be used to calculate an estimate of the number (k) of samples need to provide $100(1-\alpha)\%$ assurance that the final population estimate $\hat{N}(U_{(k)})$ will not be in error by more than $100\,\epsilon\,\%$. The problem is to compute and graph the solution (k) as a function of N and p for various combinations of α and ϵ, and to compare these solutions obtained from the normal approximation with exact solutions obtained for a few specific cases. A further useful result would be the preparation of a computer program for (iteratively) calculating the maximum likelihood estimate $\hat{N}(U_{(k)})$ and the confidence limits for N.

Biometrics Unit, Cornell University, Ithaca, New York

A M.S. thesis problem on sample size determination
for mark-recapture experiments

D. S. Robson

BU-348-M December, 1970

The classical mark-recapture method of estimating the number of elements in a finite population consists of drawing a sample of size n_1 from the population, marking these n_1 elements and returning them to the population, then drawing a second sample of size n_2 and counting the number of mark "recaptures" R appearing among these n_2 elements. The unknown population size N is estimated by equating the proportion marked in the second sample to the proportion marked in the population,

$$\frac{R}{n_2} = \frac{n_1}{\hat{N}} .$$

If at least one of the two samples is drawn randomly (without replacement) then for fixed n_1, n_2, N, the probability distribution of R is hypergeometric,

$$P_N\{R=r\} = \frac{\binom{n_1}{r}\binom{N-n_1}{n_2-r}}{\binom{N}{n_1}}$$

In practice this method is often extended from two to k samples, always marking any unmarked elements appearing in a sample and returning the entire sample to the population before drawing the next sample. Letting U_i denote the number of unmarked elements appearing in the i'th sample then for fixed $n_1, n_2, \ldots, n_k, N$, the joint probability distribution of $^{*}U_2, \ldots, U_k$ is easily obtained as

$$P_N(U_2=u_2, \ldots, U_k=u_k) = P_N(U_2=u_2) \cdot P_N(U_3=u_3 \mid U_2=u_2) \cdots P_N(U_k=u_k \mid U_2=u_2, \cdots, U_{k-1}=u_{k-1})$$

* $U_1 \equiv n_1$

$$= \frac{\binom{n_1}{n_2-u_2}\binom{N-n_1}{u_2}}{\binom{N}{n_2}} \cdot \frac{\binom{n_1+u_2}{n_3-u_3}\binom{N-n_1-u_2}{u_3}}{\binom{N}{n_3}} \qquad (1)$$

$$\cdots \frac{\binom{n_1+u_2+\cdots+u_{k-1}}{n_k-u_k}\binom{N-n_1-u_2\cdots u_{k-1}}{u_k}}{\binom{N}{n_k}}$$

when sampling is random. This expression simplifies to

$$P_N(u_2,\ldots,u_k) = \frac{\binom{n_1}{n_2-u_2}}{\binom{N}{n_2}} \frac{\binom{n_1+u_2}{n_3-u_3}}{\binom{N}{n_3}} \cdots \frac{\binom{n_1+u_2+\cdots+u_{k-1}}{n_k-u_k}}{\binom{N}{n_k}} \binom{N-n_1}{u_2,\ldots,u_k}$$

and application of the Meyman factorization theorem shows that

$$U = n_1 + u_2 + \ldots + u_k$$

is a sufficient statistic with respect to $P_N(u_2,\ldots,u_k)$.

The moments of U, obtainable from (1), are more easily derived by expressing U as a sum of N indicator variables

$$\delta_i = \begin{cases} 1 \text{ if the i'th population element appears in at least one of the k samples} \\ 0 \text{ otherwise} \end{cases}$$

$$U = \sum_{i=1}^{N} \delta_i .$$

Thus,

$$E_N(U) = \sum_{i=1}^{N} E_N(\delta_i) = NE_N(\delta_i)$$

$$V_N(U) = NE_N(\delta_i) + N(N-1)E_N(\delta_i\delta_i) - \left[NE_N(\delta_i)\right]^2$$

The marginal distribution of δ_i is seen to be

$$P_N(\delta_i=0) = 1 - P_N(\delta_i=1)$$

$$= \prod_{\nu=1}^{k} \left(1-\frac{n_\nu}{N}\right)$$

so, letting $p_\nu = n_\nu/N$,

$$E_N(\delta_i) = 1 - \prod_{\nu=1}^{k} (1-p_\nu)$$

and

$$E_N(U) = N\left[1- \prod_{\nu=1}^{k} (1-p_\nu)\right] .$$

Similarly,

$$E_N(\delta_i\delta_j) = P_N(\delta_i=1,\ \delta_j=1)$$

$$= P_N(\delta_i=1) - P_N(\delta_i=1,\ \delta_j=0)$$

$$= P_N(\delta_i=1) - P_N(\delta_i=0)P_N(\delta_i=1|\delta_j=0)$$

$$= P_N(\delta_i=1) - P_N(\delta_j=0)P_{N-1}(\delta_i=1)$$

$$= \left[1- \prod_{1}^{k}(1-\frac{n_\nu}{N})\right] - \prod_{1}^{k}(1-\frac{n_\nu}{N}) \left[1- \prod_{1}^{k}(1-\frac{n_\nu}{N-1})\right]$$

