Journal of Official Statistics
Vol. 10. No. 3, 1994, pp. 245–256

Analysis of Sample Based Capture–Recapture Experiments[1]

Juha M. Alho[2]

Abstract: The estimation of population totals in sample based capture–recapture experiments is considered. We permit heterogeneous capture probabilities and use logistic regression to model them. A Horvitz–Thompson type estimator is introduced and an estimator for its variance is given. These formulas appear to be new even when specialized to the homogeneous case. Some aspects of experiments that combine a census with a sample (such as the U.S. Post Enumeration Survey) are studied. In particular, we discuss the role of sampling weights in the estimation of capture probabilities, and we show how purely sample based estimators can be combined with the census estimates to reduce variance. Logistic catch effort models are applied to the optimal allocation of resources to the sampling part and the capture–recapture part of the experiments. We show with an analytical example that there can be a genuine trade-off between the two sources of error.

Key words: Catch effort; Horvitz–Thompson; optimization; post enumeration survey; variance estimation.

1. Introduction

Consider capture–recapture experiments such as the post enumeration surveys (PES) related to the decennial U.S. censuses (cf., Hogan 1992), or estimates of fish stocks or other wildlife in a given geographic area (Seber 1982, pp. 328–340). Narrowly defined, the purpose of such experiments is to estimate the unknown population size. More generally, we may be interested in the sum of the values Z_i of some variable in the population. If $Z_i = 1$ for every individual i, then we obtain an estimate of population size. But Z_i can also be some numerical characteristic, such as income in an economic survey designed to study taxation, or it can be a weight in a study of fish stocks. Both are *population totals* of Z_i.

In many applications it is impracticable to attempt to cover the whole geographic area for which a population total needs to be estimated. Instead, the area is divided into plots ("blocks" in the PES), and a random sample of the plots will be used to estimate the population total (Seber 1982, pp. 19–28). Although the estimation of population totals is a well-studied area in cluster sampling (Cochran 1977, ch. 9A), and the estimation of population size ($Z_i = 1$) is the classical object of capture–recapture experiments, the estimation of population totals in a capture–recapture context appears to have been neglected.

We will first extend the results of Huggins

[1] A version of the paper was presented at the Joint Statistical Meetings in Boston, August 10, 1992.

[2] Department of Statistics, University of Joensuu, P.O. Box 111, 80101 Joensuu, Finland.

(1989) and Alho (1990) that permit the estimation of heterogeneity in capture probabilities using conditional logistic regression. In particular, we will define a Horvitz–Thompson type estimator for a general population total in a sampling context and derive an estimator for its variance.

Second, we will address some aspects of an experiment that combines a census with a sample. Alho, Mulry, Wurdeman, and Kim (1993) proved the basic feasibility of the logistic approach with the 1990 PES. In this paper we will concentrate on two other issues. We will first show that one aspect of the weighting scheme used by the U.S. Bureau of the Census (USBC) leads to an unnecessary inflation of variance. Then, we will compare the efficiency of a purely sample based approach to that of the census-sample approach. This issue is of interest, if (unlike in most national censuses) fine geographic detail is not required. It will also lead to the definition of alternative estimators that have a smaller variance than either the purely sample based estimators or the census based estimators.

Third, we will consider the planning of sample based capture–recapture experiments. These contain two independent sources of error, one due to sampling variation, the other due to the uncaptured individuals in the sample area. This suggests that there may be a trade-off in the design of such experiments between the number of plots sampled and the effort spent in catching individuals within the sample area. We will use logistic catch effort models to highlight the issue.

1.1. *Notation for a capture–recapture experiment*

Consider a non-sample capture–recapture experiment in a closed population of size N. Following the notation of Alho (1990), let us define indicator variables for $i = 1, \ldots, N$ as follows. Let $u_{ji} = 1$, if i is captured the jth time but not the other time, $j = 1, 2$; and $m_i = 1$, if i is captured twice. Then, $n_{ji} = u_{ji} + m_i$ is the indicator of the jth capture with $P(n_{ji} = 1) \equiv p_{ji}$, and $M_i = u_{1i} + u_{2i} + m_i$ is the overall capture indicator with $P(M_i = 1) \equiv \phi_i$. The captures are assumed to be independent, so $\phi_i = p_{1i} + p_{2i} - p_{1i}p_{2i}$. Furthermore, we assume that the capture probabilities follow logistic models, $\text{logit}(p_{ji}) = \mathbf{X}_{ji}^{\mathrm{T}}\mathbf{a}_j$, where $\mathbf{X}_{ji}$ are vectors of covariates relating to $i = 1, \ldots, N$, and $\mathbf{a}_j$ are vectors of parameters. Conditional maximum likelihood estimators of the $\mathbf{a}_j$ can be obtained based on observed individuals (i.e., those with $M_i = 1$) only, denote the estimators by $\hat{\mathbf{a}}_j$. The corresponding maximum likelihood estimators of p_{ji} and ϕ_i are $\hat{p}_{ji}$ and $\hat{\phi}_i$. The resulting estimator of population size is $\tilde{N} = M_1/\hat{\phi}_1 + \ldots + M_N/\hat{\phi}_N$. Define also $n_j = n_{j1} + \ldots + n_{jN}$, $j = 1, 2$, and similarly the other sums over individuals i as u_j, m, and M. Under the homogeneity of capture probabilities the estimator $\tilde{N}$ reduces to the usual capture–recapture (or dual system) estimator $\hat{N} = n_1 n_2/m$ (Alho 1990, p. 628).

