On the design of closed recapture experiments

Danilo Alunni Fegatelli∗,1 and Alessio Farcomeni **,1

1 Department of Public Health and Infectious Diseases (Sapienza - University of Rome)

Received zzz, revised zzz, accepted zzz

We propose a method to plan the number of occasions of recapture experiments for population size estimation. We do so by fixing the smallest number of capture occasions so that the expected length of the profile confidence interval is less than or equal to a fixed threshold. In some cases we solve the optimization problem in closed form. For more complex models we use numerical optimization. We detail models assuming homogeneous, time-varying, subject-specific capture probabilities, behavioural response to capture and combining behavioural response with subject-specific effects. The principle we propose can be extended to plan any other model specification. We formally show the validity of the approach by proving distributional convergence. We illustrate with simulations and challenging examples in epidemiology and ecology. We report that in many cases adding as few as two sampling occasions may substantially reduce the length of confidence intervals.

Key words: Profile confidence interval; planning of experiments; population size estimation.

1 Introduction

Capture-recapture is concerned with estimation of the size \( N \) of a target population based on the capture histories of \( n \leq N \) individuals over \( K \) repeated occasions. Capture probabilities may depend on the specific capture occasion (\( M_t \) models, where \( t \) stands for time), previous capture occasions (\( M_b \) models), and unobserved heterogeneity (\( M_h \) models). These sources of variability/heterogeneity may be combined. See for instance Otis et al. (1978).

Prevalence estimation studies that are based on a random sample are often planned by pre-specifying a desired precision of the estimate, which can be summarized for instance with the expected length of a confidence interval. A necessary ingredient is a previous guess for the true prevalence, which can be obtained from similar studies on different areas/populations, pilot studies, or simply via expert knowledge. Surprisingly enough, prevalence estimation studies based on capture-recapture are seldom formally planned. When this happens, planning involves simulation (e.g., Chang et al. (1999); Devinau et al. (2006)) or it is based on benchmark data sets (Hay, 1997; Ismail et al., 2000). The only exception to this rule are given by methods in Lloyd and Chaiyapong (1999), which is restricted to model \( M_b \), and Xi et al. (2008). The paper by Xi et al. (2008) covers \( M_t \), \( M_b \) and \( M_h \), but is devoted mostly to planning the sampling fraction \( n/N \), which can not be fixed before the experiment. In fact, in capture-recapture studies the sample size \( n \) is random, and can not be fixed by design. Hence, with the Xi et al. (2008) method the number of capture occasions is also random at the planning stage and the researcher does not really know in advance the exact details of the study plan. A difference of our approach with respect to previous works is that these are focused on MSE in estimation of the true population size, while our approach is focused on a prescribed length of the confidence interval. Of course, the two quantities are strongly related. An underlying assumption (and limitation) of our general approach is that the user specifies the correct model, therefore making estimators asymptotically unbiased.

∗e-mail: danilo.alunnifegatelli@uniroma1.it
**Corresponding author: e-mail: alessio.farcomeni@uniroma1.it
The problem we are facing is rather simple in principle. The ideas can be fixed as follows: suppose we can obtain samples from a population of interest, whose size is unknown. Suppose additionally that sampling is repeated \( K > 1 \) times, and that at each occasion some individuals already observed may appear again. At each occasion we do not know how many subjects we will be able to detect, and how many will be new detections. We have control only on the number of sampling occasions. By specifying the modeling assumptions, expected population size and expected detection probabilities, we obtain below a measure of the expected length of the confidence interval for \( \hat{N} \), the estimate of the population size. For instance, if we assume homogeneity and independence of capture probabilities for each subject and at each occasion (model \( M_0 \), see below), that these probabilities correspond to 5\%, and that the true population size is \( N = 1000 \), our formulas predict that \( K = 10 \) occasions should yield a confidence interval of length 330.

In summary, we plan discrete-time capture-recapture experiments by fixing the smallest number of capture occasions \( K \) so that the expected length of the profile confidence interval is less than or equal to a fixed threshold. In some cases we solve the optimization problem in closed form. For more complex models we use numerical optimization, which still does not involve any simulation and is very fast.

It must be mentioned that for certain model specifications there is the possibility of likelihood failure, that is, that the MLE is unbounded. For these models we not only control the expected length of the confidence interval, but also the probability of failure (and set the number of occasions as the larger between the one that should lead to the desired length and the one guaranteeing low enough risk of failure). Of course, lack of failure is not guaranteed but can be achieved with arbitrarily high probability, provided the researcher is prepared to have enough sampling occasions.

In the following we illustrate our method in general, and then give explicit formulas conditionally on several relevant modeling assumptions. We provide an asymptotic theoretical justification of our approach, and illustrate with simulations and real data examples that, provided the assumptions are correct, we are able to accurately predict the length of the profile confidence interval even when the true population size is relatively small. Finally, we give throughout hints and guidelines, especially when discussing real data examples, on how to select modeling assumptions at the planning stage.

The rest of the paper is as follows: in Section 2 the general method based on the profile likelihood approach is described and then specified for each specific source of variability, that is, for different models hypothesized before the experiment. The validity of the method is formally shown in the Supporting Information available online. We illustrate with simulation studies and real data examples in Sections 3 and 4, respectively. Some conclusions and further developments are exposed in Section 5. R code with the methods developed in this article is available as Supplementary Material.

## 2 General method

Let us consider a closed population where the population size is fixed for all the experiment (no birth/death, no immigration/emigration). The most general way of representing a discrete time experiment is by considering a \( N \times K \) binary matrix \( X \) such that the generic element \( x_{ij} \) is equal to 1 if unit \( i \) is captured in the \( j \)-th occasion and 0 otherwise. Let \( T \) denote a (multidimensional) sufficient statistic and \( L(N, p; T) \) denote the likelihood, which is also a function of \( p \), a (multidimensional) parameter. Recognizing that the population size is a discrete parameter, we obtain the confidence interval using a profile likelihood approach (Evans et al., 1996). Let \( l(N; T) = \sup_p \log(L(N, p; T)) \). Let now \( N_1(T) \) and \( N_0(T) \) denote the limits of the confidence interval, as the solutions to the equation, in \( \hat{N} \), of \( 2(l(\hat{N}; T) - l(N; T)) = z_{\alpha/2}^2 \), where \( z_{\alpha/2} \) is the upper \( \alpha/2 \) quantile of a standard normal, and \( 1 - \alpha \) is the desired coverage of the confidence interval. The length of the confidence interval can then approximately be obtained as \( L_{CI}(N; T) = N_0(T) - N_1(T) \). The common practice in classical statistical design of experiments is to specify expected pre-experimental values of the parameters, call them \( N_0 \) and \( p_0 \), and replace \( T \) in \( L_{CI}(N; T) \) with its expected value, which is only a function of \( p_0 \) and \( N_0 \). Similarly, we propose to obtain \( L_{CI}(N_0; E[T]) \), and forecast the length.
of the confidence interval obtained as \(2(l(\hat{N}; E[T]) - l(N_0; E[T])) = z_{\alpha/2}^2\), as a function of \(\hat{N}\). Formally, we will evaluate \(L_{C1}(N_0; E[T])\) for different values of \(K\) and plan \(K\) so that \(L_{C1}(N_0; E[T])/N_0\) is small enough. Note that for ease of notation we suppress the dependence of the above expressions on \(K\), which is the target of our planning of experiment (and the only quantitative feature that can be set in advance).

It is worth noting a fundamental difference with the common (post data collection) pratice of data analysis: in that situation, confidence intervals are computed by substituting \(N\) with \(\hat{N}\) in the expression above. Here we have a pre-data guess for \(N\), as customary in planning of experiments.

