Role and Use of Composite Sampling and Capture–Recapture Sampling in Ecological Studies

M. T. Boswell, K. P. Burnham and G. P. Patil

Introduction

The physical mixing of samples with other samples or with the population has turned out to be a basis of some important sampling procedures. Sampling with replacement may be interpreted as returning a sample to the original population and thoroughly mixing it before the next sample is selected. This type of sampling has been quite common in practice.

A relatively recent sampling procedure, called composite sampling, involves physically mixing of samples before measuring, counting, or otherwise analyzing the composite sample. Pertinent statistical analysis is able to extract most of the information from the composite sample that can otherwise be extracted from the measurements on the individual original samples before they are physically mixed. The savings in the cost of measuremental analyses can be substantial.

Another sampling procedure, called capture–recapture sampling, involves physical mixing of a sample back into the original population. While composite sampling, and sampling with replacement, are used to estimate the population density/abundance, capture–recapture sampling is used to estimate population size and survival at of individuals.

Both composite sampling and capture–recapture sampling techniques have been refined and adapted in response to the varying needs involving different kinds of parameters of the populations of interest. The purpose of this paper is to provide a perspective of these sampling procedures. The material for composite sampling is taken from Boswell and Patil (1987).

1 Prepared in part as a Visiting Professor of Biostatistics, Department of Biostatistics, Harvard School of Public Health and Dana–Farber Cancer Institute, Harvard Medical School, Harvard University, Boston, MA.
1. Composite sampling

1.1. Introduction

A composite sample is formed by taking a number of individual samples and physically mixing them. Composite samples are used for different purposes. Usually the goal is to obtain the desired information in the original samples but at reduced cost or effort.

There are two general types of applications: estimating the mean of a stochastic process and identifying the individuals with a certain trait. Composite sampling can also be used to estimate the fraction of a population that possesses a given trait while maintaining the privacy or confidentiality of the individuals.

Composite samples have been formed from water (Becker, 1977; Roskopf, 1968; Schaeffer and Janardan, 1978; Schaeffer et al., 1980) and from bulk materials such as soils and fertilizers (Rohde, 1979; Brown and Robson, 1975; Hueck, 1976). In these cases, a single measurement on the composite sample is used in place of the average of individual measurements of the original samples.

In most of the applications, it is hoped that the mean of the process can be estimated by a single measurement taken on the composite sample, thereby saving the cost of taking many measurements. Considerable savings in the overall cost may be realized. What is lost is the ability to estimate the variability in the measurement process. However, by processing several composite samples, the variance may be estimated, still at considerable savings.

Instead of analyzing the entire composite, subsamples or aliquots are drawn and processed separately. This allows the measurement errors to be taken into account. Further, several composite samples may be formed from subsamples or 'increments' drawn from the original samples. This allows the variance of the original observation to be estimated.

1.2. Composite samples of known proportions

1.2.1. Composite samples for the detection of a certain trait

Laboratory procedures (such as blood tests for a disease), to see if a certain trait is present or not, can use compositing of samples to reduce costs (see, for example, Feller, 1968, p. 239; Garner, Stapanian and Williams, 1987). If the problem is to identify every individual with a rare trait, compositing several individual samples and testing the individual samples only when the composite sample tests are positive, has the potential of greatly reducing the number of tests required. Feller gives references to several generalizations involving two stage sampling.

The expected number of tests required for a procedure that composites \( n \) samples and tests all \( n \) samples when the composite sample exhibits the trait, can be calculated explicitly in terms of the unknown incidence of the trait, \( p \). The expected number of tests is

\[
E[N] = 1 \cdot (1 - p)^n + (n + 1)[1 - (1 - m)^n] = (n + 1) - n(1 - p)^n,
\]

and the relative factor (RCF) equals \( 1 + 1/n - (1 - p)^n \) tests per individual.

For \( p \) less than about 0.29, the graph of RCF versus \( n \) first decreases to a global minimum less than 1, then increases to a local maximum, and then decreases to an asymptote greater than the minimum. For larger values of \( p \) the graph steadily decreases but never falls below 1; thus, composite sampling is not advantageous for large \( p \). If the RCF = 0.5, then compositing results in a saving of 50% of the tests required if compositing is not done.

1.2.2. Estimation of incidence of a trait with composite sampling for confidentiality

When testing for a trait is likely to cause embarrassment to an individual, confidentiality can be assured by compositing samples from several individuals and only testing the composite. If the composite has the trait, then the individual(s) who have the trait are still unknown. Let \( p \) be the unknown incidence of the trait. A composite sample of \( n \) individual samples will exhibit the trait with probability \( 1 - (1 - p)^n \). Test a (large) number, \( m \), of composite samples, and let \( x \) be the total and \( p_m \) be the fraction of these that exhibit the trait. Then

\[
\hat{p} = 1 - (1 - p_m)^{1/m}
\]

is an estimator of \( p \) with a positive bias (see Boswell and Patil, 1987).

1.2.3. Composite sampling to test for compliance

As in Section 1.2.1, it is desired to identify cases having a certain trait. The measurement is a continuous variable which must satisfy a certain criterion. All cases that do not satisfy this criterion are to be identified. If the compliance rate is high, considerable savings can be realized by composite sampling.

Let \( c \) be a criterion value that should not be exceeded. For example, the measure may be the level of a toxicant in fish or other food products which are routinely tested. If the level exceeds \( c \), then the product can not be used.

For a composite of \( n \) samples, if any one sample exceeds \( c \), then the composite sample of \( n \) individuals will exceed \( c/n \). Of course, it is possible that none of the samples will exceed \( c \) and the composite will still exceed \( c/n \). If the composite exceeds \( c/n \), then it is necessary to test each individual sample. On the other hand, if the composite does not exceed \( c/n \), then none of the individual samples need to be tested.