We may generalize the definition of $\tilde{N}$ to give a Horvitz–Thompson estimator of the population total $Z = Z_1 + \ldots + Z_N$ as $\tilde{Z} = Z_1 M_1/\hat{\phi}_1 + \ldots + Z_N M_N/\hat{\phi}_N$. When $Z_i \equiv 1$, a sufficient condition for the consistency and asymptotic normality of the estimators is (essentially) the boundedness of the $\mathbf{X}_{ji}$ (Alho 1990, p. 628). A perusal of the proof shows that sufficient condition for the asymptotic results for a general $\tilde{Z}$ is, roughly speaking, that the Z_i are also bounded.

2. A Sample Based Estimator of Z

Let N be the (unknown) size of a closed population in a geographic area. Suppose

the area is partitioned into plots. The object is to estimate the population total Z based on a random sample of the plots.

We assume that based on sample design we know for every plot the probability of its being included in the sample. This will also be the probability that the cluster of individuals residing in the plot will be included in the capture–recapture experiment. In other words, we know the probabilities $f_i = P(i$ is included in the sample area), for every individual $i = 1, \ldots, N$ we happen to capture. A Horvitz–Thompson estimator of the population total is then

$$\tilde{Z} = \sum_{i=1}^{N} Z_i M_i / (f_i \hat{\phi}_i) \tag{2.1}$$

This estimator generalizes the estimator for population size in Alho et al. (1993, eq. 4, p. 1132). Note that the definition (2.1) agrees with the one given in Section 1.1, if we take $f_i = 1$. This is the way other formulas in this section simplify to the non-sampling case.

Suppose the sampling design is such that there are S different possible samples. Let s be a random variable indicating which sample was chosen, so s takes values in $I = \{1, \ldots, S\}$. Let $I_s \subset \{1, \ldots, N\}$ be the set of individuals belonging to the sample area. Define $N_s =$ number of elements in I_s. With these notations we can make the dependency of $\tilde{Z}$ on the particular sample more obvious by writing

$$\tilde{Z} = \sum_{i \in I_s} Z_i M_i / (f_i \hat{\phi}_i). \tag{2.2}$$

Note also that conditioning on s, we have asymptotically (keep the sample plots fixed and let the population of the sample area increase) that $E[\tilde{Z} \mid s] = F_s$, where

$$F_s = \sum_{i \in I_s} Z_i / f_i. \tag{2.3}$$

Example 2.1. Suppose $Z = Z(1) + \ldots + Z(K)$, where $Z(k) =$ population total of region $k = 1, \ldots, K$. Suppose there are $Q(k)$ plots in region k, $q(k)$ of which are selected into the sample with equal probability. Let $Z_s(k)$ be the population total in the sample area in region k. In this case $f_i = q(k)/Q(k)$ for all individuals in region k, and

$$F_s = \sum_{k=1}^{K} Z_s(k) Q(k) / q(k). \tag{2.4}$$

Each of the $q(k)$ sample plots from region k has the expected total $Z(k)/Q(k)$. Asymptotic unbiasedness follows.

For future reference, let us calculate the asymptotic variance of $E[\tilde{Z} \mid s]$ using (2.4). We have

$$\operatorname{Var}(E[\tilde{Z} \mid s]) = \sum_{k=1}^{K} \operatorname{Var}(Z_s(k)) \times Q(k)^2 / q(k)^2. \tag{2.5}$$

Assume now that $Z(k) = Z(k,1) + \ldots + Z(k, Q(k))$ and denote the variance of the numbers $Z(k,1), \ldots, Z(k, Q(k))$ by S_k^2 (with $Q(k) - 1$ rather than $Q(k)$ in the denominator; cf. Cochran 1977, p. 23); then, (2.5) becomes

$$\operatorname{Var}(E[\tilde{Z} \mid s]) = \sum_{k=1}^{K} S_k^2 [Q(k) - q(k)] \times Q(k)/q(k). \tag{2.6}$$

We see that (2.6) is minimized if the plots can be defined so that the population totals $Z(k,j)$, $j = 1, \ldots, Q(k)$, are equal within each region, a familiar result from stratified sampling.

In practice we can estimate S_k^2 by the sample variance of the capture–recapture estimates $\tilde{Z}(k,j)$ from the $q(k)$ sample plots. This gives an overestimate because the $Z(k,j)$ are only known up to the error $\epsilon(k,j) \equiv \tilde{Z}(k,j) - Z(k,j)$. A correction is easily obtained if we subtract from the sample variance of the $\tilde{Z}(k,j)$, the

average of the capture–recapture variances $\mathrm{Var}(\epsilon(k,j)\,|\,s) = \mathrm{Var}(\tilde{Z}(k,j)\,|\,s)$.■

In the general case all plots may have different sampling probabilities. This is the case if the probabilities are proportional to the areas of the plots, for example. Then, an estimate of $\mathrm{Var}(E[\tilde{Z}\,|\,s])$ can be obtained from the Corollary of Theorem 9A.5 of Cochran (1977, p. 261, with the identification $\pi_j = f_j$ and $Y_j =$ capture–recapture estimate of the total of plot j; π_{ij} depends on the sampling method used). A correction to the variance along the lines of Example 2.1 is also feasible.

3. Estimating the Variance of $\tilde{Z}$

Strictly speaking the variance of $\tilde{Z}$ does not exist. The easiest way to see this is to consider the case of population estimation under homogeneous capture probabilities, when $\tilde{Z} = \tilde{N} = n_1 n_2/m$. Since there is a positive probability that $m = 0$, no moments exist. Therefore, all the variances and expectations we consider refer to those of the asymptotic normal distribution of $\tilde{Z}$. We will calculate $\mathrm{Var}(\tilde{Z})$ by using the well known decomposition

$$\mathrm{Var}(\tilde{Z}) = E[\mathrm{Var}(\tilde{Z}\,|\,s)] + \mathrm{Var}(E[\tilde{Z}\,|\,s]). \tag{3.1}$$

The second term in the decomposition (3.1) can be written as $\mathrm{Var}(F_s)$. As shown in Example 2.1, this can be readily estimated from sample data.