In our experience the method is very accurate (as testified by the simulations below). The profile log-likelihood, conditionally on \(N\), can be obtained in closed form in many cases, some of which we describe in the following. When it is not available in closed form one can proceed through numerical maximization (e.g., in models \(M_{lb}\), \(M_{lb}\) and \(M_{lbh}\), or when occasion specific covariates are used to parameterize the detection probabilities). In the Supplementary Material we formally show that the profile likelihood can be approximated with a normal random variable as \(\hat{N}\) (and consequently \(n\)) grows. The latter provides a theoretical justification to the procedure proposed, as the asymptotic expected value of the profile likelihood (and of the length of the confidence interval) will correspond to what is obtained after plug-in of the expected sufficient statistic \(E[T]\). A consequence is that \(L_{C1}(N; E[T])\), which we have used so far to plan the number of occasions \(K\), will asymptotically correspond to \(E[L_{C1}(N; T)]\) as desired. These results could also be used to obtain a tighter control of the sampling design by requiring at least \(\eta\)% of the experiments yield a confidence interval of at most the desired length. To do so one could use the normal approximation and require that the \(\eta\) quantile of the asymptotic distribution of \(L_{C1}(N; X)/N\) is as close as possible to the desired relative length. Note that in \(\sigma_L\), the gradient of \(L_{C1}(N; T)\) would have to be evaluated numerically. We have anyway seen that the standard deviation of the confidence interval length is often rather small, and given that \(K\) is a discrete parameter, control of the mean expected length or of quantiles of the expected length of the confidence intervals almost always lead to the same experimental plan. This happens even for relatively small \(N\) and/or small capture probabilities.

### 2.1 \(M_0\) model

In the context of closed capture-recapture analyses, when no sources of variability are considered we have the simplest model \(M_0\). It is based only on two parameters, \(N\) and \(p = Pr(X_{ij} = 1)\). The sufficient statistic is \(T_{M_0} = (n, \sum_{ij} X_{ij})\), the number of units seen at least once and the total number of captures regardless of subject and occasion. The likelihood function for model \(M_0\) is

\[
L(N, p; T_{M_0}) = \left(\frac{N}{n}\right) p^{\sum_{ij} X_{ij}} (1 - p)^{KN - \sum_{ij} X_{ij}}.
\]

Obviously, we cannot know in advance the values for the sufficient statistic \(T_{M_0}\). Our idea is to substitute the expected values of such quantities, which depend on the pre-specified \(N_0\) and \(p_0\). These will be specified according to the researcher’s expectation to yield a plan for the number of occasions \(K\). For \(M_0\) we have that \(E[\sum_{ij} X_{ij}] = K N_0 p_0\) and \(E[n] = N_0 \Pr(\sum_{ij} X_{ij} > 0) = N_0 P_0\), where \(P_0 = 1 - (1 - p_0)^K\) is the probability of being captured at least once. Indeed, \(n\) is a binomial random variable with parameters \(N_0\) and \(P_0\). In order to work with the profile likelihood, we simply note that for each \(\hat{N}\) the profile log-likelihood is obtained by substituting the true value of the nuisance parameter \(p\) with \(\hat{p}_{M_0} = \sum_{ij} X_{ij}/NK\). After substitution of \(E[T_{M_0}] = K N_0 p_0\), \(\hat{p}_{M_0} = E[\sum_{ij} X_{ij}]/N_0 K = p_0\). Hence, after substitution and passing to logarithms we have that the profile log-likelihood function is

\[
\begin{align*}
l(\hat{N}; E[T_{M_0}]) &= \log \Gamma(\hat{N} + 1) - \log \Gamma \left(\left[N_0(1 - (1 - p_0)^K)\right] + 1\right) \\
&\quad - \log \Gamma \left(\left[\hat{N} - N_0(1 - (1 - p_0)^K)\right] + 1\right) + K N_0 p_0 \log(p_0) \\
&\quad + (K\hat{N} - K N_0 p_0) \log(1 - p_0)
\end{align*}
\]
One can now find the expected length of the confidence interval numerically for each possible value of $K$ as the difference between the two solutions of $2(l(\tilde{N}; E[T_{M_0}]) - l(N_0; E[T_{M_0}])) = z^2_{\alpha/2}$. A crucial point is that in computing the confidence interval $E[\sum_{ij} X_{ij}]$ is fixed throughout (that is, it is not a function of $\tilde{N}$ but only of $N_0$ and $p_0$ as discussed at the beginning of this section). The expected confidence interval lengths decrease monotonically as $N$, $p$ and $S$ increase (see Figure 1).

**Figure 1**  Model $M_0$: Expected relative confidence interval length ($\hat{\mu}_L/N$) for different values of $N$, $S$ and $p$

### 2.2 $M_b$ model

In the standard behavioural model $M_b$ the capture probability at the $j$-th occasion depends on the previous, partial, capture history. A well known issue with model $M_b$ is the so called likelihood failure, that is, under model $M_b$ the likelihood may become unbounded as a function of $N$ (Lloyd and Chaiyapong, 1999; Alunni Fegatelli and Tardella, 2013). In model $M_b$ we have two capture probabilities. The parameter $p = \Pr(X_{ij} = 1|\sum_{l=1}^{j-1} x_{il} = 0)$ represents the initial capture probability and $r = \Pr(X_{ij} = 1|\sum_{l=1}^{j-1} x_{il} > 0)$ is the recapture probability. The sufficient statistic is $T_{M_b} = (n, n_{0p}, n_{0r}, n_{1r})$, where

\[
\begin{align*}
n_{0p} &= \sum_{i=1}^{n} \sum_{j=1}^{K} I(\sum_{l=1}^{j} x_{il} = 0) \\
n_{0r} &= \sum_{i=1}^{n} \sum_{j=1}^{K} I(\sum_{l=1}^{j-1} x_{il} > 0, x_{ij} = 0) \\
n_{1r} &= \sum_{i=1}^{n} \sum_{j=1}^{K} I(\sum_{l=1}^{j-1} x_{il} > 0, x_{ij} = 1),
\end{align*}
\]

and $I(\cdot)$ denotes the indicator function. In words, $n_{0p}$ is the number of times that observed units never captured before are not captured at the current occasion. Similarly, $n_{0r}$ and $n_{1r}$ are the number of times
that observed units with at least one capture are not captured and captured, respectively, at the current occasion. Likelihood function for model $M_b$ can be expressed as follows

$$L(N, p, r; T_{M_b}) = \left[\frac{N}{n}\right] p^n (1 - p)^{K(N - n) + n_0r} \times [p^{n(1 - r)n_0r}] .$$  \hspace{1cm} (2)

It is straightforward to check that only the initial capture probability $p$ is involved in the estimate of the population size $N$, and only the first factor of (2) is involved in the computation of the profile confidence interval of $N$. The profile likelihood is obtained after plug-in of $\hat{p} = n/(n + n_0p + K(N - n))$. As in model $M_0$, we have that the probability of never being captured is $(1 - p)^K$. Hence, we have the same expected value for $n$, while $n_0p$ conditionally on $n > 0$ is a sum of $n$ truncated geometric random variables with parameters $p$ and $K$, and support $\{1, 2, \ldots, K\}$. We therefore have that

$$E[n_0p] = E[E(n_0p|n > 0)] = E\left[\sum_{j=0}^{K-1} (1 - p_0)^j \frac{1}{1 - (1 - p_0)^K}\right] = N_0p_0 \left(\sum_{j=0}^{K-1} j(1 - p_0)^j\right).$$

After plug-in we have that the profile log-likelihood function for model $M_b$ corresponds, up to an additive constant, to

$$\log \Gamma(\hat{N} + 1) - \log \Gamma\left([N_0(1 - (1 - p_0)^K)] + 1\right) - \log \Gamma\left(\hat{N} - N_0(1 - (1 - p_0)^K)\right) + 1$$

$$+ N_0 \left[1 - (1 - \hat{p}_{M_b})^K\right] \log(\hat{p}_{M_b})$$

$$+ \left[K\hat{N} - KN_0(1 - (1 - p_0))^K + N_0p_0 \sum_{j=0}^{K-1} j(1 - p_0)^j\right] \log(1 - \hat{p}_{M_b})$$

where

$$\hat{p}_{M_b} = \frac{1 - (1 - p_0)^K}{1 + p_0 \sum_{j=0}^{K-1} j(1 - p_0)^j + (K - 1)(1 - p_0)^K}.$$

As before, the profile confidence interval can now be obtained numerically for each $K$.