Let \( F(x) \) be the distribution function of the individual samples; then the \( n \)-fold convolution of \( F \), \( F^{(n)}(x) = F(x) \ast F(x) \ast \cdots \ast F(x) \), is the distribution function of the sum. The probability that a composite sample of \( n \) individual samples exceeds the sum is \( 1 - F^{(n)}(c) \). The cost per individual is

\[
\frac{F^{(n)}(c) + (n + 1)\left[1 - F^{(n)}(c)\right]}{n} = 1 + \frac{1}{n} - F^{(n)}(c).
\]

As \( n \) increases, \( F^{(n)}(c) \) decreases to a limiting value of zero. The optimal composite size depends on the distribution of the sum.
1.2.4. Compositing to reduce variance in the presence of inexact analytical procedures

In the above discussions, the only variance of the estimator was that associated with the heterogeneity of the population and sampling. If the outcome of the test applied to a sample has an error with mean zero and variance $\sigma^2$, and if the population has a variance of $\sigma^2$, then the variance of $\bar{X}$, the average of $n$ samples, is $(\sigma^2 + \sigma^2)/n$. The variance of a composite sample of $n$ individual samples is $\sigma^2/n + \sigma^2$. Thus, the variance of the composite sample is larger than the variance of the average of $n$ samples; however, the composite involves only 1 measurement. Further, if the variance of $\sigma^2$ of the test is small in comparison to $\sigma^2$, the variance due to the heterogeneity of the population, then these two methods of estimating the mean may have variances that are not very different. If $k$ composite samples of size $n$ are formed, then the estimates based on the average of these $k$ measurements will have variance $\sigma^2/(kn) + \sigma^2/k$ which can be significantly smaller than $(\sigma^2 + \sigma^2)/n$, and the cost of estimation may still be reduced.

1.2.5. Compositing to reduce the cost of estimating the mean

The usual method of estimating the mean of a stochastic process is to take $n$ observations and calculate the sample mean. Nearly the same information can be obtained by mixing the $n$ observations and taking a single measure from the composite sample. Some laboratory procedures are time consuming or costly. In such cases, being able to make a single measurement instead of $n$ measurements, is a great benefit.

Let $X_1, \ldots, X_n$ be measurements on $n$ independent samples, and let $Z$ be the measurement of the composite of the $n$ samples. Then

$$Z = \sum_{i=1}^{n} w_i X_i, \quad E[Z] = \sum_{i=1}^{n} w_i \mu = \mu, \quad V(Z) = \sum_{i=1}^{n} w_i^2 \sigma^2,$$

where $w_i$ is the fraction of the composite sample coming from the $i$th sample. If each sample consists of the same amount of material, then $w_i = 1/n$ and $Z = \bar{X}$, and the estimator of $\mu$ based on the composite sample is exactly the same as that based on the $n$ samples.

The procedure described above assumes the fractions $w_1, w_2, \ldots, w_n$ to be fixed and known. Examples of this kind of compositing include:

1. Stratified random samples are of this form, showing that a measurement of the composite sample formed by combining the individual samples gives the same result.
2. Compositing of soil samples to obtain the overall or average fertility of the soil. Other bulk samples include sampling of fertilizers, coffee beans and concrete mixes.
3. Compositing of samples taken from waste sites to obtain the overall concentration of hazardous material.
4. Compositing of filtrate (assuming a known amount of water is filtered) to estimate the abundance of various plankton species.

1.3. Estimation of the mean using composite samples of random proportions

Example 4 above could be a situation where a net is towed through a water body for a constant time period at a constant speed. The amount of water filtered depends on many random conditions, such as wind speed, wave height and water currents. The fraction of each sample in the composite is then a random quantity. The measurement on the composite sample is $Z = \sum_{i=1}^{n} W_i X_i$, where $W_i$ is a random variable representing the fraction of the composite sample from the $i$th sample. This same formula holds if the composite sample is formed by combining only a part of a subsample ('increment') from each sample. Using this procedure, several composite samples can be formed from the same collection of samples. Figure 1 illustrates the formation of composite samples taken from 'increments' of the original samples. Subsamples taken from each composite sample and tests or measurements made on these subsamples are also illustrated (heterogeneity within all of these samples can be accounted for).

![Diagram](image-url)

Rohde (1976) develops the theory for a single composite sample formed from $n$ samples. The proportions of each sample used to make the composite sample are either fixed (known) or random. Elderd (1977) (see also Elderd, Thompson, and Myers, 1980) generalizes to $r$ composite samples, each formed from $n$ increments, and $s$ subsamples taken from each composite sample. Further, $t$ analyses or tests are done on each subsample. This generalization takes into account the variability.
in dividing the original samples into increments before compositing, variability due to non-perfect mixing of the composite samples and the selection of subsamples, and, finally, the variability of the test procedure itself.

The properties of the various composite-sample procedures are studied in the sections below. Formulas for the mean and variance of the measurement or average of measurements from composite samples are obtained. If the means of the observations vary, as they do in stratified sampling, then stratified means are estimated. On the other hand, if the means are constant, then these procedures give unbiased estimators of the common means.

It would be desirable to estimate the variance of the observations. This information is lost when composite samples are made. A possible solution is discussed in Section 1.3.2.

The means and variances of the composite-sample estimators depend on the distributions of the weights and the observations. Two statistical distributions for the weights are studied. Rohde argues for the Dirichlet distribution, and Elder (1977) formulates a model for the weights which gives the multivariate hypergeometric distribution which is the model originally used by Brown and Fisher (1972). Elder points out that both this distribution and the Dirichlet distribution converge to singular multivariate normal distribution under certain suitable conditions. He suggests this to be a reasonable approximation in many cases because of the physical averaging that occurs in blending a composite sample.