Consider the first term of the decomposition. In practice, we would use the sample based estimator of $\mathrm{Var}(\tilde{Z}\,|\,s)$ to estimate $E[\mathrm{Var}(\tilde{Z}\,|\,s)]$. The estimator can be derived following the steps of the argument for $Z_i = 1$ and $f_i = 1$ in Alho (1990, pp. 629–630; we omit the details here). First we get a representation for the theoretical variance of the form $\mathrm{Var}(\tilde{Z}\,|\,s) = \nu_2(s) + \nu_3(s)$, where $\nu_2(s) = E[\mathrm{Var}(\tilde{Z}\,|\,\mathbf{M}, s)]$ and $\nu_3(s) = \mathrm{Var}(E[\tilde{Z}\,|\,\mathbf{M}, s])$, or the conditioning is with respect to the M_i values represented by $\mathbf{M}$, in addition to the sample s. This can be estimated by $V_0(s) = V_2(s) + V_3(s)$, where

$$V_2(s) = \tilde{\psi}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\tilde{\psi} \tag{3.2}$$

and

$$V_3(s) = \sum_{i \in I_s} Z_i^2 M_i(1 - \hat{\phi}_i)/(f_i\hat{\phi}_i)^2. \tag{3.3}$$

In (3,2), we have that $\tilde{\psi} = (\tilde{\psi}_1, \ldots, \tilde{\psi}_{2N})^{\mathrm{T}}$ with

$$\tilde{\psi}_i = (Z_i M_i/f_i)\exp(\mathbf{X}_{1i}^{\mathrm{T}}\hat{\mathbf{a}}_1) \times (1 + \exp(\mathbf{X}_{2i}^{\mathrm{T}}\hat{\mathbf{a}}_2))/\hat{K}_i^2$$

$$\tilde{\psi}_{N+i} = (Z_i M_i/f_i)\exp(\mathbf{X}_{2i}^{\mathrm{T}}\hat{\mathbf{a}}_2) \times (1 + \exp(\mathbf{X}_{1i}^{\mathrm{T}}\hat{\mathbf{a}}_1))/\hat{K}_i^2$$

where

$$\hat{K}_i = \exp(\mathbf{X}_{1i}^{\mathrm{T}}\hat{\mathbf{a}}_1) + \exp(\mathbf{X}_{2i}^{\mathrm{T}}\hat{\mathbf{a}}_2) + \exp(\mathbf{X}_{1i}^{\mathrm{T}}\hat{\mathbf{a}}_1 + \mathbf{X}_{2i}^{\mathrm{T}}\hat{\mathbf{a}}_2)$$

for $i = 1, \ldots, N$. Furthermore,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{bmatrix}, \quad \tilde{\mathbf{W}} = \begin{bmatrix} \tilde{\mathbf{W}}_1 & \tilde{\mathbf{W}}_3 \\ \tilde{\mathbf{W}}_3 & \tilde{\mathbf{W}}_2 \end{bmatrix}$$

where $\mathbf{X}_j = [X_{j1}, \ldots, X_{jN}]^{\mathrm{T}}$, and $\tilde{\mathbf{W}}_j = (\tilde{w}_{ik}^j)$ are diagonal $N \times N$ matrices, with

$$\tilde{w}_{ii}^j = M_i(\hat{p}_{ji}/\hat{\phi}_i - \hat{p}_{ji}^2/\hat{\phi}_i^2), \quad j = 1, 2$$

$$\tilde{w}_{ii}^3 = M_i(\hat{p}_{1i}\hat{p}_{2i}/\hat{\phi}_i - \hat{p}_{1i}\hat{p}_{2i}/\hat{\phi}_i^2).$$

Note that the M_i omit those individuals from the above expressions who were not in the sample area or were not observed in the capture–recapture experiment. When $Z_i = f_i = 1$ for $i = 1, \ldots, N$, then the formulas reduce to those in Alho (1990). In particular, in the case of homogeneous capture probabilities they reduce to the classical variance estimator (3.1.2) given below.

The formulas given above define an estimator of $\mathrm{Var}(\tilde{Z}\,|\,s)$. It was obtained by estimating certain of the ϕ_i by M_i for $i \in I_s$

and the parameters a_j by their maximum likelihood estimators. Conversely, when we need an expression for the theoretical variance itself, $\mathrm{Var}(\tilde{Z}|s)$, it is obtained by replacing certain M_i by ϕ_i and the parameter estimators by the true parameter values. Details of this and an expression for $E[\mathrm{Var}(\tilde{Z}|s)]$ are given in the Appendix.

3.1. *Variance under homogeneity*

Assume that $p_{ji} = p_j$ for $i = 1, \ldots, N$ and $j = 1, 2$. It was shown in Alho (1990, p. 628) that the conditional maximum likelihood estimators of capture probabilities agree with the classical estimators $\hat{p}_1 = m/n_2$ and $\hat{p}_2 = m/n_1$. In the sampling case $\hat{N} = n_1 n_2/m$ is a consistent estimator of the population size N_s of the sample area. Its variance is $N_s\Phi$, where

$$\Phi = (1 - p_1)(1 - p_2)/(p_1 p_2). \tag{3.1.1}$$

This follows from the formulas described in the Appendix or it can be shown directly (Alho 1991, p. 125). The well-known moment estimator of the variance is

$$V_1(s) = n_1 n_2 u_1 u_2 / m^3. \tag{3.1.2}$$

Recall that in the homogeneous case $\phi_i = p_1 + p_2 - p_1 p_2$. Substituting in the estimators $\hat{p}_j$ we get that $\hat{\phi}_i = mM/(n_1 n_2)$. Therefore, (2.1) can be written as