### 2.2.1 Likelihood failure

The likelihood failure phenomenon, which consists in an unbounded estimate of $N$, was studied for the first time in Chapman (1951). See also Seber and Whale (1970). It is a characteristic of removal models, like $M_b$, but can occur also for other models. We consider $M_0$ below. Failure conditions were proposed also for the profile likelihood approach in Carle and Strub (1978) and reviewed in a more general class of behavioural models in Alunni Fegatelli and Tardella (2013).

The failure condition for the observed data provided in Carle and Strub (1978) is

$$R = \frac{K - 1}{2} (n + 1) - (n_0p - 1) < 0.$$  \hspace{1cm} (3)

The first and second moments of $R$ can be found as Lloyd and Chaiyapong (1999) as

$$E[R] = \frac{K - 1}{2} (E[n] + 1) - (E[n_0p] - 1)$$

$$\text{Var}[R] = \left(\frac{K - 1}{2}\right)^2 \text{Var}[n] + \text{Var}[n_0p] - (K - 1)\text{Cov}[n, n_0p]$$

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

www.biometrical-journal.com
where
\[
\text{Var}[n] = N_0(1 - (1 - p_0)^K)(1 - p_0)^K \\
\text{Var}[n_{00}] = \text{Var}[X|E[n]] + E^2[X|\text{Var}[n]] \\
\text{Cov}[n, n_{00}] = (E^2[X|\text{Var}[N]] E[X] - E[n]E[n_{00}] ,
\]
where \( X \) is a geometric random variable with parameter \( p_0 \), truncated at \( K - 1 \). The probability of likelihood failure can be estimated using a normal approximation as
\[
\Pr(\text{Likelihood failure}|M_b) = \Phi \left( \frac{-E(R)}{\sqrt{\text{Var}(R)}} \right) \tag{5}
\]
where \( \Phi(.) \) is the cumulative distribution function of the standard normal distribution. The normality assumption is only an approximation, and might not be good for certain model specifications and small sample sizes. We recommend routinely checking the assumption, for instance via simulations as we outline in Section 4.2. For the \( M_b \) model we not only evaluate \( L_{CL}(N_0; E[T_{M_b}]) \) for each \( K \geq 1 \), but also (5).

For model \( M_0 \), likelihood failure occurs when there are no recaptures \( (\sum_{i=1}^{K} X_{ij} \leq 1 \forall i = 1, ..., N) \). The probability of this event can be estimated as
\[
\Pr(\text{Likelihood failure}|M_0) = \left[ (1 - p_0)^K + p_0(1 - p_0)^{K-1} \right]^{N_0},
\]
which can be seen to be often negligible.

2.3 \( M_t \) model

In the \( M_t \) model time-heterogeneity is considered. For \( j = 1, \ldots, K \) we have \( \Pr(X_{ij} = 1) = p_j \quad \forall i = 1, \ldots, N \). The sufficient statistic is \( T_{M_t} = (n, n_K) \) where \( n_K = (n_1, \ldots, n_K) \) and \( n_j = \sum_{i=1}^{N} x_{ij} \) is the number of units captured at the \( j \)-th occasion. The likelihood function is
\[
\log L(N, p_1, \ldots, p_K; T_{M_t}) = \binom{N}{n} \prod_{j=1}^{K} p_j^{n_j} (1 - p_j)^{N-n_j} \tag{6}
\]

Call \( p_{j0}, j = 1, \ldots, K \), the pre-specified parameters at the design stage. The expected value of \( n \) is
\[
E[n] = N_0 \left( 1 - \prod_{j=1}^{K} (1 - p_{j0}) \right),
\]
while the expected value of the generic \( j \)-th element of \( n_K \) is \( E[n_j] = N_0 p_{j0} \). It can also be shown that the profile likelihood is simply obtained by plug-in of \( \hat{p}_j = n_j/N \). Once we plug-in the expected values into the profile likelihood we have the following expression:
\[
\log \Gamma(\hat{N} + 1) - \log \Gamma \left( N_0 \left( 1 - \prod_{j=1}^{K} (1 - p_{j0}) \right) + 1 \right) - \log \Gamma \left( \hat{N} - N_0 \left( 1 - \prod_{j=1}^{K} (1 - p_{j0}) \right) + 1 \right)
\]
\[
+ \sum_{j=1}^{K} \left[ (N_0 p_{j0}) \log(p_{j0}) + (\hat{N} - N_0 p_{j0}) \log(1 - p_{j0}) \right]. \tag{7}
\]
A peculiar situation in planning under the \( M_t \) model is that the number of parameters and the dimensionality of the sufficient statistic increase with \( K \). This may be cumbersome at the planning stage. We in general have to pre-specify a long vector of nuisance parameters (as many as the largest \( K \) considered).
In case the general approach is too cumbersome, there are two separate routes we can suggest to simplify planning of model $M_t$.

One possibility is to specify capture probabilities by assuming $p_{j0}$ is a function of $j$, e.g., $\logit(p_{j0}) = \alpha + j\beta$ for some fixed $\alpha$ and $\beta$. This leads to a general rule for $p_{j0}$ and does not affect the strategy above. Another possibility is to specify a working a priori distribution for $p_{j0}$, and results obtained after marginalization. We are not assuming $p_{j0}$ is random, but only specifying $p_{j0}$ before the experiment in a convenient way. For instance, one may set

$$\logit(p_{j0}) \sim N(p_0, \tau_0)$$

for some $p_0$ and $\tau_0$, with $p_{j0}$ is independent of $p_{h0}$. Independence can be easily relaxed using for instance an auto-regressive structure if appropriate. A simple Monte Carlo strategy, Laplace approximations, or numerical quadrature can be used to integrate out $p_0$.

### 2.4 Occasion specific covariates

We now briefly mention the case of occasion specific covariates, which is closely linked to model $M_t$. If we let $Z_j$ denote a vector of covariates specific to the occasion (and known in advance), a popular modeling is

$$\logit\left(\frac{p_{j0}}{1 - p_{j0}}\right) = \alpha + Z_j\beta,$$

and consequently

$$p_{j0} = \frac{\alpha + Z_j\beta}{1 + \exp\{\alpha + Z_j\beta\}}$$

The expression above shall be substituted in expression (6). The case of occasion specific covariates is a good example of a situation where a closed form expression is not available for the profile likelihood. The profile likelihood shall be obtained using any general purpose numerical maximizer.

### 2.5 $M_h$ model

In model $M_h$, unobserved heterogeneity is assumed to lead to subject-specific capture probabilities $\Pr(X_{ij} = 1) = p_i$, with $p_i \sim F(\alpha)$. The sufficient statistic depends on the actual functional form for the parametric distribution $F(\cdot)$. Common assumptions include point masses, Beta distribution, or Gaussian assumptions on the logit scale. In all cases, the likelihood for model $M_h$ is

$$L(N, \alpha; X) = \prod_{i=1}^{N} \int p^{\sum_j X_{ij}} (1 - p)^{K - \sum_j X_{ij}} dF(\alpha).$$

Considering all possible unit label arrangements that give rise to the same capture counts, the same can be expressed as

$$L(N, \alpha; T_{M_h}) = \binom{N}{n} P_{0,F}^{N-n} \prod_{k=1}^{K} P_{k,F}^{f_k},$$

where $T_{M_h} = (n, f_1, \ldots, f_K)$ is the sufficient statistic for model $M_h$ with the generic element $f_k$ representing the number of rows of $X$ such that $\sum_j X_{ij} = k$ and

$$P_{k,F} = \binom{K}{k} \int p^k (1 - p)^{K-k} dF(\alpha).$$

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

www.biometrical-journal.com
Table 1 Some mixing distributions for model $M_h$ yielding closed form expressions for $P_{k,F}$.