1.3.1. A single composite sample

Let \( Z = \sum_{i=1}^{r} W_{j}X_{j} \) be the measurement on the composite where \( W_{j} \) is the random proportion of the \( i \)th sample in the composite and \( X_{j} \) is the measurement on the \( i \)th individual sample. We assume the \( W_{j} \)'s and the \( X_{j} \)'s are independent.

In the matrix notation \( Z = W^{T}X \) Let the means be \( E[W] = \mu_{w} \) and \( E[X] = \mu \); let the variance-covariance matrices be \( C_{w} = \text{Cov}(W) \) and \( C_{x} = \text{Cov}(X) \). Then \( E[Z] = \mu_{w} \mu \). Further, the variance of \( Z \) is

\[
\text{Var}(Z) = \mu_{w}^{2} C_{w} \mu_{w} + \mu_{x}^{2} C_{x} \mu_{x} + \text{tr}(C_{w} C_{x}).
\]

1.3.2. Several composite samples

If composite samples are formed by taking subsamples, also called increments or aliquots, from the \( n \) original samples, the joint distribution of the fractions of the original measurements represented in the composite sample must have the properties given above. If \( Z = W^{T}X \) and \( Y = U^{T}X \) are two composite samples formed from aliquots of the original samples, then

\[
\text{Cov}(Y, Z) = \mu_{u}^{T} C_{u} \mu_{u} + \mu_{w}^{T} C_{w} \mu_{w} + \text{tr}(C_{u} C_{w})
\]

where \( C_{u,w} \) is the covariance \( \text{Cov}(U, W) = (\text{Cov}(U_{i}, W_{j})) \).

In many cases, it is reasonable to assume the random weights are exchangeable random variables. Then \( E[W_{j}] = 1/n \) and \( \text{Cov}(W_{i}, W_{j}) = \sigma_{w}^{2}/r \) is a constant. This implies that

\[
C_{w} = \sigma_{w}^{2} \begin{bmatrix}
1 & \rho & \cdots & \rho \\
\rho & 1 & \cdots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \cdots & \rho & 1
\end{bmatrix}.
\]

As Rohde (1976) pointed out, the symmetric Dirichlet distribution has these properties. Now suppose that \( r \) measurements are all that can be made. We then construct \( r \) composite samples and compare the results with \( r \) measurements on original samples. As before, let \( X_{1}, X_{2}, \ldots, X_{n} \) be the values associated with \( n \) samples. By subsampling \( r \) times from each sample we form composite samples by combining one subsample from each sample. Then \( Z_{i} = W_{i}^{T}X \) is the \( i \)th composite sample for \( i = 1, 2, \ldots, r \). We compare \( \bar{Z} = (1/r) \sum_{i=1}^{r} Z_{i} \) with the average \( \bar{X} = (1/r) \sum_{i=1}^{r} X_{i} \) of the original samples. Since

\[
\bar{Z} = (1/r)(W_{1} + \cdots + W_{r})X = V^{T}X,
\]

\( \bar{Z} \) represents a composite sample measurement and the formulas presented above apply. Both \( \bar{X} \) and \( \bar{Z} \) are unbiased estimators of \( \mu \). The variance \( \text{Var}(\bar{X}) = \sigma_{x}^{2}/r \).

Let \( C_{x}, C_{w}, C_{z} \) be the variance-covariance matrices of \( X, W \) and \( Z \), respectively. The variance of \( \bar{Z} \) is

\[
\text{Var}(\bar{Z}) = \mu_{w}^{2} C_{w} \mu_{w} + \mu_{x}^{2} C_{x} \mu_{x} + \text{tr}(C_{w} C_{x})/r.
\]

Now assume \( X_{1}, X_{2}, \ldots, X_{n} \) to be independent identically distributed with mean \( \mu \) and variance \( \sigma^{2} \). It follows from (4), above, that

\[
\text{Var}(\bar{Z}) = \sigma_{w}^{2} n \sigma_{x}^{2} \sigma_{w}^{2} \frac{n^{2} \sigma_{w}^{2}}{r (n + 1)}.
\]

Assume \( W \) to have a Dirichlet distribution with parameters \( 1 \); then

\[
\text{Var}(\bar{Z}) = \sigma_{w}^{2} \frac{n^{2} \sigma_{x}^{2} \sigma_{w}^{2}}{r (n + 1)}.
\]

Also \( \text{Var}(\bar{X}) = \sigma_{x}^{2}/r \). The ratio of the variances is

\[
\frac{\text{Var}(\bar{Z})}{\text{Var}(\bar{X})} = \frac{r}{n} \left[ 1 + \frac{n - 1}{r (n + 1)} \right].
\]

If \( n \) is large compared to \( r \), then the variance of \( \bar{Z} \) is much smaller than the variance of \( \bar{X} \).
It seems that choosing \( n \) large in comparison to \( r \) would result in the best composting situation. However, when a composite sample is made, often the physical mixing is imperfect, resulting in more variability. The optimal choice may also depend on the cost of sampling and the cost of analysis of samples.

Rohde (1976) points out that the sample variance of the composite samples is a biased estimator. In fact, \( E[S^2] = n \sigma^2 \). Rohde suggests two approaches. If an independent estimate of \( \sigma^2 \) is available, then \( S^2 \) can be used to estimate \( \sigma^2 \). The other approach is to assume some model, such as the Dirichlet distribution, which reduces the number of parameters that need to be estimated. For further details, see Rohde (1976) and Elder (1977).