$$\tilde{Z} = \hat{N}\tilde{M}/M \tag{3.1.3}$$

where

$$\tilde{M} = \sum_{i \in I_s} Z_i M_i / f_i. \tag{3.1.4}$$

Based on the Appendix it is easy to see that $\mathrm{Var}(\tilde{Z}|s) = \nu_2(s) + \nu_3(s)$ has

$$\nu_3(s) = G_s \Phi p_1 p_2 / \phi \tag{3.1.5}$$

where

$$G_s = \sum_{i \in I_s} (Z_i/f_i)^2. \tag{3.1.6}$$

A straightforward, but tedious calculation shows that

$$\nu_2(s) = (F_s^2/N_s)\Phi(\phi - p_1 p_2)/\phi. \tag{3.1.7}$$

Note that when $Z_i = f_i = 1$, then $\mathrm{Var}(\tilde{Z}|s) = N_s\Phi$, as it should.

A formula for $E[\mathrm{Var}(\tilde{Z}|s)]$ is obtained by taking the expectation

$$E[\mathrm{Var}(\tilde{Z}|s)] = E[G_s]\Phi p_1 p_2/\phi + E[F_s^2/N_s] \times \Phi(\phi - p_1 p_2)/\phi. \tag{3.1.8}$$

Example 3.1.1. In the simple random sampling of q plots from Q equally likely plots $f_i = q/Q$. Define $\Gamma_j = Z_1^j + \ldots + Z_N^j$, $j = 1, 2$. Then, we see that $E[G_s] = (Q/q)\Gamma_2$. Unfortunately, the term $E[F_s^2/N_s]$ in (3.1.8) does not simplify without further assumptions. To get an idea of what it might look like let us resort to a randomization argument. Suppose we take first a sample of the plots, and thereby of the individuals, with N_s individuals selected. Suppose we then assign the Z_i values by taking a random permutation of some known vector $(Z_1, \ldots, Z_N)$. In this case a sample of plots with N_s individuals is expected to have a population total $N_s\Gamma_1/N$. Furthermore, for any $i \in I_s$, $E[Z_i^2] = \Gamma_2/N$; and for any $i, j \in I_s$, $i \neq j$, $E[Z_i Z_j] = (\Gamma_1^2 - \Gamma_2)/[N(N-1)]$. Therefore,

$$E[F_s^2/N_s] = (Q/q)^2\{\Gamma_2/N + (\Gamma_1^2 - \Gamma_2) \times (qN/Q - 1)/[N(N-1)]\}.$$

It follows that $E[\mathrm{Var}(\tilde{Z}|s)] = T_1$, where

$$T_1 = (Q/q)\Gamma_2\Phi p_1 p_2/\phi + [\Phi(\phi - p_1 p_2)/\phi] \times (Q/q)^2\{\Gamma_2/N + (\Gamma_1^2 - \Gamma_2) \times (qN/Q - 1)/[N(N-1)]\}. \tag{3.1.9}$$

Take $K = 1$ in Example 2.1 and write $S_1^2 = S^2$ for short. From (2.6) we get that $\mathrm{Var}(E[\tilde{Z}|s]) = T_2$, where

$$T_2 = S^2[Q - q]Q/q. \tag{3.1.10}$$

In simple random sampling under homogeneous capture probabilities, $\text{Var}(\tilde{Z}) = T_1 + T_2$. Note that this reduces to $N\Phi$, if $q = Q$ and $Z_i = 1$. We will illustrate the use of (3.1.9) and (3.1.10) in Section 6.

An extension to stratified sampling is immediate. All parameters above can be taken to be region specific: $q = q(k)$, $Q = Q(k)$, $\Phi = \Phi(k)$, $p_j = p_j(k)$, $\Gamma_j = \Gamma_j(k)$, $S^2 = S^2(k)$, and $T_j = T_j(k)$, $k = 1, \ldots, K$. Then, $\text{Var}(\tilde{Z})$ is obtained by summing $T_1(k) + T_2(k)$ over the regions k.■

4. Estimation in a Census-Sample Experiment

In the decennial U.S. censuses a complete count is attempted at census time. In the PES, capture–recapture (or dual system) techniques are used on a sample basis to assess the accuracy of the census. The census count is subject to various types of data errors (duplications, fictitious enumerations, clerical errors, etc.) and it contains imputations for missing data (cf. Hogan 1992). A major part of the PES consists of estimating the net effect of such factors. The resulting undercount estimates are smoothed using regression techniques (Hogan 1992, p. 267).

In Section 4.1. we will address one aspect of this complex procedure, and show that the use of sampling weights in the estimation of capture probabilities unnecessarily inflates variance. Section 4.2. will show that purely sample based estimation of population size can, in some circumstances, be more accurate than a census-sample approach. This leads to the definition of more efficient estimators.

4.1. *Variance of the weighted estimator*

The analysis of the PES data is currently carried out by subgroups of population (defined by age, sex, race, region, etc.) to permit the assumption that capture probabilities are homogeneous (cf. Sekar and Deming 1949). We will confine attention to one such homogeneous subpopulation with the size N. The maximum likelihood estimator of the population N_s of the sample area is $\hat{N} = n_1 n_2 / m$. The problem of estimating the population outside the sample area, $N - N_s$, remains. Denote the census count of that area by Π. One alternative is $\tilde{N}_1 = \Pi / \hat{p}_1$, where $\hat{p}_1 = m/n_2$ is the maximum likelihood estimator of p_1. A close analogue of the estimator the USBC uses[2], is $\tilde{N}_2 = \Pi \tilde{n}_2 / \tilde{m}$, where

$$\tilde{n}_2 = \sum_{i \in I_s} n_{2i}/f_i \tag{4.1.1}$$

and

$$\tilde{m} = \sum_{i \in I_s} m_i/f_i. \tag{4.1.2}$$

The rationale of this estimator is that it attempts to recreate the "target dual system estimator" (cf. Mulry and Spencer 1991) that would be obtained if the whole country were covered by the second count. We will prove now that *the variance of $\tilde{N}_1$ is always smaller than that of $\tilde{N}_2$.*