<table>
<thead>
<tr>
<th>Model</th>
<th>$P_{k,F}$ ∝</th>
</tr>
</thead>
<tbody>
<tr>
<td>binomial mixture</td>
<td>$\sum_{j=1}^f g_j \pi^j (1-\pi)^{K-k}$ with $\sum_j g_j = 1$</td>
</tr>
<tr>
<td>beta-binomial</td>
<td>$\prod_{i=0}^{\alpha_1+j\alpha_2} \prod_{i=0}^{\alpha_1+k\alpha_2}$</td>
</tr>
<tr>
<td>binomial-beta-binomial mixture</td>
<td>$g_1 \pi^k (1-\pi)^{K-k} + (1-g_1) \prod_{j=0}^{\alpha_1+j\alpha_2} \prod_{j=0}^{\alpha_1+k\alpha_2}$</td>
</tr>
</tbody>
</table>

It is clear from the expression above that the sufficient statistics are given by $n$ and $f_1, \ldots, f_K$. The user must here pre-specify $N_0$ and $\alpha_0$. Straightforward algebra gives

$$E[n] = \sum_i \Pr \left( \sum j X_{ij} > 0 \right) = N_0 \left( 1 - \int (1-p)^K dF(\alpha_0) \right) = N_0 (1 - P_{0,F,0}), \quad (11)$$

and

$$E[f_k] = N_0 P_{k,F,0}, \quad (12)$$

where $P_{k,F,0}$ corresponds to (10) after plug-in. The expressions above and the expected values of the quantities involved as a function of $X$ can be easily computed through numerical quadrature. It shall be noted that the integrals involved are unidimensional. As soon as $F(\alpha)$ is such that the integrals involved in (10) can be solved, closed form expressions will be available for the likelihood and the expected sufficient statistics. For instance, if $F(\alpha)$ is a Beta distribution with parameters $\alpha_1$ and $\alpha_2$,

$$E[n] = \frac{\Gamma(\alpha_{10} + \alpha_{20}) \Gamma(\alpha_{20} + K)}{\Gamma(\alpha_{10} \alpha_{20} + K) \Gamma(\alpha_{20})}.$$

Additionally,

$$P_{k,F,0} = \binom{K}{k} \frac{\Gamma(\alpha_{10} + \alpha_{20}) \Gamma(\alpha_{10} + k) \Gamma(\alpha_{20} + K - k)}{\Gamma(\alpha_{10}) \Gamma(\alpha_{20}) \Gamma(\alpha_{10} + \alpha_{20} + K)}.$$

More common choices include the logistic-normal binomial model, and finite and infinite mixture models, including binomial mixtures, beta-binomial models, point masses. See Coull and Agresti (1999); Pledger (2000); Dorazio and Royle (2003); Morgan and Ridout (2008). With the exception of the first case, closed form expressions are available for $P_{k,F,0}$, making planning under $M_h$ straightforward in all these cases. For reference, we list them in Table 1. The values of $P_{k,F,0}$ can then be substituted in (11) and (12). When $F(\alpha)$ on the other hand is a Gaussian on the log-scale (logistic-normal model), the only viable option is to perform numerical integration, e.g., via quadrature. See also Pledger and Phillpot (2008).

It shall be noted that, unless one wants to assume $\alpha$ is known in the analysis (which is unusual), the profile likelihood must be obtained by maximizing (9) numerically. This can be accomplished using a general purpose maximizer, while other strategies (e.g., the EM algorithm) might be more appropriate in some cases (e.g., when point masses are assumed for the mixing distribution).

2.6 $M_{bh}$ model

We now combine the models considered previously by allowing capture probabilities depend on whether the subject was ever captured, and on unobserved factors. For the resulting $M_{bh}$ model capture and recapture probabilities are expressed as $\Pr(X_{ij} = 1 | \sum_{s=1}^{u} X_{is} = 0) = p_i$; and $\Pr(X_{ij} = 1 | \sum_{s=1}^{u} X_{is} > 0) = r_i$, respectively, where $r_i \sim F_r(\alpha_r)$ and $p_i \sim F_p(\alpha_p)$. The recapture part of the likelihood can be ignored when computing profile confidence intervals. The sufficient statistic is given by $T_{M_{bh}} = (n, u_1, \ldots, u_K)$,
where \( u_j \) denotes the number of unmarked subjects observed at occasion \( j \). It shall be underlined that for planning purposes, models \( M_{bh} \) and \( M_b \) (similarly to models \( M_0 \) and \( M_b \)) are rather similar. Additionally, the same latent probability distributions considered for \( M_b \) can be considered for \( M_{bh} \). The relevant part of the likelihood is

\[
L(N, \alpha; T_{M_{bh}}) \propto \left( \frac{N}{n} \right) P_{0,F}^{N-n} \prod_{j=1}^{K} \tilde{P}^{u_j}_{j,F}
\]

where \( P_{0,F} \) is as in (11) and \( \tilde{P}_{j,F} = \int (1-p)^{j-1} p \, dF_p \). It can be seen that \( E[\mu] \) has also the same expression as (11), while \( E[\nu_j] = N_0 \tilde{P}_{j,F,0} \) with obvious notation. As in model \( M_b \) the log profile likelihood must be obtained numerically by maximizing the logarithm of \( L(N, \alpha; E[T_{M_{bh}}]) \). The same algebra used for model \( M_b \) leads to a very similar expression. The expected length of the confidence interval coincides with the one that could be obtained for an \( M_b \) model based on \( F_p(\alpha_{p0}) \). The other moments of the distribution do not coincide, but this is not an issue for our purposes (see also the Supporting Information).

3 Simulations

We carried out a simulation study to verify the validity of the proposed procedure (based on the profile likelihood approach) for predicting the expected length of the confidence interval for \( N \). For each model \( (M_0, M_b, M_b, \text{ and } M_{bh}) \), twelve alternative scenarios are considered, based on three values for \( N \) (100, 200 and 1000), and four different settings for the nuisance parameters.

For model \( M_0 \), the capture probability \( p \) was set to 0.05 (trials 1,2,3), 0.1 (trials 4,5,6), 0.2 (trials 7,8,9), and 0.4 (trials 10,11,12). The same values were used for the first capture probability \( p_1 \) of model \( M_b \) in trials 13-24 while the recapture probability \( r \) (which does not affect the estimate of the population size) was set to 0.2, 0.3, 0.4 and 0.2, respectively. For model \( M_b \) capture probabilities at each occasion were equally spaced in \([p_1, p_n] \) (as reported in Table 2), and permuted randomly at each iteration. The beta-binomial distribution was considered for the \( M_b \) model, based on parameters \( \alpha_1 \) and \( \alpha_2 \) for the beta mixing distribution. The values considered are also reported in Table 2. We treated \( \alpha_1 \) and \( \alpha_2 \) as known at the data collection. For each setting and for each \( K \in \{2,...,10\} \) 1000 data sets were generated and the average confidence interval length \( L_{CI}/N \) was computed. We compare it with our predicted interval length \( \hat{\mu}_L/N \).

The results for the fortyeight trials summarized in Table 2 are reported in Tables 3 and 4. Moreover, for trials 1-24, Table 5 reports the empirical failure proportions and the expected failure rate for each value of \( K \). It can be seen that in all cases the expected confidence interval length before seeing the data is very close to the empirical one. This happens regardless of the actual empirical length, as in our settings these vary wildly from almost 10\( N \) to less than 0.01\( N \). For the classical behavioural model \( M_b \), in accordance with a relevant likelihood failure rate, very wide confidence intervals are observed for small values of \( N \) (\( \leq 200 \)). Indeed, in these cases, even if the failure conditions are not satisfied, the flatness of the profile likelihood leads to a quasi-failure pathology. A careful planning of the experiment in presence of behavioural effects can substantially decrease the risks of likelihood failures and the length of the confidence intervals.