1.3.3. Subsampling of the composite samples.

The comparisons given above are based on the models discussed by Rohde (1976). Elder, Thompson and Myers (1980) generalize Rohde’s work to allow for within aliquot variability. That is, when the composite samples are formed by subsampling from the original samples, different concentrations are possible. Thus, different composite samples can be formed from different random variables.

Further, Elder et al. also generalize to include the testing error. Suppose \( r \) composite samples are formed as before. From each composite sample, select \( s \) subsamples and run \( t \) analyses on each subsample. The estimator of the mean of the original population is the overall average. Let the result of the \( j \)th test from the \( i \)th subsample from the \( k \)th composite sample be represented by \( Z_{ijk} \). Then \( Z_{ijk} = \frac{1}{n} \sum_{i=1}^{n} W_{ij} X_{ij} + \epsilon_{ijk} \), where \( X_{ij} \) depends on the composite index \( i \). This assumption incorporates the variability from dividing a sample into \( r \) aliquots; this is called within increment variability. The variability of the testing procedure results in \( \epsilon_{ijk} \). The \( r \times s \times t \) tests are assumed to have independent identically distributed errors with \( E[\epsilon] = 0 \) and \( \text{Cov}(\epsilon) = \sigma^2 I \). The technique used to find the mean and variance of the estimator \( \bar{Z} \) is similar to that described above. The final result is that \( \bar{Z} \) is an unbiased estimator of \( \mu \), with variance

\[
V(\bar{Z}) = \frac{\sigma^2}{rn} + \frac{n \sigma^2}{rs} + \frac{rs}{rst} \sigma^2.
\]

If the increments used to form the composite are either made up of discrete (equal sized) pieces or are divided into equal sized pieces then a hypergeometric distribution can be used to model the number of pieces from the various increments that make up the composite samples that finally end up in a subsample (Brown and Fisher, 1972; see also Elder, 1977). Assume each increment is composed of \( g \) equal-sized pieces and that each subsample is formed from a random selection of \( G \) pieces. Taking the limit as \( g \) and \( G \) go to infinity, while keeping \( g/G \) constant results in the distributions of the weights approaching a singular multivariate normal distribution, see Elder (1977).

The best values of \( n, r \) and \( t \) depend on many factors. Elder (1977) discusses this, pointing out that closed form solutions do not exist. Generally, if the cost of sampling is relatively small, then \( n \) can be much larger than \( r \). If the within increment variability is small, then \( r \) is small. If the test procedure has small errors, then \( t \) is small. Often \( r \) and \( t \) can be taken as one or two.

1.4. Examples

1.4.1. Power plant effect on environment

Consider a situation presented by Rohde (1979) where a small scale study was made to see if composting would be useful in studying the effect of power plant construction on the environment. For the Crane Power Plant near Baltimore City, data was collected on the density of plankton for one time period. A transect near the intake to the plant was selected starting in deep water and ending in shallow water near shore. Samples were taken at six locations along the transect. The first two, in deep water, were sampled near the top and near the bottom. The density of plankton seemed to be smaller near the shore and near the bottom. All locations were sampled with equal effort by pumping water for 10 minutes at a rate of 37.8 gal/minute into a collecting net. The filtrate was washed into a 250 ml bottle from which 1 ml subsamples were selected. Actual measurements of the samples and of composite samples are given in Table 1. It is clear that about the same results were obtained from the composite samples as from the average of the original samples. The average of the eight observations was 203.2 organisms per liter while the density in the composite sample was 209.3 organisms per liter.

<table>
<thead>
<tr>
<th>Location</th>
<th>Density organisms/liter</th>
<th>Density of average/of composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 top</td>
<td>269.3</td>
<td>260.8/253.3</td>
</tr>
<tr>
<td>1 bottom</td>
<td>253.3</td>
<td></td>
</tr>
<tr>
<td>2 top</td>
<td>195.0</td>
<td>167.2/170.3</td>
</tr>
<tr>
<td>2 bottom</td>
<td>139.3</td>
<td>262.8/286.0</td>
</tr>
<tr>
<td>3</td>
<td>282.3</td>
<td>122.0/131.3</td>
</tr>
<tr>
<td>4</td>
<td>243.3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>126.0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>118.0</td>
<td></td>
</tr>
</tbody>
</table>

1.4.2. Heavy metal pollution

A study of heavy metal pollution of aquatic environments was reported by Hueck (1976). In this study, mussels from relatively unpolluted water were transplanted to various locations under study. After a period of time, tissue from the mussels was analyzed for heavy metal content. Since there was a great local variation within a region, pooling of results from adjacent localities seemed reasonable. Composite samples were formed by homogenizing the tissue of several specimens. The amount of heavy metal in the sample was used as an indicator
of the pollution of the region. A map of coastal region of France, Belgium, Netherlands, Germany, Denmark and Great Britain giving the results for mercury is given in Hueck (1976). The measurements varied from 60 to 1650 ppb. Other maps show the regional change in mercury, copper and zinc.

2. Capture–recapture sampling

2.1. Introduction

Capture–recapture sampling is used in ecology and wildlife studies to estimate animal population size (N) and associated demographic parameters such as survival rate. The basic procedure is to mark and release a sample of individuals into the target population and then, assuming a thorough mixing of this sample with the population, recapture marked and unmarked individuals in one or more subsequent samples. Capture rates and marked-to-unmarked ratios are the basis for parameter estimation and inference.