It is easy to see that both $\tilde{N}_1$ and $\tilde{N}_2$ are asymptotically unbiased. Note that Π is statistically independent of the n_{1i} and n_{2i}. Therefore, conditioning on Π we have

[2] Based on Hogan (1992, p. 267) we can infer that in the 1990 PES the estimator implied for the non-imputed population outside the sample area can be written as $\tilde{N} = c \times \Pi \tilde{n}_2 / \tilde{m}$, where Π = census count of non-sample area excluding whole person imputations; $\tilde{n}_2$ = weighted P-sample total; $\tilde{m}$ = weighted estimate of P-sample cases that could be matched with the census; and c = estimated fraction of correct non-imputed census enumerations, which is based on the E-sample. (Here, P-sample is an independent population sample designed to measure the rate of omission in the census. E-sample is a sample of census enumerations designed to estimate the rate of erroneous inclusions.) The weights reflected sampling probabilities, corrections for whole household non-interviews, and estimated probabilities of resolvedness. In our analysis we will simplify by taking $c = 1$, and by assuming that the weights represent sampling probabilities only.

asymptotically that

$$\mathrm{Var}(\tilde{N}_1) = E[\Pi^2]\,\mathrm{Var}(n_2/m) + \mathrm{Var}(\Pi/p_1)$$

$$\mathrm{Var}(\tilde{N}_2) = E[\Pi^2]\,\mathrm{Var}(\tilde{n}_2/\tilde{m}) + \mathrm{Var}(\Pi/p_1).$$

Conditioning on s, we have asymptotically $E[n_2/m\,|\,s] = 1/p_1 = E[\tilde{n}_2/\tilde{m}\,|\,s]$ for all s. Therefore,

$$\mathrm{Var}(n_2/m) = E[\mathrm{Var}(n_2/m\,|\,s)]$$

$$\mathrm{Var}(\tilde{n}_2/\tilde{m}) = E[\mathrm{Var}(\tilde{n}_2/\tilde{m}\,|\,s)].$$

To compare the variances, we will use a linear Taylor series expansion (or the *delta method*; cf. Rao 1973, pp. 388–389) for $\tilde{n}_2/\tilde{m}$ at $E[\tilde{n}_2\,|\,s] = F_s p_2$ and $E[\tilde{m}\,|\,s] = F_s p_1 p_2$. We get the asymptotic expansion

$$\begin{aligned}\mathrm{Var}(\tilde{n}_2/\tilde{m}|s) &= [1/(F_s p_1 p_2)]^2\,\mathrm{Var}(\tilde{n}_2\,|\,s) \\ &\quad + [F_s p_2/(F_s p_1 p_2)^2]^2 \mathrm{Var}(\tilde{m}\,|\,s) \\ &\quad - 2[F_s p_2/(F_s p_1 p_2)^3]\,\mathrm{Cov}(\tilde{n}_2, \tilde{m}\,|\,s).\end{aligned}$$

Take $Z_i = 1$ in (3.1.6). We have that $\mathrm{Var}(\tilde{n}_2\,|\,s) = G_s p_2(1-p_2)$, $\mathrm{Var}(\tilde{m}\,|\,s) = G_s p_1 p_2(1 - p_1 p_2)$ and

$$\begin{aligned}\mathrm{Cov}(\tilde{n}_2, \tilde{m}\,|\,s) &= \sum_{i \in I_s} \mathrm{Cov}(n_{2i}, m_i\,|\,s)/f_i^2 \\ &\quad + \sum_{i \neq j} \mathrm{Cov}(n_{2i}, m_j\,|\,s)/(f_i f_j).\end{aligned}$$

When $i \neq j$, then $\mathrm{Cov}(n_{2i}, m_j\,|\,s) = 0$ by the independency of the individuals. For $i = j$ we have that

$$\mathrm{Cov}(n_{2i}, m_i\,|\,s) = p_1 p_2(1 - p_2).$$

It follows that asymptotically

$$\mathrm{Var}(\tilde{n}_2/\tilde{m}\,|\,s) = (G_s/F_s^2)H \qquad (4.1.3)$$

where

$$H = [(1 - p_1 p_2)p_1 - p_2(1 - p_2)]/(p_1 p_2)^2.$$

A corresponding variance calculation for n_2/m is obtained by taking $f_i = 1$ for all i. Or, we have asymptotically

$$\mathrm{Var}(n_2/m) = (1/N_s)H. \qquad (4.1.4)$$

It is a direct consequence of the Cauchy–Schwartz inequality that $G_s N_s > F_s^2$. Therefore, (4.1.3) is always larger than (4.1.4), with equality only in the case of simple random sampling when the f_i are constant over i. This completes the proof.

We have shown that the asymptotic variance of $\tilde{N}_2$ is always larger than or equal to the variance of $\tilde{N}_1$. It is possible that for reasons of data correction there are motives for using $\tilde{N}_2$ instead of $\tilde{N}_1$. However, one should realize that this is done at the cost of increased variance[3]. Our result supports the idea (cf. Alho et al. 1993, p. 1133) that when there are data errors, one can estimate the parameters of the capture–recapture model (here p_1) using somewhat different data than the data used to estimate population (here $N - N_s$). What is an appropriate correction for one purpose may not be optimal for the other.