The risks of formal failure are reported in Table 5. It can be seen that our estimates work well also in this respect. It can also be seen that for model \( M_0 \) the failure rate is very high for small values of the capture probability \( p = 0.05 \) and population size \( (N = 100) \) but it decreases quickly as \( p \) and \( N \) increase. On the other hand, for model \( M_b \), when the probability of first capture \( p \) is low, the failure rate is not too high but it decrease less quickly than in model \( M_0 \) as \( p \) and \( N \) increase. Moreover the failure rate for model \( M_b \) does not decrease monotonically as \( S, p \) and \( N \) increase (see the surface plots in Figure 2 and in Figure 3).
Table 2  Parameter configurations for the 48 simulation experiments. For the $M_t$ model (trials 25-36) for each $K \in \{2, \ldots, 10\}$ we define the set of capture probabilities by a regular sequence of length $K$ where $p_l$ and $p_u$ represent the lower and upper limit respectively.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Model</th>
<th>$N$</th>
<th>Parameters</th>
<th>Trial</th>
<th>Model</th>
<th>$N$</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$M_0$</td>
<td>100</td>
<td>$p = 0.050$</td>
<td>13</td>
<td>$M_0$</td>
<td>100</td>
<td>$p = 0.050; r = 0.200$</td>
</tr>
<tr>
<td>2</td>
<td>$M_0$</td>
<td>200</td>
<td>$p = 0.050$</td>
<td>14</td>
<td>$M_0$</td>
<td>200</td>
<td>$p = 0.050; r = 0.200$</td>
</tr>
<tr>
<td>3</td>
<td>$M_0$</td>
<td>1000</td>
<td>$p = 0.050$</td>
<td>15</td>
<td>$M_0$</td>
<td>1000</td>
<td>$p = 0.050; r = 0.200$</td>
</tr>
<tr>
<td>4</td>
<td>$M_0$</td>
<td>100</td>
<td>$p = 0.100$</td>
<td>16</td>
<td>$M_0$</td>
<td>100</td>
<td>$p = 0.100; r = 0.300$</td>
</tr>
<tr>
<td>5</td>
<td>$M_0$</td>
<td>200</td>
<td>$p = 0.100$</td>
<td>17</td>
<td>$M_0$</td>
<td>200</td>
<td>$p = 0.100; r = 0.300$</td>
</tr>
<tr>
<td>6</td>
<td>$M_0$</td>
<td>1000</td>
<td>$p = 0.100$</td>
<td>18</td>
<td>$M_0$</td>
<td>1000</td>
<td>$p = 0.100; r = 0.300$</td>
</tr>
<tr>
<td>7</td>
<td>$M_0$</td>
<td>100</td>
<td>$p = 0.200$</td>
<td>19</td>
<td>$M_0$</td>
<td>100</td>
<td>$p = 0.200; r = 0.400$</td>
</tr>
<tr>
<td>8</td>
<td>$M_0$</td>
<td>200</td>
<td>$p = 0.200$</td>
<td>20</td>
<td>$M_0$</td>
<td>200</td>
<td>$p = 0.200; r = 0.400$</td>
</tr>
<tr>
<td>9</td>
<td>$M_0$</td>
<td>1000</td>
<td>$p = 0.200$</td>
<td>21</td>
<td>$M_0$</td>
<td>1000</td>
<td>$p = 0.200; r = 0.400$</td>
</tr>
<tr>
<td>10</td>
<td>$M_0$</td>
<td>100</td>
<td>$p = 0.400$</td>
<td>22</td>
<td>$M_0$</td>
<td>100</td>
<td>$p = 0.400; r = 0.200$</td>
</tr>
<tr>
<td>11</td>
<td>$M_0$</td>
<td>200</td>
<td>$p = 0.400$</td>
<td>23</td>
<td>$M_0$</td>
<td>200</td>
<td>$p = 0.400; r = 0.200$</td>
</tr>
<tr>
<td>12</td>
<td>$M_0$</td>
<td>1000</td>
<td>$p = 0.400$</td>
<td>24</td>
<td>$M_0$</td>
<td>1000</td>
<td>$p = 0.400; r = 0.200$</td>
</tr>
<tr>
<td>25</td>
<td>$M_1$</td>
<td>100</td>
<td>$p_l = 0.010; p_u = 0.090$</td>
<td>37</td>
<td>$M_2$</td>
<td>100</td>
<td>$\alpha_1 = 0.25; \alpha_2 = 4.75$</td>
</tr>
<tr>
<td>26</td>
<td>$M_1$</td>
<td>200</td>
<td>$p_l = 0.010; p_u = 0.090$</td>
<td>38</td>
<td>$M_2$</td>
<td>200</td>
<td>$\alpha_1 = 0.25; \alpha_2 = 4.75$</td>
</tr>
<tr>
<td>27</td>
<td>$M_1$</td>
<td>1000</td>
<td>$p_l = 0.010; p_u = 0.090$</td>
<td>39</td>
<td>$M_2$</td>
<td>1000</td>
<td>$\alpha_1 = 0.25; \alpha_2 = 4.75$</td>
</tr>
<tr>
<td>28</td>
<td>$M_1$</td>
<td>100</td>
<td>$p_l = 0.025; p_u = 0.175$</td>
<td>40</td>
<td>$M_2$</td>
<td>100</td>
<td>$\alpha_1 = 0.50; \alpha_2 = 4.50$</td>
</tr>
<tr>
<td>29</td>
<td>$M_1$</td>
<td>200</td>
<td>$p_l = 0.025; p_u = 0.175$</td>
<td>41</td>
<td>$M_2$</td>
<td>200</td>
<td>$\alpha_1 = 0.50; \alpha_2 = 4.50$</td>
</tr>
<tr>
<td>30</td>
<td>$M_1$</td>
<td>1000</td>
<td>$p_l = 0.025; p_u = 0.175$</td>
<td>42</td>
<td>$M_2$</td>
<td>1000</td>
<td>$\alpha_1 = 0.50; \alpha_2 = 4.50$</td>
</tr>
<tr>
<td>31</td>
<td>$M_1$</td>
<td>100</td>
<td>$p_l = 0.100; p_u = 0.300$</td>
<td>43</td>
<td>$M_2$</td>
<td>100</td>
<td>$\alpha_1 = 1.00; \alpha_2 = 4.00$</td>
</tr>
<tr>
<td>32</td>
<td>$M_1$</td>
<td>200</td>
<td>$p_l = 0.100; p_u = 0.300$</td>
<td>44</td>
<td>$M_2$</td>
<td>200</td>
<td>$\alpha_1 = 1.00; \alpha_2 = 4.00$</td>
</tr>
<tr>
<td>33</td>
<td>$M_1$</td>
<td>1000</td>
<td>$p_l = 0.100; p_u = 0.300$</td>
<td>45</td>
<td>$M_2$</td>
<td>1000</td>
<td>$\alpha_1 = 1.00; \alpha_2 = 4.00$</td>
</tr>
<tr>
<td>34</td>
<td>$M_1$</td>
<td>100</td>
<td>$p_l = 0.300; p_u = 0.500$</td>
<td>46</td>
<td>$M_2$</td>
<td>100</td>
<td>$\alpha_1 = 2.00; \alpha_2 = 3.00$</td>
</tr>
<tr>
<td>35</td>
<td>$M_1$</td>
<td>200</td>
<td>$p_l = 0.300; p_u = 0.500$</td>
<td>47</td>
<td>$M_2$</td>
<td>200</td>
<td>$\alpha_1 = 2.00; \alpha_2 = 3.00$</td>
</tr>
<tr>
<td>36</td>
<td>$M_1$</td>
<td>1000</td>
<td>$p_l = 0.300; p_u = 0.500$</td>
<td>48</td>
<td>$M_2$</td>
<td>1000</td>
<td>$\alpha_1 = 2.00; \alpha_2 = 3.00$</td>
</tr>
</tbody>
</table>

Figure 2  Model $M_0$: probability of likelihood failure for different values of $N$, $S$ and $p$

4  Examples

We revisit in this section some published datasets, in order to assess how well the number of occasions were originally calibrated and how the sampling design could be possibly improved. We then conclude the section by planning a new experiment.
Table 3  Trial 1-24: Expected value ($\hat{\mu}_L/N$) and empirical median ($L_{CI}/N$) of the relative confidence intervals length for different values of $K$ and different settings as described in Table 2. Results are averaged over 1000 replications.