In the simplest study, an initial capture effort results in the release of n1 marked animals. A short interval of time later (relative to the population dynamics of survival, recruitment and dispersal), a second sample is captured, of size n2. The second sample will consist of both previously marked animals (m2) and unmarked individuals (u2). Let p2 be the average capture probability for the second sample. Assuming that marking has not affected recapture, then E(n1) = Np1, E(m2|n1|n2) = p2, and the classical Petersen estimator of N is \( \hat{N} = n_1n_2/m_2 \). The Petersen estimator provides the heuristic basis of most estimators of wildlife population size, and from this humble beginning (Petersen, 1896) a very large capture–recapture literature has developed (see, e.g., Seber, 1982).

We use capture–recapture here in a generic sense to refer to any ecological methodology that uses the release and recapture of marked individuals in order to estimate a wildlife population size or a survival rate. Statistical models for these studies are said to be either closed or open. Often, the sampling occurs during a short enough time interval that the wildlife population can be assumed to be closed: no unknown additions or removals. Then only the parameter of interest is the population size, N. In contrast, the open population models emphasize estimating survival rates. For example, bird banding studies involve the yearly release of thousands of banded waterfowl. Bands are returned by hunters and these release-recapture data are used to estimate waterfowl annual survival rates, S. The statistical models for the band 'recaptures' are conditional on the known releases (see, e.g., Brownie et al., 1985). However, if we can also model the process of capturing unmarked individuals, which are then marked and released, we can also estimate yearly population sizes.

Capture–recapture does involve sampling populations. However, this methodology does not correspond well to the usual finite population sampling paradigm of statistics. Definition of the sampled population is a major problem. There is no frame, and estimation of N is often a primary study objective. For more critical than unknown N is that the sampling probabilities (p) cannot be determined in advance, nor can they be exactly determined after sampling. Rather, these probabilities must be estimated, and factors that can affect them must be considered in the models. Instead of classical statistical sampling theory, the basis for analysis of capture–recapture data is sampling models (primarily, binomial, multinomial, hypergeometric and multiple hypergeometric) incorporating some, or all, of the three types of parameters N, p and S. The central issue in the analysis of capture–recapture data is what model to use (see e.g., Cormack, 1979, p.248).

2.2. Closed models

The literature on closed models for capture–recapture has been recently summarized by White et al. (1982). They elaborate a series of increasingly general models and provide a computer program to fit these models to data. In wildlife and ecology we can have k = 5 to 10, and sometimes more, capture occasions on successive days before the assumption of closure becomes untenable.

In the case of 2 occasions the basic statistical model for N is the hypergeometric distribution:

\[ \Pr \{ m_2 | n_1, n_2, N \} = \binom{n_1}{m_2} \binom{N-n_1}{n_2-m_2} / \binom{N}{n_2} ; \]

here we condition on the initial releases n1 and total captures at time 2, n2. For a complete sampling model of the data we can start with

\[ \Pr \{ n_1, n_2, m_2 | N, p_1, p_2 \} = \left[ \binom{N}{n_1} \binom{N-n_1}{n_2-m_2} \binom{N-n_1}{n_2-m_2} \right] \frac{(p_1)^{n_1}(1-p_1)^{N-n_1}}{m_2} \frac{(p_2)^{m_2}(1-p_2)^{N-m_2}}{(n_2-m_2)} \]

which factors into two binomials and the hypergeometric given above.

Chapman (1948, 1951) has thoroughly investigated the hypergeometric model. In particular his recommended point estimator is \( \tilde{N} = (n_1 + 1)(n_2 + 1)/(m_2 + 1) - 1 \). Confidence limits should be computed as \( N_{U} \) and \( N_{L} \) satisfying

\[ \sum_{i = 0}^{m} \binom{n_1}{i} \binom{N_U-n_1}{n_2-i} / \binom{N_U}{n_2} = \alpha_U , \quad \sum_{i = m}^{n_1} \binom{n_1}{i} \binom{N_L-n_1}{n_2-i} / \binom{N_L}{n_2} = \alpha_L \]

(Chapman, 1948).

We illustrate this estimator with data used by Buckland (1984). The size of the workers component of an ant colony was estimated by initially capturing, marking
and releasing 500 ants. A second sample had \( n_2 = 189 \) and \( m_2 = 17 \) recaptures: \( \hat{N} = (501)(190)/(18) - 1 = 5287 \). Using \( a_2 = a_4 = 0.025 \), the solution of the nonlinear equations for \( N_1 \) and \( N_4 \) gives 3603 to 9330 as the 95% CI on \( N \).

When there are multiple recapture occasions the model can allow for possible variations in capture probabilities due to factors such as time, animal behavior (response to capture) and individual differences in capture probabilities (heterogeneity). White et al. (1982) elaborate a sequence of models which incorporate these features. For example, if there are \( k \) capture occasions, the Petersen model generalizes to Darroch's model (Darroch, 1958) with parameters \( N, p_1, \ldots, p_k \), where \( p_i = \) the capture probability on occasion \( i \). The complete data are the capture histories of each animal captured, provided unique marks are used. A minimal sufficient statistic is \( n_i, \ldots, n_k \) and \( M \) where \( n_i = \) the number of animals captured on occasion \( i \) and \( M = \) the total number of different individuals captures over all days. The maximum likelihood estimator is not closed-form. An approximation to the MLE is the solution of

\[
M = N \left[ 1 - \left( 1 - \frac{n_i}{N} \right)^k \sum_{i=1}^{k} \prod_{j=1}^{i-1} (1 - p_j) \right],
\]

which derives from \( E(M) = N[1 - \prod (1 - p_i)] \) and \( E(n_i) = Np_i \). Complete theory on this model is in Darroch (1958).