4.2. *Efficiency of purely sample based estimation*

Since a census is often a very costly operation one might inquire how well one can estimate population size if one omitted the census altogether and based estimates on a purely sample based capture–recapture experiment. We note that somewhat surprisingly, in certain circumstances, the purely sample estimator can be more accurate than the census-sample based estimator.

Consider the case in which a simple random sample of size q is taken from among Q plots, so $f_i = q/Q$. Let us investigate the conditions under which

[3] Having assumed that $c = 1$ does not change the conclusion as far as the PES is concerned. Having probabilities of resolvedness available does not change it either, but it leads to several alternative procedures, some of which are reviewed in Alho et al. (1993, p. 1133).

$\tilde{N}_1 = \Pi n_2/m$ is a more efficient estimator of the population $N - N_s$ (i.e., the population outside the sample area) than the purely sample based estimator $\tilde{N} - \hat{N} = [(Q-q)/q]n_1n_2/m$. Both estimators are of the form $f(X,Y,Z)$, where $f(x,y,z) = xy/z$. In both cases we may take $Y = n_2$ and $Z = m$. In the case of $\tilde{N}_1$ take $X = \Pi$. In the case of $\tilde{N} - \hat{N}$ take $X = [(Q-q)/q]n_1$. Using the delta method one can show (lengthy details omitted) that $\mathrm{Var}(\tilde{N} - \hat{N}) \geq \mathrm{Var}(\tilde{N}_1)$, if and only if

$$Q(Q-2q)/[q(Q-q)] \geqslant \mathrm{Var}(\Pi + n_1)/\mathrm{Var}(N_s p_1). \quad (4.2.1)$$

If the left hand side of (4.2.1) is less than the right hand side, then the purely sample based estimator is more efficient. This can happen if p_1 is small and the populations of the plots can be made very nearly equal. Such circumstances might occur in a census of wildlife. Considerations of cost might make a purely sample based approach preferable to a census for a human population as well, if great geographic detail is not required. Conceivably some of the geographic detail of a census could be preserved if administrative records, or other lists, could be used outside the sample area to provide a frame for *pro rata* estimates.

The result (4.2.1) is also of theoretical interest. When (4.2.1) does not hold, then $\mathrm{Var}(\tilde{N} - \hat{N}) < \mathrm{Var}(\tilde{N}_1)$. But having the census count available we know n_1, and hence $\tilde{N} - \hat{N}$. Therefore, it would seem, paradoxically, that it is more efficient to ignore the data Π than to use it! The resolution is to use a weighted average. Define $\gamma_1 = \mathrm{Var}(\tilde{N}_1)$, $\gamma_2 = \mathrm{Var}(\tilde{N} - \hat{N})$, and $\gamma_{12} = \mathrm{Cov}(\tilde{N}_1, \tilde{N} - \hat{N})$. Then,

$$\tilde{N}_3 = [(\gamma_2 - \gamma_{12})\tilde{N}_1 + (\gamma_1 - \gamma_{12})(\tilde{N} - \hat{N})] / (\gamma_1 + \gamma_2 - 2\gamma_{12}) \quad (4.2.2)$$

has a smaller variance than $\tilde{N} - \hat{N}$ or $\tilde{N}_1$, irrespective of (4.2.1). The weights depend on p_1, p_2, and population size (in addition to Q and q), but an iterative estimator based on (4.2.2) appears always feasible.

The result extends directly to the general estimation of population totals

$$\tilde{Z}_1 = Z_1 n_{11}/\hat{p}_{11} + \ldots + Z_N n_{1N}/\hat{p}_{1N} \quad (4.2.3)$$

which generalizes the census based population estimator proposed in Alho et al. (1993, eq. 5, p. 1132). A weighted average of (4.2.3) and (2.2) in analogy with (4.2.2) can be superior to both (4.2.3) and (2.2).

5. Logistic Catch Effort Models

Both in a purely sample based experiment and in a census-sample experiment there potentially is a trade-off between the intensity of the catch effort within the sample area and sample size. Suppose that a given baseline expenditure $C^* = C_1^* + C_2^*$, with $C_j^* > 0$ going to the jth capture, $j = 1,2$, has yielded in the past capture probabilities $\mathrm{logit}(p_{ji}) = \mathbf{X}_{ji}^{\mathrm{T}} a_j$, $j = 1,2$. Postulating the availability of such information, we will assume that in general the expenditure $C = C_1 + C_2$ with C_j going to the jth capture will yield capture probabilities of the form

$$\mathrm{logit}(p_{ji}(C_1, C_2)) = \mathbf{X}_{ji}^{\mathrm{T}} a_j + g_j(C_j), \quad j = 1,2, \quad (5.1)$$

where the g_j are known, differentiable, strictly increasing functions such that $g_j(C_j^*) = 0$. In other words, the expenditure C_j versus the expenditure C_j^* will yield *an odds ratio of capture* equal to $\exp(g_j(C_j))$, no matter what the individual covariates X_{ji} are. For this reason we will call (5.1) a *logistic catch effort model.*

Example 5.1. Perhaps the simplest model for g_j assumes that $g_j = G_{1j}$, where

$$G_{1j}(C_j) = \kappa_j(C_j - C_j^*), \quad j = 1,2, \quad (5.2)$$

and $\kappa_j > 0$ are *catch effort parameters*. A model of this type implies a positive probability of capture at $C_j = 0$, so it cannot hold for all values of C_j. It might hold approximately when C_j is close to C_j^*, however. Another defect in the model is that it implies increasing returns in the probability of capture per unit investment, when $p_{ji}(C_1, C_2) < 1/2$, and decreasing returns only when $p_{ji}(C_1, C_2) > 1/2$. It follows that the model (5.2) might be applicable only in cases in which capture probabilities are large, such as the PES.■