<table>
<thead>
<tr>
<th>$K$</th>
<th>$L_{CI}/N$</th>
<th>$\hat{\mu}_L/N$</th>
<th>$L_{CI}/N$</th>
<th>$\hat{\mu}_L/N$</th>
<th>$L_{CI}/N$</th>
<th>$\hat{\mu}_L/N$</th>
<th>$L_{CI}/N$</th>
<th>$\hat{\mu}_L/N$</th>
<th>$L_{CI}/N$</th>
<th>$\hat{\mu}_L/N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.43</td>
<td>9.88</td>
<td>6.21</td>
<td>9.82</td>
<td>4.90</td>
<td>4.69</td>
<td>8.41</td>
<td>9.71</td>
<td>4.59</td>
<td>4.66</td>
</tr>
<tr>
<td>3</td>
<td>7.46</td>
<td>9.77</td>
<td>5.01</td>
<td>7.68</td>
<td>1.65</td>
<td>1.61</td>
<td>3.08</td>
<td>3.10</td>
<td>1.70</td>
<td>1.76</td>
</tr>
<tr>
<td>4</td>
<td>5.03</td>
<td>7.61</td>
<td>3.44</td>
<td>3.33</td>
<td>1.06</td>
<td>1.03</td>
<td>1.77</td>
<td>1.72</td>
<td>1.10</td>
<td>1.09</td>
</tr>
<tr>
<td>5</td>
<td>4.03</td>
<td>3.97</td>
<td>2.17</td>
<td>2.13</td>
<td>0.76</td>
<td>0.77</td>
<td>1.20</td>
<td>1.20</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>6</td>
<td>2.85</td>
<td>2.67</td>
<td>1.60</td>
<td>1.57</td>
<td>0.62</td>
<td>0.61</td>
<td>0.92</td>
<td>0.92</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>7</td>
<td>2.08</td>
<td>2.01</td>
<td>1.28</td>
<td>1.25</td>
<td>0.51</td>
<td>0.50</td>
<td>0.74</td>
<td>0.74</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>8</td>
<td>1.69</td>
<td>1.62</td>
<td>1.09</td>
<td>1.04</td>
<td>0.43</td>
<td>0.43</td>
<td>0.61</td>
<td>0.62</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>9</td>
<td>1.34</td>
<td>1.35</td>
<td>0.89</td>
<td>0.89</td>
<td>0.37</td>
<td>0.37</td>
<td>0.51</td>
<td>0.53</td>
<td>0.36</td>
<td>0.37</td>
</tr>
<tr>
<td>10</td>
<td>1.19</td>
<td>1.16</td>
<td>0.79</td>
<td>0.77</td>
<td>0.33</td>
<td>0.33</td>
<td>0.46</td>
<td>0.46</td>
<td>0.32</td>
<td>0.32</td>
</tr>
</tbody>
</table>

4.1 Injecting drug users in Canada

A simple example in Xu et al. (2014) regards a two-list experiment for estimation of the number of injecting drug users in Victoria, Canada. Two lists were gathered by screening consenting participants in the needle exchange programme run by AIDS Vancouver Island and at shelter services run by the Victoria Cool Aid Society. The screening was repeated twice at months distance. The first screening yielded 254 subjects, the second 250; with an overlap of 19 subjects. The model selected by information criteria is $M_9$, yielding $\hat{N} = 3330$, with a 95% CI corresponding to $(2278, 5256)$. The resulting relative length is 0.894, with $\hat{p} = 0.0757$.

If the planning phase we had specified $N = 3330$, $K = 2$ and $p = 0.075$, using our approach based on profile likelihood confidence intervals we could expect a relative length of 0.889, remarkably close to the truly observed one. Note that in some cases previous hypotheses of $N$ might be substituted with assumptions on $n$. For model $M_0$, for instance, if with $K = 2$ we assumed $n = 480$ we would then expect,
by inversion of the formulas for the expected value of $n$, $N = 3325$, leading to similar results. This is not very practical in general, though, as the expected $n$ should, of course, change with the number of sampling occasions.

Table 6 shows the expected relative length of the confidence interval for different values of $K$. Accordingly, by including an additional list the relative length would have decreased to 49%, and with two additional lists it would have been one third of the observed one.

An open issue with our planning strategy is that, similarly to all other design approaches, one must specify a model and model parameters in advance. In our context, an idea of the size of the population is often available, while the most appropriate model not necessarily is. For these data, we could rule out $M_h$ in advance as screening is performed at the centers. Hence it is unlikely that subject’s characteristics might influence the detection probability. For the same reason we rule out model $M_k$. For model $M_I$ to hold there should have been differences in the access to the needle exchange programmes or shelters between the two sampling occasions. This was known in advance not to hold.

### Table 4

<table>
<thead>
<tr>
<th>Trial 25</th>
<th>Trial 26</th>
<th>Trial 27</th>
<th>Trial 28</th>
<th>Trial 29</th>
<th>Trial 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>$L_{C1}/N$</td>
<td>$\hat{\mu}_L/N$</td>
<td>$L_{C2}/N$</td>
<td>$\hat{\mu}_L/N$</td>
<td>$L_{C3}/N$</td>
</tr>
<tr>
<td>4</td>
<td>9.79</td>
<td>9.79</td>
<td>9.79</td>
<td>9.77</td>
<td>9.87</td>
</tr>
<tr>
<td>5</td>
<td>9.61</td>
<td>9.06</td>
<td>4.42</td>
<td>3.73</td>
<td>1.10</td>
</tr>
<tr>
<td>6</td>
<td>4.62</td>
<td>4.29</td>
<td>2.42</td>
<td>2.24</td>
<td>0.78</td>
</tr>
<tr>
<td>7</td>
<td>2.99</td>
<td>2.78</td>
<td>1.60</td>
<td>1.62</td>
<td>0.62</td>
</tr>
<tr>
<td>8</td>
<td>2.15</td>
<td>2.06</td>
<td>1.25</td>
<td>1.27</td>
<td>0.51</td>
</tr>
<tr>
<td>9</td>
<td>1.30</td>
<td>1.37</td>
<td>0.89</td>
<td>0.90</td>
<td>0.37</td>
</tr>
<tr>
<td>10</td>
<td>1.21</td>
<td>1.17</td>
<td>0.79</td>
<td>0.78</td>
<td>0.33</td>
</tr>
</tbody>
</table>

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
www.biometrical-journal.com

Alunni Fegatelli, D. and Farcomeni, A.: Design of recapture experiments.
Figure 3  Model $M_0$: probability of likelihood failure for different values of $N$, $S$ and $p$

Table 5  Likelihood failure rate occurred in 1000 replicates under models $M_0$ and $M_b$. Values inside the brackets are the estimates of the failure rate.