Another common situation is removal sampling, especially electrofishing studies (White et al., 1982, Chapter 4). In removal sampling the data are \( u_1, \ldots, u_k \), where \( u_i = \) number of unmarked individuals captured on occasion \( i \). Under the assumption that removal probability is constant, the data is multinomial:

\[
\Pr \{u_1, \ldots, u_k|N, p\} = \binom{N}{u_1, \ldots, u_k} \left( \prod_{i=1}^{k} (1 - p)^{u_i} \right) (1 - p)^{kN - u_1 - \cdots - u_k}
\]

The MLE is not closed form if \( k > 2 \). For \( k = 2 \), \( \hat{N} = u_1/(1 - (u_2/u_1)) \). This removal model was first rigorously studied by Zippin (1956).

More complex closed models exist, in particular ones that allow for heterogeneity of individual capture probabilities (see, e.g., Burnham and Overton, 1978), and both behavioral response to first capture and heterogeneity (White et al., 1982). Program CAPTURE (White et al., 1982) incorporates all these models and considers model selection, and thus provides for near-comprehensive analysis of closed population capture data.

Log-linear methodology also provides a basis for comprehensive analysis of closed population capture data (Cormack, 1979). The two approaches overlap to a large extent, but not totally. There are log-linear models that have no counterparts in White et al. (1982); also some models of White et al. have no log-linear counterparts. However, our key point is that comprehensive model systems do exist for the analysis of capture-recapture data from closed wildlife populations.

2.3. Open models

It has been mostly since the early 1960's that a thorough foundation has appeared in the literature for the estimation of parameters based on capture-recapture sampling of open populations. One branch of the literature includes bird-banding and fish-tagging studies that yield data from a single, terminal, harvest-related recovery, as reviewed by Brownie et al. (1985). Another branch deals with multiple (live) recaptures of marked animals; we will refer to this as Jolly-Seber sampling (and Jolly-Seber models) after two of the leading contributors. Seber (1982, 1986) reviewed this literature. The links between these two major developments were outlined by Brownie et al. (1985). Until recently, the two approaches were developed separately, usually with different notations and contexts. Recently, however, there has been growing recognition that most open capture-recapture models are special cases of a more general theory. Now the entire subject of open populations, capture-recapture can be unified under one umbrella of theory (see, e.g., Burnham et al., 1987).

2.3.1. Bird-banding

In a typical banding study a sample of birds is captured, banded and released just before the annual hunting season for a number of consecutive years. The data are symbolized as

\[
N_i = \text{the number of birds released (banded) in year } i,
\]

\[
R_{ij} = \text{the number of band recoveries from birds shot in hunting season } j \text{ from birds released in year } i, \quad i = 1, \ldots, k; \quad j = i, \ldots, k.
\]

A convenient tabular representation is

<table>
<thead>
<tr>
<th>Releases</th>
<th>Recoveries</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_1 )</td>
<td>( R_{11} ) ( R_{12} ) ( R_{13} ) ( \cdots ) ( R_{1k} )</td>
</tr>
<tr>
<td>( N_2 )</td>
<td>( R_{22} ) ( R_{23} ) ( \cdots ) ( R_{2k} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \ddots )</td>
</tr>
<tr>
<td>( N_k )</td>
<td>( R_{kk} )</td>
</tr>
</tbody>
</table>

For example, from Brownie et al. (1985, p. 21):

<table>
<thead>
<tr>
<th>Year</th>
<th>Releases</th>
<th>Recoveries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1964</td>
<td>1603</td>
<td>124 44 37</td>
</tr>
<tr>
<td>1965</td>
<td>1595</td>
<td>62 76</td>
</tr>
<tr>
<td>1966</td>
<td>1197</td>
<td>82</td>
</tr>
</tbody>
</table>

These are data from a banding study on adult male wood duck.
The model structure under the assumption that recovery rate and survival rate parameters are time-specific only is

\[ E(R_j) = \begin{cases} N_i f_i, & j = i; \ i = 1, \ldots, k, \\ N_i S_i, \ldots, S_{j-1} f_j, & j = i + 1, \ldots, k. \end{cases} \]

We assume that the \( k \) cohorts are each multinomial random variables, i.e., \( R_{i1}, \ldots, R_{ik} \), where \( N_i - R_i \) are uncertain, \( j = 1, \ldots, k \), where \( \pi_{ij} = E(R_{ij})/N_i \), \( \lambda_i = 1 - \sum_{j=1}^{k} \pi_{ij} \), and \( R_i = \sum_{j=1}^{k} R_{ij} \).

The statistical theory of this, and other banding models, is well developed in, e.g., Brownie et al. (1985). The minimal sufficient statistic can be taken as \( R_{i1}, \ldots, R_{ik}, C_i, \ldots, C_{k-1} \), where \( C_i = \sum_{j=1}^{i} R_{ij} \) are the column sums, \( i = 1, \ldots, k \). Given these row and column sums, we define block totals \( T_i = R_{i1}, T_2 = T_1 - C_1 + R_{12}, \ldots, T_{i+1} = T_i - C_i + R_{i+1}, \) and \( T_k = C_k; T_i \) is all recoveries, at and after recovery year \( i \), from birds released in, or before, year \( i \).