Example 5.2. Assume that $g_j(C_j) = G_{2j}(C_j)$, where

$$G_{2j}(C_j) = \kappa_j \log(C_j/C_j^*), \quad j = 1, 2, \quad (5.3)$$

where $\kappa_j > 0$ are catch effort parameters. Note that in this case $p_{ji}(0, 0) = 0$ and $p_{ji}(C_1, C_2) \to 1$, as $C_j \to \infty$. From the second derivative of p_{ji} we see that (5.3) is guaranteed to define p_{ji} as a concave function of C_j, if $\kappa_j < 1$, no matter what C_j^* and $\mathbf{X}_{ji}^{\mathrm{T}}\mathbf{a}_j$ are. Therefore, if it is realistic to assume that $0 < \kappa_j < 1$, (5.3) might be taken to hold globally, for all $C_j > 0$, in some applications. We will illustrate its use in Section 6.■

In practice, (5.2) and (5.3) require that the κ_j are known from past data. Rough estimates may be obtained even from a single PES type experiment, if one can determine how much more intensive the sample study was as compared to the standard census (thus yielding two probabilities of capture), and how much extra one had to pay for the improvement (C^* = census cost per plot; C = sample cost per plot). The accounting of shared costs (permanent personnel, basic training, computers, maps, etc.) is a major challenge in such a calculation. A more refined analysis requires a better replication of experiments at different levels of expenditure. Then, standard techniques of variable selection and goodness-of-fit testing (analogous to those proposed for standard logistic regression in Hosmer and Lemeshow 1989, chs. 4 and 5) can be used in modeling.

Above, the logistic catch effort models were introduced as a mathematically convenient representation. In some applications the following argument may be used to lend further credibility to the models.

Define $p(C|X)$ as the probability that a person with covariates X is captured when the level of expenditure is C. Suppose that an important factor in capturing individuals is the flow of information from those captured to those not yet captured. Then the effect of increased expenditure from C to $C + h$ could be to intensify the flow of information. In such circumstances we might assume that, as in the theory of epidemics

$$p(C + h|X) - p(C|X) = \nu(C)p(C|X)(1 - p(C|X))h + o(h) \quad (5.4)$$

where $\nu(C)$ is some continuous, nonnegative function, and $o(h)/h \to 0$, as $h \to 0$. Dividing (5.4) by h and letting $h \to 0$, we get a differential equation

$$p'(C|X) = \nu(C)p(C|X)(1 - p(C|X)).$$

Suppose we have the initial condition

$$p(C^*|X) = \exp(A(X))/(1 + \exp(A(X))),$$

where A is a known function. Then, the well-known solution (e.g., Boyce and DiPrima, 1970, pp. 48–49) is

$$p(C|X) = \exp(g(C) + A(X))/(1 + \exp(g(C) + A(X)))$$

where

$$g(C) = \int_{C^*}^{C} \nu(y)\, dy.$$

By taking $A(X) = \mathbf{X}_{ji}^{\mathrm{T}}\mathbf{a}_j$ and $g = g_j$, we get (5.1). For example, if $\nu(y) = \kappa_j$, then $g = G_{1j}$, and if $\nu(y) = \kappa_j/y$, then $g = G_{2j}$ defined by (5.2) and (5.3), respectively.

6. Optimization

The practical planning of sample based capture–recapture experiments involves both sampling design and the design of the capture experiment. Based on the results of Sections 2 and 3 (Examples 2.1 and 3.1.1, for example) we get a formula for the variance of the estimator of the population total. A priori values for the parameters permit the numerical evaluation of loss functions for different designs provided that the losses are functions of the variance.

Analytical solutions appear to be feasible only in the simplest case of population size estimation under homogeneous capture probabilities and simple random sampling. We will go through one analysis in this simple setting and show that there can be a genuine trade-off in the allocation of funds to the sampling part and capture–recapture part of the experiment.

Let us assume that $Z_i = 1$, $p_{ji} = p_j$, and that a sample of q is selected from among Q equally likely plots. In this case $\mathrm{Var}(E[\tilde{N}\,|\,s]) = T_2$ as given by (3.1.10). Correspondingly, $E[\mathrm{Var}(\tilde{N}\,|\,s)] = T_1$ as defined by (3.1.9), where $Z_i = 1$ implies that $\Gamma_1 = \Gamma_2 = N$. It follows that

$$\mathrm{Var}(\tilde{N}) = (Q/q)\{N\Phi + S^2(Q-q)\}. \tag{6.1}$$

Suppose now that an amount $C = C_1 + C_2$ is spent for catch effort per plot, with C_j going to the jth capture. Assume that the capture probabilities depend on expenditures through (5.3). How should C be allocated to the captures? We see from (6.1) that the allocation has an effect on the variance through Φ only. We have

$$\Phi = (C_1^*)^{\kappa_1}(C_2^*)^{\kappa_2}C_1^{-\kappa_1}C_2^{-\kappa_2}. \tag{6.2}$$

Substituting $C_2 = C - C_1$, and by differentiating we get that the minimum of Φ occurs at

$$C_j = \kappa_j C/\kappa, \quad j = 1,2, \tag{6.3}$$

where $\kappa = \kappa_1 + \kappa_2$. Or, the optimal allocations are proportional to the catch effort parameters κ_j. The minimum value of Φ is $\xi C^{-\kappa}$, where

$$\xi = (C_1^*)^{\kappa_1}(C_2^*)^{\kappa_2}\kappa_1^{-\kappa_1}\kappa_2^{-\kappa_2}\kappa^{\kappa}.$$