<table>
<thead>
<tr>
<th>Model $M_0$</th>
<th>$p = 0.05$</th>
<th>$p = 0.10$</th>
<th>$p = 0.20$</th>
<th>$p = 0.40$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N=100$</td>
<td>$N=200$</td>
<td>$N=1000$</td>
<td>$N=100$</td>
<td>$N=200$</td>
</tr>
<tr>
<td>2</td>
<td>17.6 (14.9)</td>
<td>25.9 (13.4)</td>
<td>29.9 (28.9)</td>
<td>22.2 (17.9)</td>
</tr>
<tr>
<td>3</td>
<td>26.5 (21.0)</td>
<td>27.1 (24.8)</td>
<td>26.0 (24.0)</td>
<td>22.1 (17.9)</td>
</tr>
<tr>
<td>4</td>
<td>25.6 (22.2)</td>
<td>25.9 (23.7)</td>
<td>16.2 (17.2)</td>
<td>15.7 (14.2)</td>
</tr>
<tr>
<td>5</td>
<td>23.3 (21.5)</td>
<td>22.4 (21.1)</td>
<td>11.6 (10.9)</td>
<td>12.8 (10.0)</td>
</tr>
<tr>
<td>6</td>
<td>20.5 (19.8)</td>
<td>16.2 (17.9)</td>
<td>5.8 (6.1)</td>
<td>6.3 (6.4)</td>
</tr>
<tr>
<td>7</td>
<td>17.2 (17.7)</td>
<td>16.8 (14.6)</td>
<td>3.5 (2.9)</td>
<td>3.7 (3.7)</td>
</tr>
<tr>
<td>8</td>
<td>14.7 (15.4)</td>
<td>10.6 (11.5)</td>
<td>1.1 (1.2)</td>
<td>1.8 (1.9)</td>
</tr>
<tr>
<td>9</td>
<td>13.1 (13.1)</td>
<td>9.9 (8.7)</td>
<td>0.5 (0.4)</td>
<td>1.5 (0.9)</td>
</tr>
<tr>
<td>10</td>
<td>9.8 (10.9)</td>
<td>6.5 (6.3)</td>
<td>0.6 (0.1)</td>
<td>0.1 (0.4)</td>
</tr>
</tbody>
</table>

4.2 Deer Mouse Data

Deer Mouse data were considered in Lloyd and Chaiyapong (1999) as an example of $M_b$ design. It is based on $K = 10$ sampling occasions, with $n = 69$. The population size estimate is $\hat{N} = 112$ and the estimated probability of first capture $0.9$. The 95% profile CI is different from the one estimated, with a different approach, in Lloyd and Chaiyapong (1999). With the profile likelihood we estimate a 95% CI $(81, 600)$, with corresponding relative confidence interval length of 4.63 (quasi-failure pathology). These values
are not totally in agreement with what we would predict before the experiment for $K = 10$, assuming $N = 112$ and $p = .09 (\hat{\mu}_L / \hat{N} = 2.87)$. The difference between observed and expected confidence interval length is due to the high variability in the estimates in correspondence of small values of $N$ and $p$ (near flat profile likelihood). Moreover, as shown in Figure 4, the distributions of the CI lengths are asymmetric in correspondence of several values of $K$. Figure 4 is obtained by simulating data based on the planned parameters. The normal approximation for these distributions is not accurate for $K$ small, and this might lead to a bias and discrepancy between planned and actually estimated confidence interval length. For small values of $N$ and $p$ we always recommend to check the normal approximation via simulations, as in Figure 4. However, although different, both the predicted and the really observed CI are too large. In our experience, there seldom is disagreement when the likelihood is stable and the predicted CI length is relatively narrow. For these data we believe it might be desirable therefore to obtain a narrower CI by increasing $K$. We estimate that in order to overcome the near failure a good choice is $K = 18$. For $K \geq 18$, the expected relative confidence interval length is $< 0.5$. Additionally, as shown in Figure 4, for $K \geq 18$ we not only have a substantial reduction of the variability (most of the observed CI lengths are very close to the expected ones), but also better approximation to the Gaussian distribution for the CI lengths. Lloyd and Chaiyapong (1999) conclude that including an additional 9 unseen animals (that is, raising $n$ to 78) would lead to a substantial improvement of the relative squared error. Assuming $N = \hat{N}$ and $p = \hat{p}$ we can obtain an expected value for $n$ greater than 78 by setting $K = 13$. When $K = 13$ the expected confidence interval length is indeed improved, as suggested by Lloyd and Chaiyapong (1999), but still too large ($\hat{\mu}_L / N = 0.98$) in our opinion.

### 4.3 Amyotrophic Lateral Sclerosis among Gulf War veterans

Coffman et al. (2005) use four independent lists of cases of amyotrophic lateral sclerosis (ALS) to investigate its incidence in the U.S. military. This is an important issue as incidence of ALS among Gulf war (1990-1991) veterans was reported to be doubled. Capture-recapture is used to evaluate and correct for ascertainment bias, which might be differential leading to biased relative risk estimates. Incidentally, after data collection the conclusions are that even after adjusting for a mild but differential ascertainment bias, the 10 year incidence of ALS among Gulf war veterans is indeed increased. Four sources were used: the veterans affairs database, the department of defense database, the database of the ALS patients association, and screening of calls to a toll-free telephone number that individuals could call if they believed were eligible for the study. The toll-free number was publicized in military and non-military media. In order to choose a model for planning, we note that behavioural response can be certainly ruled out, and that differential probability of detection among the lists might be very likely present. This leaves us with two possible models, $M_t$ and $M_{th}$.

We focus here on model $M_t$, and the overall number of cases (both among Gulf war veterans and non-deployed militaries). It is likely that only a fraction of affected subjects joined the ALS association, which therefore should have the smallest detection probability, followed by the phone interviews. The veterans and defense databased should have the highest detection probabilities. As a matter of fact, after data collection we can estimate these probabilities to be 0.23, 0.47, 0.67 and 0.58, respectively; $\hat{N} = 113$, with 95% CI (108 – 120) and therefore a relative length of 11%. This is exactly the expected relative length before data collection, when specifying the probabilities and population size above.
Figure 4  Empirical distribution of $L_{C1}/N$ for different numbers of capture occasions, based on the parameters estimated for the Deer Mouse data. For each value of $K$ 1000 datasets were generated from model $M_b$ considering $N = 112$ and $p = 0.09$. The black triangles represent the expected relative confidence interval length $\mu_L/N$. 
For these data we can claim that the expensive and time-consuming effort of setting up and advertising a toll-free phone has lead to a decrease of about 6% of the confidence interval relative length. It would have been 17% without the phone list. The use of three lists excluding the ALS patients association database would have lead to a relative length of 13%, instead.

4.4 Chipmunk Data

Another example from Otis et al. (1978) regards chipmunks. An $M_t$ model was suggested for this experiment, which is based on $K = 6$ and yielded 7, 15, 16, 24, 19 and 7 captures, respectively at each occasion. The final population size estimate is $\hat{N} = 50$, with 95% CI (46, 58). Therefore $\hat{\mu}_L/\hat{N} = .24$. Would have we specified the estimated parameters, the expected length would have been .25. By holding these fixed, it can be estimated that an additional capture with parameter set as the average of the estimates would lead to $\hat{\mu}_L/\hat{N} = .19$, another one to .14. Therefore, a decrease of about 10% in the confidence interval length can be expected with only two additional sampling occasions.

4.5 Taxicabs Data

Xi et al. (2008) consider the Taxicabs data, which are assumed to be based on a Beta-binomial $M_h$ design and regard the number of taxicabs in Edinburgh. It is straightforward to realize that behavioural responses from taxicabs might not be expected. Additionally, the study was designed to achieve constant observed numbers each day, so no time-varying effects should be expected. For these data, $K = 10$ and $n = 283$. They conclude that the study is appropriately designed, in that even a value of $n = 210$ would have led to a good relative error for $\hat{N}$. The true population size is known to be $N = 420$. For these data, under the $M_h$ model considered we have $\hat{N} = 391$ and 95% CI (344,476); therefore $L_{CI}/N = 33.8\%$. We actually conclude that $K = 9$ would have sufficed, as the expected $L_{CI}/N$ would have been 37%, thus having a very mild loss in efficiency. Additionally, increasing $K$ does not lead to substantial improvements for $K \leq 19$; therefore reasonable additional sampling efforts are not likely to improve the estimates. Given our results, also assuming an $M_{bh}$ model with same (initial) mixing distribution would have led to the same results.