The ML estimators and their sampling variances are

\[
\hat{f}_i = \frac{R_{i1} C_i}{N_i T_i}, \quad i = 1, \ldots, k, \\
\hat{S}_i = \frac{R_i}{N_i} \left[ 1 - \frac{C_i}{T_i} \right] / \left[ R_{i+1} / N_{i+1} \right], \quad i = 1, \ldots, k-1, \\
\text{var}(\hat{f}_i) = (\hat{f}_i)^2 \left[ \frac{1}{R_i} - \frac{1}{N_i} + \frac{1}{C_i} - \frac{1}{T_i} \right], \quad i = 1, \ldots, k, \\
\text{var}(\hat{S}_i) = (\hat{S}_i)^2 \left[ \frac{1}{R_i} + \frac{1}{R_{i+1}} - \frac{1}{N_{i+1}} + \frac{1}{T_{i+1} - R_{i+1}} - \frac{1}{T_i} \right].
\]

For the wood duck example

<table>
<thead>
<tr>
<th>( i )</th>
<th>( R_i )</th>
<th>( C_i )</th>
<th>( T_i )</th>
<th>( \hat{f}_i )</th>
<th>( \text{se}(\hat{f}_i) )</th>
<th>( \hat{S}_i )</th>
<th>( \text{se}(\hat{S}_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>208</td>
<td>127</td>
<td>208</td>
<td>0.079</td>
<td>0.0067</td>
<td>0.58</td>
<td>0.079</td>
</tr>
<tr>
<td>2</td>
<td>138</td>
<td>106</td>
<td>219</td>
<td>0.042</td>
<td>0.0045</td>
<td>0.63</td>
<td>0.094</td>
</tr>
<tr>
<td>3</td>
<td>82</td>
<td>195</td>
<td>195</td>
<td>0.071</td>
<td>0.0076</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

Numerous other models are well developed for analysis of band-recovery data. For birds banded as adults, one can consider the case of all annual survival rates being constant \( (S_i = S, f_i = f) \). Generalizations also exist to allow for possible response of birds to being captured, banded and released; primarily this could cause a different recovery or survival rate in the first year after release. Age-specific models exist for the case of two or more age classes being banded at the same time. Consideration of model selection and tests of assumptions are summarized in Brownie et al. (1985). There is a well-developed literature for the statistical analysis of band-recovery data.

### 2.3.2. The Jolly–Seber model

The extension of capture–recapture to fully stochastic, open models was not entirely successful until 1965 (Jolly, 1965; Seber, 1965). The Jolly–Seber model assumes time-specific parameters; it is analogous to the time-specific banding model considered in Section 2.3.1. The complete specification of the model entails a substantial amount of notation; we present here only the summary statistics needed for parameter estimation. Moreover, we present the model in a manner that leads to a unified theory of open models, but using here the notation that has arisen in Jolly–Seber modelling (see, e.g., Jolly, 1965; Pollock, 1975). That notation is very different from the banding notation.

Let releases of marked animals be \( R_{i1}, \ldots, R_{ik} \) at occasions \( i = 1, \ldots, k-1 \). The released \( R_i \) animals constitute a cohort from which animals are lost by either being recaptured, dying, or permanently leaving the study area. Let \( m_{ij} \), \( j = i + 1, \ldots, k \) be the number of animals recaptured for the first time from the \( R_i \) releases at time \( i \). Upon recapture at time \( j \), an animal may be re-released as part of the cohort releases \( R_p \).

As typically done, capturing and releasing occur simultaneously for \( k \) occasions. The parameters of interest are population sizes \( N_i \) and capture probabilities \( p_i, i = 1, \ldots, k \), and \( \phi_i \) the probability of surviving from occasion \( i \) to \( i + 1 \), for \( i = 1, \ldots, k - 1 \). The \( \phi_i \) and \( p_i \) apply to all animals in the population at risk of capture on occasion \( i \). The estimable parameters are \( \phi_1, \ldots, \phi_{k-2}, p_2, \ldots, p_k \) and \( (\phi_1 p_1), (\phi_2 p_2), \ldots, (\phi_{k-1} p_{k-1}) \) and \( (N_1, p_1) \) and \( (N_k, p_k) \). The estimation of \( N_2, \ldots, N_{k-1} \) requires assuming that marked and unmarked animals have equal capture probabilities, this often fails.

In addition to the releases, \( R_i \) and (first) recaptures, \( m_{ij} \) we define \( u_i \), \( i = 1, \ldots, k \), the number of unmarked animals captured at times \( i = 1, \ldots, k \). These summary data are conveniently represented as

<table>
<thead>
<tr>
<th>( i )</th>
<th>( u_i )</th>
<th>( R_i )</th>
<th>( m_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( u_1 )</td>
<td>( R_1 )</td>
<td>( m_{12} ) ( m_{13} ) \ldots ( m_{1k} )</td>
</tr>
<tr>
<td>2</td>
<td>( u_2 )</td>
<td>( R_2 )</td>
<td>( m_{23} ) \ldots ( m_{2k} )</td>
</tr>
<tr>
<td>k-1</td>
<td>( u_{k-1} )</td>
<td>( R_{k-1} )</td>
<td>( m_{k-1,k} )</td>
</tr>
</tbody>
</table>
The minimal sufficient statistic is the row and column sums, \( r_1, \ldots, r_k \), \( m_1, \ldots, m_k \), and the initial captures \( u_1, \ldots, u_k \). Here \( r_j = \sum_{i=1}^{k} m_{ij} \), \( m_j = \sum_{i=1}^{k} m_{ij}, j = 2, \ldots, k \). We also define a block total, \( z_{ij} = r_i + z_{j(i)}, j = 2, \ldots, k - 1, z_{j} = r_j - m_j, j = 2, \ldots, k - 2 \), and in general, \( z_{i+1} = r_i + z_{i+1}, i = 2, \ldots, k - 2 \).

The total number of animals captured on occasion \( j \) is \( n_j = m_j + u_j \) (\( m_1 = 0 \)); the difference between captures and releases, \( n_j - R_j \), is losses on capture.

The row and column sums here are exactly like the row and column sums for banding data. Also, the block total \( m_j + z_j = T_j \) is exactly like the block total defined for banding data.