Assuming that a given amount C is allocated optimally between the captures we get that

$$\mathrm{Var}(\tilde{N}) = (Q/q)\{N\xi C^{-\kappa} + S^2(Q-q)\}. \tag{6.3}$$

Suppose we have a total expenditure $D > 0$ that can be used for the sample based capture–recapture experiment. With C going to each plot and q plots sampled, we have $D = qC$. Substituting $C = D/q$ into (6.3) we can express the variance in terms of sample size

$$\mathrm{Var}(\tilde{N}) = Q\{N\xi D^{-\kappa}q^{\kappa-1} + S^2[(Q/q) - 1]\}. \tag{6.4}$$

By considering the derivative of (6.4) with respect to q we note that the derivative is always negative, if $\kappa \leq 1$. For those values it is optimal to sample all plots, or take $q = Q$. If $\kappa > 1$, then the minimum occurs at

$$q = D\left\{\frac{S^2Q}{N\xi(\kappa-1)}\right\}^{1/\kappa}. \tag{6.5}$$

The optimal sample size is an increasing function of the variance S^2 and a decreasing function of N/Q. In this case there is a genuine trade-off between sample size and catch effort.

An interesting theoretical problem connected with (6.5) is caused by the fact that q is a function of uncertain parameters κ_j. The uncertainty may derive either from sampling variability, or it may be subjective, if the parameters represent expert judgement. It is important to analyze how sensitive q is to the uncertainty.

The practical application of the optimization results may require additional considerations. In the PES, for example, the so-called E-sample of census enumerations is designed to estimate the rate of erroneous enumerations. Only the independent population sample, or the P-sample, properly corresponds to a capture experiment. Hence, both the loss functions and the accounting of cost are more complex than the ones considered here.

7. Further Aspects

We have discussed some aspects of the analysis of sample based capture–recapture experiments based on the variance of the population total estimator. Future research on the optimal allocation of funds should address questions of bias for three reasons, at least. First, an important way in which increased expenditure per sample plot can influence the analysis is that it may permit the recording of more explanatory variables. If those variables are relevant, then the added expenditure would be better seen as reducing the bias of the estimator for a population total, rather than reducing its variance. Second, it is possible that an increased sample size will permit us to study the possible variation in the coefficients of the logistic regression models. If variation exists, then a more complex (perhaps region specific) parametrization is called for. In this case allocation to sample size would tend to reduce the bias of the capture–recapture estimates, in addition to reducing the sampling variance. Third, increased expenditure may be used to improve data quality, thereby reducing data errors. Data errors typically cause bias in logistic regression, so also in this case the improvements would not be seen in the variance.

8. References

Alho, J. (1990). Logistic Regression in Capture–Recapture Models. Biometrics, 46, 623–635.

Alho, J. (1991). Variance Estimation in Dual Registration Under Population Heterogeneity. Survey Methodology, 17, 123–130.

Alho, J., Mulry, M., Wurdeman, K., and Kim, J. (1993). Estimating Heterogeneity in the Probabilities of Enumeration for Dual System Estimation. Journal of the American Statistical Association, 88, 1130–1136.

Boyce, W.E. and DiPrima, R. (1970). Introduction to Differential Equations. New York: John Wiley.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed. New York: John Wiley.

Hogan, H. (1992). The 1990 Post-Enumeration Survey: An Overview. The American Statistician, 46, 261–269.

Hosmer, D.E. and Lemeshow, S. (1989). Applied Logistic Regression. New York: John Wiley.

Huggins, R.M. (1989). On the Statistical Analysis of Capture Experiments. Biometrika, 76, 133–140.

Mulry, M. and Spencer, B.D. (1991). Total Error in PES Estimates of Population. Journal of the American Statistical Association, 86, 839–855.

Rao, C.R. (1973). Linear Statistical Inference and its Applications, 2nd ed. New York: John Wiley.

Seber, G.A.F. (1982). The Estimation of

Animal Abundance, 2nd ed. New York: Griffin.

Sekar, C.C. and Deming, W.E. (1949). On a Method of Estimating Birth and Death Rates and the Extent of Registration. Journal of the American Statistical Association, 44, 101–115.

Appendix 1. Expression for $\mathrm{Var}(\tilde{Z}|s)$

Define a sampling indicator for individual i as follows. Let $\delta_i(s) = 1$, if $i \in I_s$, and $\delta_i(s) = 0$ otherwise. Consider $V_3(s)$ of (3.3) first. Extend summation to $i = 1, \ldots, N$. Replace M_i by $\phi_i \delta_i(s)$, and parameter estimates by true values, to get $\nu_3(s)$. Defining $\nu_3 = E[\nu_3(s)]$ we can uncondition the effect of sampling by further replacing $\delta_i(s)$ by f_i.

Consider (3.2) now. Replace M_i by $\delta_i(s)$ (not by $\phi_i \delta_i(s)$) in $\tilde{\psi}_i$ and $\tilde{\psi}_{N+1}$, $i = 1, \ldots, N$, and parameter estimates by true values. Furthermore, replace $\tilde{W}$ by $W(s)$ which consists of diagonal $N \times N$ matrices $W_j(s) = (w_{ik}^j(s))$, with

$$w_{ii}^j(s) = \delta_i(s)(p_{ji} - p_{ji}^2/\phi_i), \quad j = 1, 2,$$

$$w_{ii}^3(s) = \delta_i(s)(p_{1i}p_{2i} - p_{1i}p_{2i}/\phi_i).$$

This yields $\nu_2(s)$. With these notations we can write $\mathrm{Var}(\tilde{Z}|s) = \nu_2(s) + \nu_3(s)$.

In general the calculation of $E[\mathrm{Var}(\tilde{Z}|s)]$ is hard. Although the expression for ν_3 was easily obtained, an expression for $\nu_2 = E[\nu_2(s)]$ is not. A first order approximation is obtained by replacing $\delta_i(s)$ by its expected value f_i.

Received November 1992
Revised February 1994