Suppose now that a binomial mixture (as in the first row of Table 1) is used as mixing distribution. Assuming two components with equal weight, $\pi_1 = .125$ and $\pi_2 = 0.33$, an expected length of 29.5% is similarly expected. In this case we consider the study as appropriately designed, as $K \leq 9$ would lead to a near-flat likelihood. Notably, $K = 12$ would not lead to reduce the confidence interval length by a factor of about two.

4.6 A newly designed experiment

We conclude with a newly designed experiment, to give additional insights on how to proceed in practice.

In a series of papers (Attorre et al. (2011); Senan et al. (2012); Attorre et al. (2013); De Sanctis et al. (2013); Attorre et al. (2014)) the second author and colleagues have investigated biodiversity in the island of Socotra (Yemen). In De Sanctis et al. (2013) we have identified 417 different plant species over the entire island territory. These have been characterized to identify clusters. A related problem would be to estimate the actual number of plant species in the island. This would be useful especially in the light of the large number of alien species detected: in Senan et al. (2012), 88 species have been identified as alien. Invasive alien species are a threaten to endemic ones, especially in arid environments like that of Socotra. Notably, these numbers are based on a single field survey, and some undercount is expected. There is evidence that sampling has been biased towards areas close to roads and human settlements. The fraction of species not detected might also be differential comparing endemic and alien ones.

A recapture experiment therefore might be designed where additional field studies are repeated similarly to the one of De Sanctis et al. (2013). We discuss how to design the experiment for estimation of the total number of plant species in the island. We speculate an additional occasion based on remote sensing
(e.g., through automatic analysis of Google Earth satellite images) can be added to avoid bias due to field sampling. Remote sensing might though be less sensitive because of some species being not visible. We therefore plan an $M_t$ study where $K - 1$ occasions are based on field studies having the same detection probabilities, and the last (based on satellite images) with a slightly smaller one.

In order to make hypotheses for planning we can proceed in several ways. One possibility, in absence of any information, is to investigate the literature for extent of undercount in similar surveys. For instance in Guézou et al. (2010) a coverage of about 80% is reported for a single field survey for a plant inventory of San Cristobal island in Galapagos. A better approach is to obtain a preliminary estimate based on the available data (or a pilot study). Given that plants from the same species are repeatedly identified in our field study, a lower bound estimate could be obtained using the well-known Chao (Chao, 1987) estimator. Chao’s estimator can be based on repeated entries in one list, therefore being amenable to the available data with $K = 1$. The number of species sampled exactly once was 161, while 156 were sampled exactly twice. Therefore we can expect, based on a single field survey, that the total number of plant species in Socotra is at least $417 + 161^2/(2 \times 156) = 500$, leading to an estimated coverage of 83.3% at least. This is somehow in agreement with the Guézou et al. (2010) coverage. Incidentally, the relative length of the confidence interval is 21%, definitely too large to be acceptable, which motivates a project to increase the number of capture occasions. Hence we plan $N = 500$, $p = 0.8$ for $K - 1$ occasions and $p = 0.5$ for the last one. Setting $K = 2$ would lead to an estimated relative length of the confidence interval of 8.8%, $K = 3$ to 2.7% and $K = 4$ to 1%. We find the expected outcome with $K = 3$ definitely acceptable, and therefore we plan an additional field survey and a remote sensing one.

## 5 Conclusions

In this paper we have proposed a general method for planning recapture experiments. The method is based on specifying the expected population size, sources of heterogeneity and nuisance parameters; and allows the user to set the number of sampling occasions (the only non-random parameter before the experiment) by calibrating sampling efforts and expected length of the confidence interval for the population size. The latter is based on the (asymptotic) expected value of the profile likelihood. We have introduced a general approach which can be applied numerically to basically any modeling assumption for closed populations. We then have detailed results regarding few cases which do not (entirely) require numerical approaches.

It shall be stressed that our method is grounded on asymptotic considerations, and therefore only approximate. Large sample results hold, nevertheless, for large $N$. There are no requirements that $n$, the number of observed individuals, is large. Analogously, the method is valid regardless of how small are the capture probabilities. If the capture probabilities are very small but $N$ is large there are no issues with the asymptotic results: simply, a larger $K$ will be needed in order to guarantee satisfactory results and stable profile likelihood. In simulations and real data examples we have seen that $N > 30$, approximately, is usually large enough for our results to be valid. There are exceptions, though, as we have seen in the Deer Mouse data example. We therefore recommend routinely checking the normality approximation via simulations as in Figure 4.

In planning the number of sampling occasions we recommend users to take into account the assumption that the population is closed. If sampling occasions correspond to consecutive days then including two or three additional occasions is usually safe. If they are farther apart in time, birth, deaths, emigration or immigration may occur within and the models assumed may not hold anymore.

In our data examples we found that $K$ was often well calibrated, but that adding as few as two capture occasions would often have lead to a substantial improvement in terms of length of the CI (and, consequently, standard error). In one case we have identified an over-powered experiment, and in another one a near-flat profile likelihood with an under-powered experiment.

An important open issue is how to pre-specify a model. While a rough idea of the expected population size might be available in advance, selection of the most appropriate model may be slightly more difficult.
In general considerations regarding the nature of the experiment should allow the user to specify a set of possible models. For instance if variation in weather conditions are expected, then the $M_t$ model may be dominating heterogeneity. If animals are slightly different (e.g. in size), and this characteristic is expected to influence the observation probability, then the $M_h$ model might be appropriate. Unfortunately it is not possible to provide general relationships between plans obtained through different models, unless several possibly unreasonable restrictions are assumed. Our best recommendation is to plan according to each plausible model (some can be certainly ruled out), and compare different plans based on plausible values for the nuisance parameters. Finally, the largest $K$ can be taken as the target number of occasions when collecting data.

Three possible models within the $M_{tbh}$ class have not been explored: $M_{tbh}$, $M_{th}$ and $M_{tb}$. These models require some restrictions on the parameters in order to be identifiable. The approach introduced in this paper could be used to plan the number of capture occasions, but the entire procedure would require numerical methods. See Farcomeni and Tardella (2010, 2012) for a general discussion on identifiability of models with unobserved heterogeneity, and e.g. Gimenez et al. (2003) for the related issue of parameter redundancy.

There are several routes for further work. The first regards exploration of closed form expressions for additional relevant cases, including lower bound estimates (Chao, 1987; Zelterman, 1988; Navaratna et al., 2008), persistence models (Ramsey and Usner, 2003), more general behavioral models (Yang and Chao, 2005; Farcomeni, 2011, 2015; Alunni Fegatelli and Tardella, 2015). The second regards extension of the approach to continuous-time closed recapture experiments (Yip et al., 1996; Farcomeni and Scacciatelli, 2013). Continuous-time experiments are rather different in nature than discrete ones, and in many cases only one of the two kinds can be performed. Additionally, in many cases the researchers plan to include covariates in the model. Most general cases of subject-specific covariates (e.g., Bartolucci and Forcina (2006); Thandrayen and Wang (2010); Farcomeni (2015)) though do not fit directly into our approach, as (i) all covariate values shall be known in advance in general and (ii) the covariates of unobserved individuals are not measured even after data collection. Finally, an interesting sampling design is given by inverse sampling (Chapman, 1952), where the number of recaptures instead of the number of sampling occasions is fixed in advance. A possibility is to explore how our general ideas can be extended to inverse sampling.

**Supplementary Materials**

Code referenced in Section 1 and theorems referenced in Section 2 are available as supplementary materials online.

**Acknowledgments**

The authors are grateful to the editor, the Associate Editor, and three referees for kind and stimulating comments. Acknowledgements go also to the Department of Environmental Biology of Sapienza - University of Rome, for permission to use the Socotra data. Alessio Farcomeni acknowledges financial support from the grant “New methods for planning of Phase II, Phase III trials and epidemiological experiments”.

**Conflict of Interest** The authors declare no conflict of interest

**References**