and hence, after simplification,

$$V_N(U) = \left[N-E_N(U)\right] \left[E_N(U) - E_{N-1}(U)\right] .$$

Note that if N is large then the difference

$$\Delta E_N(U) = E_N(U) - E_{N-1}(U)$$

is closely approximated by the differential

$$E_N'(U) = 1 = \prod_{1}^{k}(1-\frac{n_\nu}{N}) \left[1 + \sum_{1}^{n_\nu} \frac{n_\nu}{N-n_\nu}\right]$$

so that

$$V_N(U) \doteq \left[N-E_N(U)\right]E'_N(U) \ .$$

As sample sizes and population size get large while $p_\nu = n_\nu/N$ remain fixed, the sampling distribution of

$$Z = \frac{U - E_N(U)}{\sqrt{V_N(U)}}$$

approaches the unit normal distribution, and this approximation may be exploited to construct approximate confidence limits on N once the experiment is completed. The normal approximation also provides a valuable aid in the planning and conduct of the experiment, for if preliminary estimates of N and the p_ν are available then the normal approximation enables us to estimate the number (k) of samples needed to obtain a point estimate $(\hat{N})$ of any specified degree of accuracy. In particular, if the point estimator $\hat{N}(U)$ is obtained as the solution to the moment equation *

$$U = \hat{N}\left[1 - \prod_1^k (1-\frac{n_\nu}{\hat{N}})\right] = E_{\hat{N}}(U)$$

(and this is also the maximum likelihood equation) then the probability that $\hat{N}(U)$ will not be in error by more than 100 ϵ %,

$$P_N\left\{-\epsilon < \frac{\hat{N}(U)-N}{N} < \epsilon\right\}$$

$$= P_N\left\{N(1-\epsilon) < \hat{N}_\nu < N(1+\epsilon)\right\}$$

$$= P_N\left\{E_{N(1-\epsilon)}(U) < U < E_{N(1+\epsilon)}(U)\right\}$$

$$\doteq \Phi\left(\frac{E_{N(1+\epsilon)}(U) - E_N(U)}{\sqrt{V_N(U)}}\right) - \Phi\left(\frac{E_{N(1-\epsilon)}(U) - E_N(U)}{\sqrt{V_N(U)}}\right)$$

can be fixed at any desired level $1-\alpha$ by the appropriate choice of k.

* This equation has exactly one real root which exceeds $\max(n_1,\ldots,n_k)$.

The preliminary values of N, $p_1, p_2, \ldots$ required for the solution of this equation will, in practice, be only educated guesses, correct only in order of magnitude at best. As soon as the first two samples are obtained, however, objective estimates become available for calculating the number of additional samples needed. In general, after r samples are obtained and

$$U_{(r)} = n_1 + u_2 + \ldots + u_r$$

is observed then the number k of additional samples required for $100(1-\alpha)\%$ assurance that $\hat{N}\left(U_{(r+k)}\right)$ will not be in error by more than $100\,\epsilon$ % may be estimated by solving the equation

$$1 - \alpha \doteq \Phi\left(\frac{E_{N(1+\epsilon)}(U_{(r+k)}) - E_N(U_{(r+k)}|U_{(r)})}{\sqrt{V_N(U_{(r+k)}|U_{(r)}}} \right)$$

$$- \Phi\left(\frac{E_{N(1-\epsilon)}(U_{(r+k)}) - E_N(U_{(r+k)}|U_{(r)})}{\sqrt{V_N(U_{(r+k)}|U_{(r)}}} \right)$$

at

$$N = \hat{N}(U_{(r)})$$

and, for example,

$$p_{r+1} = p_{r+2} = \cdots = 1 - \left[\prod_1^r \left(1 - \frac{n_\nu}{\hat{N}(U_r)} \right) \right]^{\frac{1}{r}} = \hat{p}$$

or, alternatively,

$$p_{r+1} = p_{r+2} = \cdots = \frac{1}{r} \sum_1^r \frac{n_\nu}{\hat{N}(U_{(r)})} = \hat{p}$$

Note that the conditional moments of $U_{(r+k)}$ are derivable from the relation

$$U_{(r+k)} = U_{(r)} + \sum_{1}^{N-U_{(r)}} \delta_i$$

where

$$\delta_i = \begin{cases} 1 \text{ if the i'th population element appears in at least one of the k additional samples} \\ 0 \text{ otherwise} \end{cases}$$

Robson and Regier (1964) published graphs of the solution to the sample size problem for the case k = 2, and these appeared to satisfy a strong need, for the graphs have been widely applied in field ecology. It would also appear, then, that a graphical solution to the above problem would likewise be well received by the ecology field.

Reference

Robson and Regier (1964). Sample size in Petersen mark-recapture experiments. Trans. Amer. Fish. Soc. 93, 215-226.