Given the releases, \( R_i \), the \( m_{i+1}, \ldots, m_k, R_i - r_i \) are multinomially distributed with cell probabilities

\[
\pi_j = \left\{ \begin{array}{ll}
\phi_j p_{j+1}, & j = i + 1 \\
\phi_j q_{j+1} \cdots (\phi_{i-1} q_{i-1}) (\phi_{i-1} p_i), & j > i + 1
\end{array} \right.
\]

If one defines \( \phi_j p_{j+1} = f_j \) and \( \phi_j q_{j+1} = S_j \) and makes allowance for the different indexing here, then this capture–recapture model is identical to the band recovery model.

The MLE's are

\begin{align*}
\hat{\phi}_i &= r_i \left[ \frac{m_{i+1}}{R_i} + \frac{z_{i+1}}{T_i} \left( \frac{R_i}{r_{i+1}} \right) \right], \quad i = 1, \ldots, k - 2, \\
p_i &= \frac{m_i}{m_i + z_i R_i / r_i}, \quad i = 2, \ldots, k - 1, \quad \hat{p}_i = \frac{r_{i+1}}{R_{i+1}}, \\
\hat{S}_i &= n_i / \hat{p}_i, \quad i = 2, \ldots, k - 1, \quad \hat{R}_i \hat{p}_i = u_i = n_i \quad \text{and} \quad \hat{R}_i \hat{p}_i = n_k.
\end{align*}

Brownie and Robson (1983) give the Jolly–Seber estimators in this formulation (see also Burnham et al., 1987). Usually, however, the Jolly–Seber model is conceptualized in terms of \( N_i, M_j, \) being the number of marked animals still alive at risk of capture just before capture occasion \( i \). The Petersen estimator of \( N_i \) is \( \hat{N}_i = n_i M_j / m_i \). A second application of the Petersen estimator uses \( E[z_i / (M_j - m_i)] = E[r_i / R_i] \). The recaptures \( z_i \) all come from the \( m_i - m_j \) animals marked in the population just after occasion \( i \), and the capture probability of any of these \( m_i - m_j \) animals is the same as that for the new releases of \( R_i \) animals. Hence, \( \hat{M}_j = m_i + z_i R_i / r_i \) and \( \hat{p}_i = m_i / \hat{M}_i \).

Just as with the banding models, there are many other Jolly–Seber (i.e., open capture–recapture) models. They are still proliferating and have not quite yet been systematized. Pollock (1975) presents a series of Jolly–Seber models which allow behavioral response to capture in both survival rates and capture rates. More recently, Pollock (1981) has produced Jolly–Seber models that incorporate age-effects. Brownie et al. (1986) consider models with survival and/or capture param-eters constant over occasions. There is also the extension by Cormack (1981) of log-linear models to open model capture–recapture. Seber (1986) reviews recent developments in open-population capture–recapture models.

2.4. Summary and discussion

A large literature on capture–recapture in ecology and wildlife has developed in the past 100 years; most of it is from about 1950 and the true proliferation of the subject started in the mid 1960's. The literature for closed models is simpler, and more complete, partly because the only biological parameter of interest is one population size, \( N \).

With open populations, the parameters of interest are survival, and population size; these are possibly different at each sampling time. When recapture is by harvest, as in bird-banding, only survival rate is estimable, then the number of models is tractable and a good comprehensive reference exists: Brownie et al. (1985). Capture-recapture on open populations has a very large and diverse literature. There has not yet been an accepted unification of this literature (let alone all open population capture literature). A cutting edge in capture-recapture research now is to produce such a unification, and statistical theory that is easier to use when developing new models.

Most of the models now in use for capture data require computer programs for their use. This will become increasingly true. A second cutting edge of research is to generate good computer programs for micro computers that allow the users to specify model structures. Flexibility and ease of use are the keys to good software implementation of currently available sophisticated capture-recapture models. There has been progress in this direction; see, for example, Arnason and Baniuk (1980), Clotert and LeBreton (1986), White (1983), Burnham et al. (1987), Cormack (1985), Conroy and Williams (1984).

We have a final set of comments on the extensions of capture-recapture models and methods to problems other than wildlife; there are numerous such applications and extensions. In ecology, capture-recapture models can be used to estimate species diversities (Burnham and Overton, 1979), and abundance of sessile objects such as nests (Magnusson et al., 1978). Both open and closed models have been applied to paleobiology problems of estimating species abundance and extinction rates (Nichols and Pollock, 1983).

Applications of capture–recapture modeling and analysis methods are frequently encountered in human populations in regards to estimating the size of 'hidden' subpopulations. Application areas include epidemiology and demography where the data base tends to be incomplete dual, or multiple record systems (El-Kharozat et al., 1977). There has been estimation of the size of criminal populations from arrest records (Greene and Stollmack, 1981). The Census Bureau uses capture–recapture models to estimate the undercount problem (Cowan and Malec, 1986). Applications of capture–recapture methods are also found in numismatology (Chao, 1984), quality control (Jewell, 1985), vocabulary of authors (Efron and Thisted, 1976) and remote sensing (Maxim et al., 1981).
Capture-recapture methods motivated and initiated by seemingly simple wildlife problems have expanded to become a major class of statistical methods with application far beyond just the estimation of wildlife numbers.

Acknowledgements

Preparation of this paper has been partially supported by the NOAA research grants to the Penn State Center for Statistical Ecology and Environmental Statistics under the auspices of the Northeast Fisheries Center. The NCI grant CA 23415 to Dana-Farber Cancer Institute provided partial support to G. P. Patil. Our thanks are due to R. C. Hennemuth and C. Taillee for interesting discussions on various themes of the paper.

References


